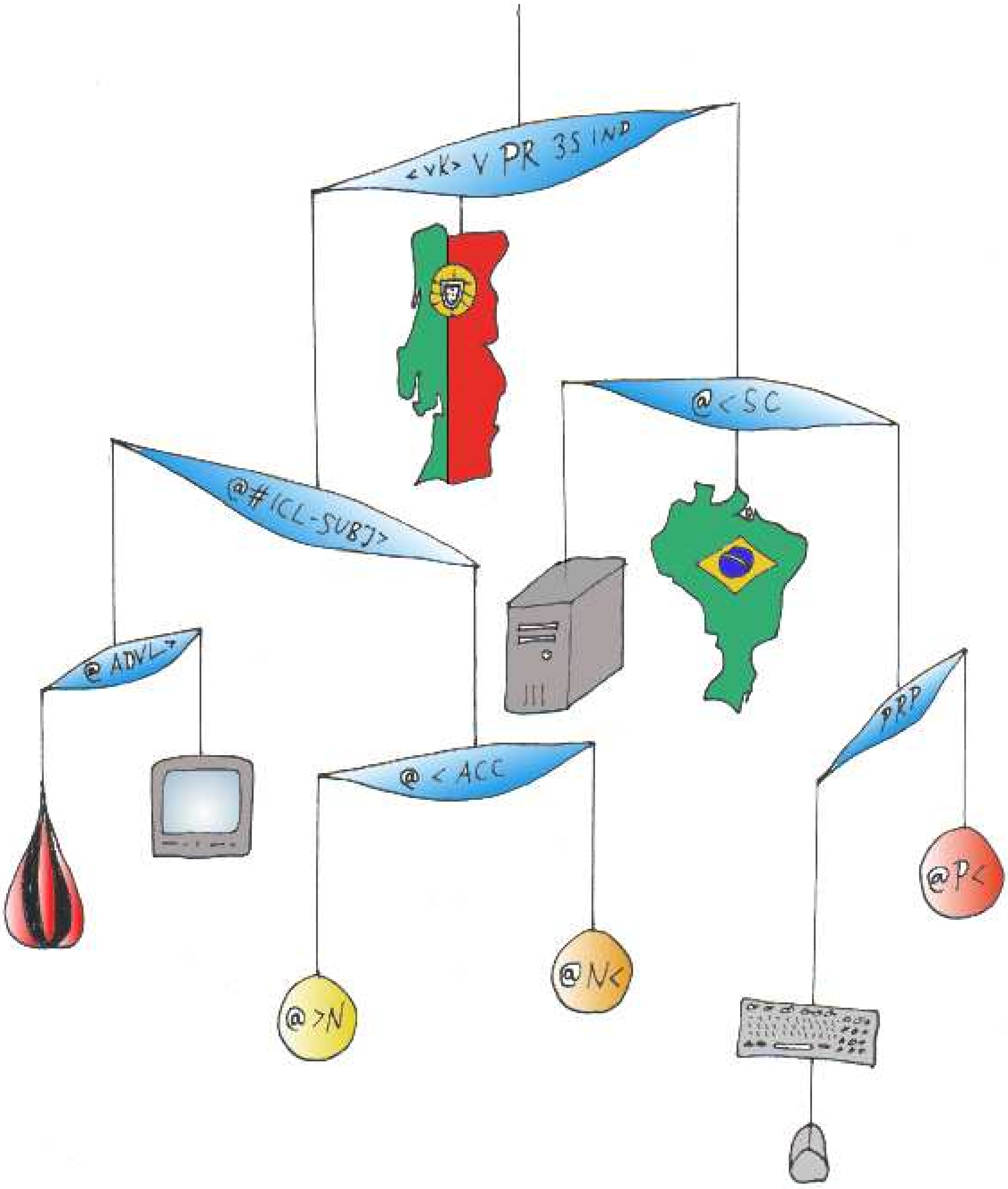
Eckhard Bick

THE PARSING SYSTEM "PALAVRAS"

Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework



Eckhard Bick

Department of Linguistics,
University of Århus, DK
lineb@hum.au.dk

♠

THE PARSING SYSTEM "PALAVRAS"

Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework

Dr. phil. thesis, defended in December 2000

(Project period 1994-1999)

Published in slightly revised form by Aarhus University Press

Cave: This postscript is based on the preliminary 1999 version with non-official page numbering! Another difference to be noted is that printed version contains updates regarding new corpus material annotated with PALAVRAS.

Abstract

The dissertation describes an automatic grammar- and lexicon-based parser for unrestricted Portuguese text. The project combines preceding and ongoing lexicographic work with a three-year Ph.D.-research effort on automatic grammatical annotation, and has since ventured into higher level syntactic and semantic analysis. Ultimately the parser is intended for applications like corpora tagging, grammar teaching and machine translation, which all have been made accessible in the form of internet based prototypes. Grammatical rules are formulated in the Constraint Grammar formalism (CG) and focus on robust disambiguation, treating several levels of linguistic analysis in a related manner. In spite of using a highly differentiated tag set, the parser yields correctness rates - for unrestricted and unknown text - of over 99% for morphology (part of speech and inflexion) and about 97% for syntactic function, even when geared to full disambiguation. Among other things, argument structure, dependency relations and subclause function are treated in an innovative way, that allows automatic transformation of the primary, "flat" CG-based syntactic notation into traditional tree structures (like in DCG and PSG). The parser uses valency and semantic class information from the lexicon, and a pilot study on disambiguation on these levels has been conducted, yielding encouraging results.

The system runs at about 400 words/sec on a 300 MHz Pentium II based Linux system, when using all levels. Morphological and PoS disambiguation alone approach 2000 words/sec.

Contents

1

Introduction

1.1 The ‘what’s, ‘why’s and ‘who’s

This dissertation is about whether and to what degree a computer program can be made to handle the grammatical analysis of natural language, in the form of ordinary, “running” text or linearly transcribed speech. The target language chosen is Portuguese, and the basic method applied in the parser to be described here is Constraint Grammar (first introduced in Karlsson, 1990), used in a context of Progressive Level Parsing[1]. Along the way, I will be concerned with the interaction between grammar system, parsing technique and corpus data, evaluating the trinity’s mutual influence, and the performance of the system as a whole. In other words, in computer linguistics, what can computers offer a linguist, and can linguistics inspire computing?

Yet before trying to answer these questions with a 400-page bore of technicalities and a load of secondary questions, it would seem relevant to balance the introduction by asking quite another type of question: Why would any of this inspire a *person*? Why would anybody want to court a computer for half a decade or more? Well, personally - and may the esteemed reader please feel free to skip the next half page or so -, I find that the most intriguing fact about computers is not their data-crunching efficiency, nor their much-appraised multimedia capability, but the plain fact that they react to stimuli in much the same half-predictable-half-unpredictable way biological entities do. Computers *communicate,* and many a nerd has found or created a social surrogate in his computer.

When I had my first naive date with a computer in 1973, the glorious glittering consumer items of today weren’t called PC’s – or even Mac’s – but went by the humble name of *Wang*. They had no hard disc, no floppies or CD-ROM’s, and 4 kB of RAM rather than 40 MB. Yet, in a subtle way, human-computer relations were superior to the uses most computers are put to today. *Nowadays*, most people treat computers as tools: Gaming devices, mail boxes, type-writers, - all of which, in different shapes, did exist before the advent of the computer. *Then*, children could not shoot their way trough a boring day by handling fire-buttons, joy-sticks and mouse-ears. They had to *program* their computer if they wanted it to play a game. And the computer would respond, as a student surpassing her teacher, by route, at

[1] Progressive Level Parsing is mirrored by the order of chapters in this book, which progresses from morphological analysis and the lexicon to morphological disambiguation, syntax, semantics and applicational considerations. This is why a discussion of the Constraint Grammar disambiguation formalism as such is “postponed” until chapters 3.5 and 3.6. Though I have tried to avoid literal CG rule quotes in the first chapters, there may be a few passages (notably 2.2.4 and 3.2-3) where readers not familiar with the basic notational conventions of CG might want to use later chapters for reference.

first, - but soon, it would move the bricks in unpredictable ways, it would be the sentient being, thinking, reacting, surprising you.

This is what has fascinated me ever since I made my school's *Wang* play checkers. With my projects evolving from the unprofessionally naive to the unprofessionally experimental, I programmed creativity by filtering random input for patterns and symmetry, I made my own *Eliza*, I built self-learning teaching tools, and I tried to make a computer *translate*. I was thrilled by the idea of a perfect memory in my digital student, the instantaneous dictionary, by never having to learn a piece of information twice.

Along the way things became somewhat less unprofessional, and I accumulated some experience with NLP, constructing machine-readable dictionaries for Danish, Esperanto and Portuguese, and – in 1986 – a morphological analyser and MT-program for Danish[2]. Then – in 1994 – I heard a highly contagious lecture by Fred Karlsson presenting his Constraint Grammar formalism for context based disambiguation of morphological and syntactic ambiguities. I was fascinated both by the robustness of the English Constraint Grammar (Karlsson et. al., 1991) and its word based notational system of *tags* integrating both morphology and flat dependency syntax in a way that allowed easy handling by a computer's text processing tools. It was not clear at the time (and still is not) up to which level of syntactic or even semantic analysis Constraint Grammar can be made to work, and it had never – at any larger scale – been applied to Romance languages. So I decided to try it out on Portuguese[3], working upwards from morphology to syntax and semantics, in the framework of a Ph.D. project in Computer Linguistics. The goal was the *automatic analysis* of free running Portuguese text, i.e. to build a computer program (a morphological tagger and a syntactic parser) that would take an ordinary text file - typed, mailed or scanned - as input and produce grammatically analysed output as unambiguous and error-free as possible. My ultimate motivation, the *raison d'être* of my digital child, has always been applicational – encompassing the production of research corpora[4], communication and teaching tools, information handling and, ultimately, machine translation. But in the process of making the digital toddler walk, I would have to fight and tame *the Beast* , as my supervisor Hans Arndt called it, the ever-changing and multi-faceted creation which is human language. I would have to chart the lexical landscape of Portuguese, to define the categories and structures I would ask my parser to recognise, and to check both tradition, introspection and grammatical intuition against raw and real corpus data. Many times, this process has turned back on itself, with the dynamics of the "tool grammar" (i.e. the growing Constraint Grammar rule set) forcing new distinctions or

[2] This system - "Danmorf" - has been revived in 1999, to become the morphological kernel of the Danish "free text" section of the VISL-project at Odense University, and can be visited at http://visl.hum.sdu.dk.

[3] Romance languages, with the possible exception of French, share much of their syntactic structure, and also most morphological categories. Even many lexical items, not least pronouns and conjunctions, can often be matched one-on-one across languages. At the time of writing (1999), I have begun to adapt my Portuguese Constraint Grammar for Spanish, with encouraging results (http://visl.hum.sdu.dk).

[4] The largest annotation task so far, completed in november 1999, has been tha annotation of a 90 million word corpus of Brazilian Portuguese, for a research group at the Catholic University of São Paulo.

definitions on the ”target grammar” (i.e. the particular grammatical description of Portuguese to be implemented by my system).

1.2 The parser and the text

This dissertation is a Janus work, both practical and theoretical at the same time, one face mirroring and complementing the other. After all, a major point was simply showing that “it could be done” - that a Constraint Grammar for a Romance langugage would work just as well as for English.

As a *practical product*, the parser and its applications can speak for themselves, and, in fact, do so every day – at http://visl.hum.sdu.dk/ - , serving users across the internet. In what could be called the *theoretical* or*text* part of this dissertation, apart from discussing the architecture and performance of the parser, I will be concerned both with the process of *building* the parser and with its linguistic spin-off for Constraint Grammar and parsing in general, and the analysis of Portuguese in particular. Both tool and target grammar will be discussed, with chapter 3 focusing on the first, and chapter 4 focusing on the second.

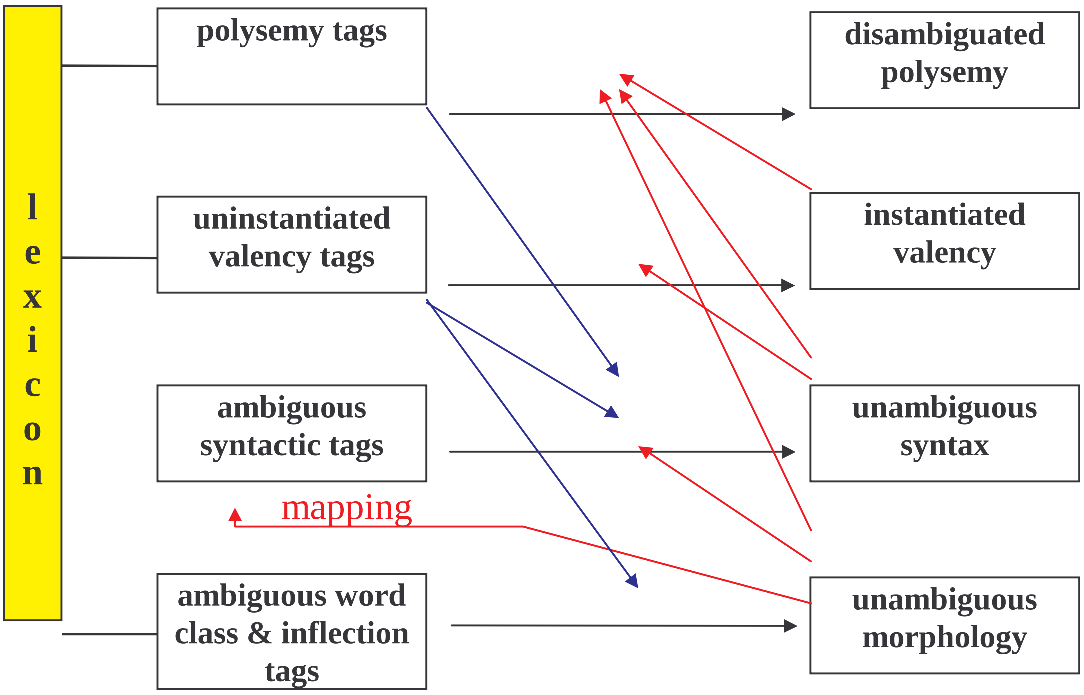
Chapter 2 describes the system’s lexicon based morphological analyser, and since the quality of any CG-system is heavily dependent on the acuracy and coverage of its lexico-morphological input base, the analyser and its lexicon constitute an important first brick in the puzzle. However, chapters 2.1, 2.2 and 2.3, which treat the architecture of the program as such, as well as the interplay of its root-, suffix-, prefix- and inflexion-lexica, are rather technical in nature, and not, as such, necessary to understand the following chapters, which may be addressed directly and individually. In 2.2.4, the Beast will raise its head in the section on the *dynamic lexicon*, where non-word words like abbreviations, enclitics, complex names and polylexical expressions are discussed, and the principle of structural morphological heuristics is explained. 2.2.5 is a reference chapter, where morphological word classes and inflexion features are defined, and 2.2.6 quantifies the analyser’s lexical coverage.

Chapter 3 introduces the Constraint Grammar formalism as a tag based disambiguation technique, compares it to other approaches, and discusses the types of ambiguity it can be used to resolve, as well as the lexical, morphological and structural information that can be used in the process. It is in chapter 3 that the ”tool grammar” as such is evaluated, both quantitatively and qualitatively, with special emphasis on level interaction and rule typology. Finally, the system’s performance is measured on different types of text (and speech) data and for different levels of analysis.

”Level interaction” is central to the concept of Incremental Parsing (or Progressive Level Parsing) and addresses the interplay between lower level tags (already disambiguated), same level tags (to be disambiguated) and higher level ”secondary” tags (not to be disambiguated at the stage in focus). Parsing is here viewed as a progression through different levels of analysis, with disambiguated morphological tags allowing syntactic mapping and disambiguation, syntactic tags allowing instantiation of valency patterns and all three contributing to semantic disambiguation.

In the illustration below, red upward arrows indicate disambiguation context provided by lower level ”primary” tags, blue downward arrows indicate disambiguation context provided by higher level ”secondary” tags.

Table (1): Parsing level interaction



Chapter 4 discusses the target grammar, especially on the syntactic level. The form and function categories used by the parser are defined and explicated, with special attention paid to verb chains, subclauses and adverbials. In the process I will sketch the outlines of a dependency grammar of Portuguese syntax that has been grown from the iterative interaction of corpus data and a dynamic CG rule system which structures such data by introducing and removing ambiguity, a process in which my linguistic perception of the object language (the Beast, so to say) had to reinvent itself continuously, on the one hand serving as a necessary point of departure for formulating *any* rule or ambiguity, on the other hand absorbing and assimilating corpus evidence of CG-elicited (or CG-disclaimed) distinctions. Finally, I will raise the question of the transformational potential of the Portuguese CG with regard to different theories of syntax. In particular I will argue that the traditional flat dependency syntax of CG can be enriched (by attachment direction markers and tags for subclause form and function) so as to allow transformation of a CG-parse into

constituent trees. Advantages and draw-backs of different notational systems of parsing output will be weighed regarding computational and pedagogical aspects as well as the expression of ambiguity.

Chapter 5 treats valency tagging, focusing not so much on valency patterns as such (which are treated in chapters 3 and 4), but rather on the role of valency tags as an intermediate CG stage linking syntactic to semantic parsing. Also, I will defend why using syntactic function tags for the instantiation of lexically derived tags for valency potential is *not* a kind of self-fulfilling prophecy, but a productive part of grammatical analysis.

In **Chapter 6**, I will discuss the highest - and most experimental - level of CG based Progressive Level Parsing, - semantics. It is the semantic level that most clearly shows the disambiguation potential residing in the interplay of tags from different levels of grammatical analysis. Thus, morphosyntactic tags and instantiated valency or dependency tags will be exploited alongside semantic tags proper and hybrid tags imposing semantic restrictions on tags for valency potential. Teleologically, polysemy resolution will be treated from a bilingual Portuguese-Danish perspective, allowing differentiation of translation equivalents. I will argue that - by using minimal distinction criteria and atomic semantic features for the delineation of semantic prototypes - semantic tagging is entirely possible without achieving full definitional or referential adequacy. However, though a complete system of semantic tagging will be presented for nouns, and a basic one for verbs and adjectives, and though the tag set has been incorporated into the whole (Portuguese) lexicon, the CG rule body concerned with semantics is still small compared to the rule sets used for lower level parsing. Therefore, definite conclusions cannot be drawn at present, and performance testing had to be sketchy and mostly qualitative at this level[5].

Chapter 7, finally, explores some of the possible applications of the parser, machine translation, corpus tools and grammar teaching programs. Corpus annotation is the traditional field of application for a parser, not much additional programming is needed, and an annotation is about as good or bad as the parser performing it[6]. In machine translation, however, parsing (even semantic parsing) solves only “half the task”, since choosing translation equivalents and performing target language generation evidently cannot be achieved without additional linguistic processing. I will show how an additional layer of CG rules can be used not for analysis, but for generation, and how CG tag context can be exploited for syntactic transformations and morphological generation. Grammar teaching on the internet, on the other hand, is an example where the parser forms not the core of a larger linguistic program

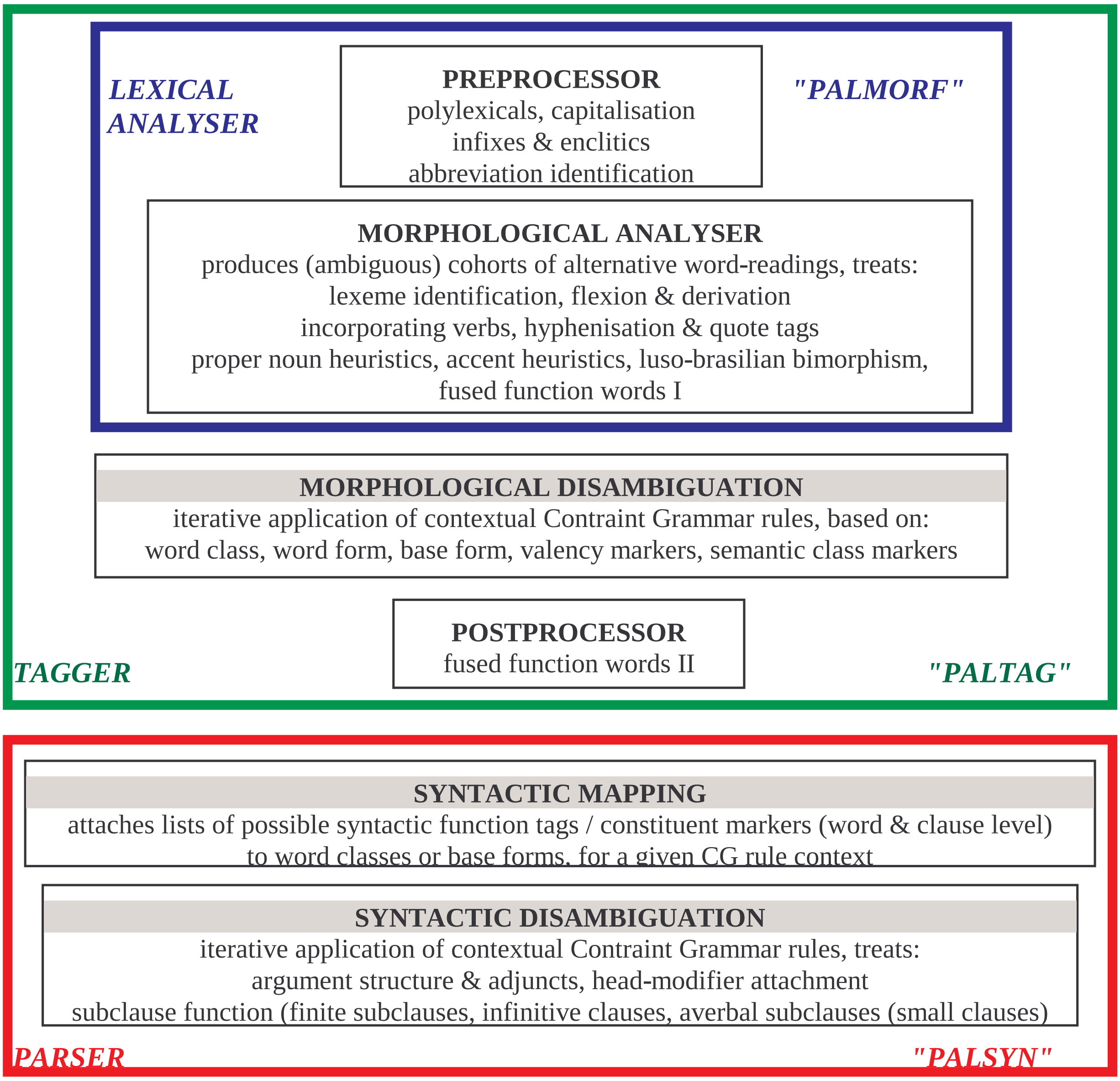
[5] A three year research grant (1999-2001) from Statens Humanistiske Forskningsråd, at Odense University, for a project involving Portuguese, English and Danish CG semantics, is hopefully going to change that.

[6] Most annotation today still means tagging with word based PoS tags, which are easy to handle with string searching tools, but lack syntactic information. The CG-approach, however, is robust and word based even on the syntactic level, allowing syntactic tag searches in the same fashion as used for PoS tags.

chain, but rather the linguistic core of a heterogeneous program chain whose other parts serve graphical and pedagogical purposes. Still, there are linguistic constraints, since an independent pedagogical application imposes a certain system of grammatical theory as well as notational conventions on the parser's output, and as an example I will discuss the automatic transformation of CG output into syntactic tree structures.

Throughout the text, frequent and unavoidable use is made of the parser's tags and symbols. Where these are not explained or clear from context, one can find the necessary definitions and examples in the "tag list" appendix. The parser's individual modules will be discussed in input-output order, i.e. in the order of the parser's program chain. The following illustration summarises module functions and sequentiality for the parser proper and its MT add-ons:

Table (2): Parsing modules



LEXICAL ANALYSER

PREPROCESSOR
polylexicals, capitalisation
infixes & enclitics
abbreviation identification

"PALMORF"

MORPHOLOGICAL ANALYSER
produces (ambiguous) cohorts of alternative word-readings, treats:
lexeme identification, flexion & derivation
incorporating verbs, hyphenisation & quote tags
proper noun heuristics, accent heuristics, luso-brasilian bimorphism,
fused function words I

MORPHOLOGICAL DISAMBIGUATION
iterative application of contextual Contraint Grammar rules, based on:
word class, word form, base form, valency markers, semantic class markers

POSTPROCESSOR
fused function words II

TAGGER

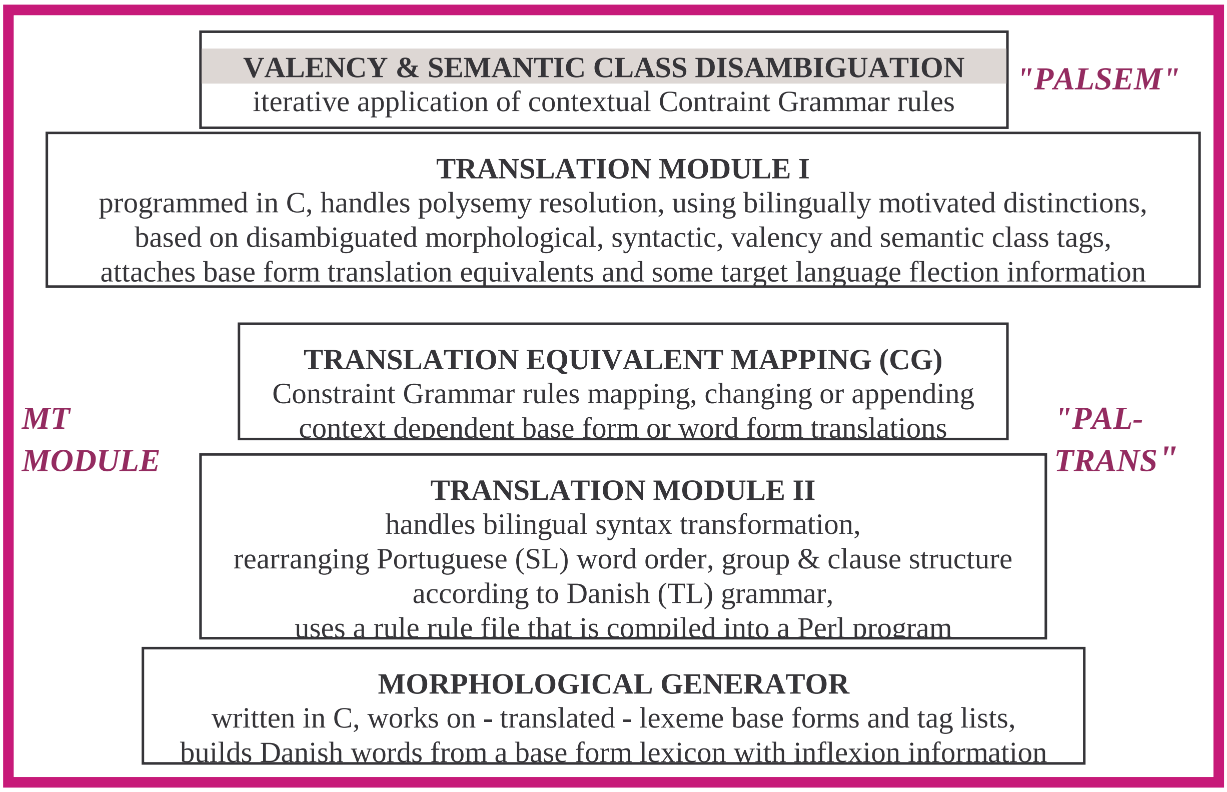
"PALTAG"

SYNTACTIC MAPPING
attaches lists of possible syntactic function tags / constituent markers (word & clause level)
to word classes or base forms, for a given CG rule context

SYNTACTIC DISAMBIGUATION
iterative application of contextual Contraint Grammar rules, treats:
argument structure & adjuncts, head-modifier attachment
subclause function (finite subclauses, infinitive clauses, averbal subclauses (small clauses)

PARSER

"PALSYN"



VALENCY & SEMANTIC CLASS DISAMBIGUATION
iterative application of contextual Contraint Grammar rules

"PALSEM"

TRANSLATION MODULE I
programmed in C, handles polysemy resolution, using bilingually motivated distinctions,
based on disambiguated morphological, syntactic, valency and semantic class tags,
attaches base form translation equivalents and some target language flection information

MT MODULE

TRANSLATION EQUIVALENT MAPPING (CG)
Constraint Grammar rules mapping, changing or appending
context dependent base form or word form translations

"PAL-TRANS"

TRANSLATION MODULE II
handles bilingual syntax transformation,
rearranging Portuguese (SL) word order, group & clause structure
according to Danish (TL) grammar,
uses a rule rule file that is compiled into a Perl program

MORPHOLOGICAL GENERATOR
written in C, works on - translated - lexeme base forms and tag lists,
builds Danish words from a base form lexicon with inflexion information

2

The lexicomorphological level: Structuring words

2.1 A lexical analyser for Portuguese: *PALMORF*

PALMORF is a so-called morphological or lexical analyser, a computer program that takes running text as input and yields an analysed file as output where word and sentence boundaries have been established, and where each word form or "word-like" polylexical unit is tagged for word class (PoS), inflexion and derivation/composition, with morphologically ambiguous words receiving multiple tag lines. The notational conventions used by PALMORF match the input conventions for a CG disambiguation grammar. With a CG-term, an ambiguous list of morphological readings, as in (1), is called a *cohort*.

(1)

WORD FORM	BASE FORM	SECONDARY TAGS	PRIMARY TAGS	
revista				
	"revista"	<+n> <rr> <CP>	N F S	‘magazine’,‘inspection’
	"revestir"	<vt> <de^vtp> <de^vrp>	V PR 1/3S SUBJ VFIN	‘to cover’
	"revistar"	<vt>	V IMP 2S VFIN	‘to review’
	"revistar"	<vt>	V PR 3S IND VFIN	
	"rever"	<vt> <vi>	V PCP F S	‘to see again’,‘to leak’

In example (1), the word form 'revista' has been assigned one noun-reading (female singular) and four verb-readings, the latter covering three different base forms, subjunctive, imperative, indicative present tense and participle readings. By convention, PoS and morphological features are regarded as primary tags and coded by capital letters. In addition there can be secondary lexical information about valency and semantic class, marked by <> bracketing, like <vi> for intransitive verbs (“rever” - ‘leak through’) , <vt> for monotransitive verbs (“rever” - ‘see again’), <+n> for pre-name distribution (“revista VEJA” - ‘VEJA magazine”), <rr> for 'readable object' or <CP> for +CONTROL and +PERFECTIVE ASCPECT (“revista” - ‘review’).

(2)

WORD FORM	BASE FORM	SECONDARY TAGS	PRIMARY TAGS
(i) *telehipnotizar*			
	"hipnotizar"	<vt> <vH> <DERP tele->	V INF 0/1/3S
	"hipnotizar"	<vt> <vH> <DERP tele->	V FUT 1/3S SUBJ VFIN
(ii) *corruptograma ALT xxxograma*			
	"corrupt"	<HEUR> <DERS -grama>	N M S

(iii) *corvos-marinhos*
"corvo-marinho" <orn> N M P
(iv) *Estados=Unidos*
"Estados=Unidos" <*> <top> PROP M P

(2) offers examples for derivational tags (DERP for prefixes and DERS for suffixes), as well as polylexical word boundaries (the '=' sign in (iv) is introduced by the tagger to mark a non-hyphen polylexical link). Also purely orthographic or procedural information can be added to the tag list, like <*> for capitalisation or <HEUR> for use of the heuristics module[7].

The morphological analyser constitutes the lowest level of the PALAVRAS parsing system, and feeds its output to Constraint Grammar morphological disambiguation, and ultimately to the syntactic and semantic modules. PALAVRAS was originally designed for written Brazilian Portuguese, but now recognises also European Portuguese orthography and grammar, either directly (lexical additions) or - if necessary - by systematic orthographic variation (pre-heuristics module).

Not all registers prove equally accessible to automatic analysis, thus phonetic dialect spelling in fiction texts or phonetically precise transcription of speech data, for instance, cause obvious problems. Scientific texts can have a very rich vocabulary, but many of the difficult words are open to systematic Latin/Greek based derivation, which has been implemented in PALAVRAS. News texts often contain many names, but name candidate words can be identified quite effectively by heuristic rules based on capitalisation, in combination with character inventory and immediate context (cp. chapter 2.2.4.4). Only words *derived* from names (e.g. adjectives) and chemical or pharmaceutical names evade this solution by not being capitalised, and need to be treated by another morphological heuristics module, also used for misspellings, foreign loan words and the few Portuguese words that are both not listed in the PALAVRAS lexicon, and underivable for the analyser (cp. 2ii).

PALAVRA's typical lexical recognition rate is 99.6-99.9% (cp. chapters 2.2.4.7 and 2.2.6). In these figures a word is counted as “recognised” if the correct base form or derivation is among those offered (ambiguity is only resolved at a later stage), and if propria are recognised as such (though without necessarily matching a lexicon entry).

2.2 The program and its data-bases

2.2.1 Program specifications

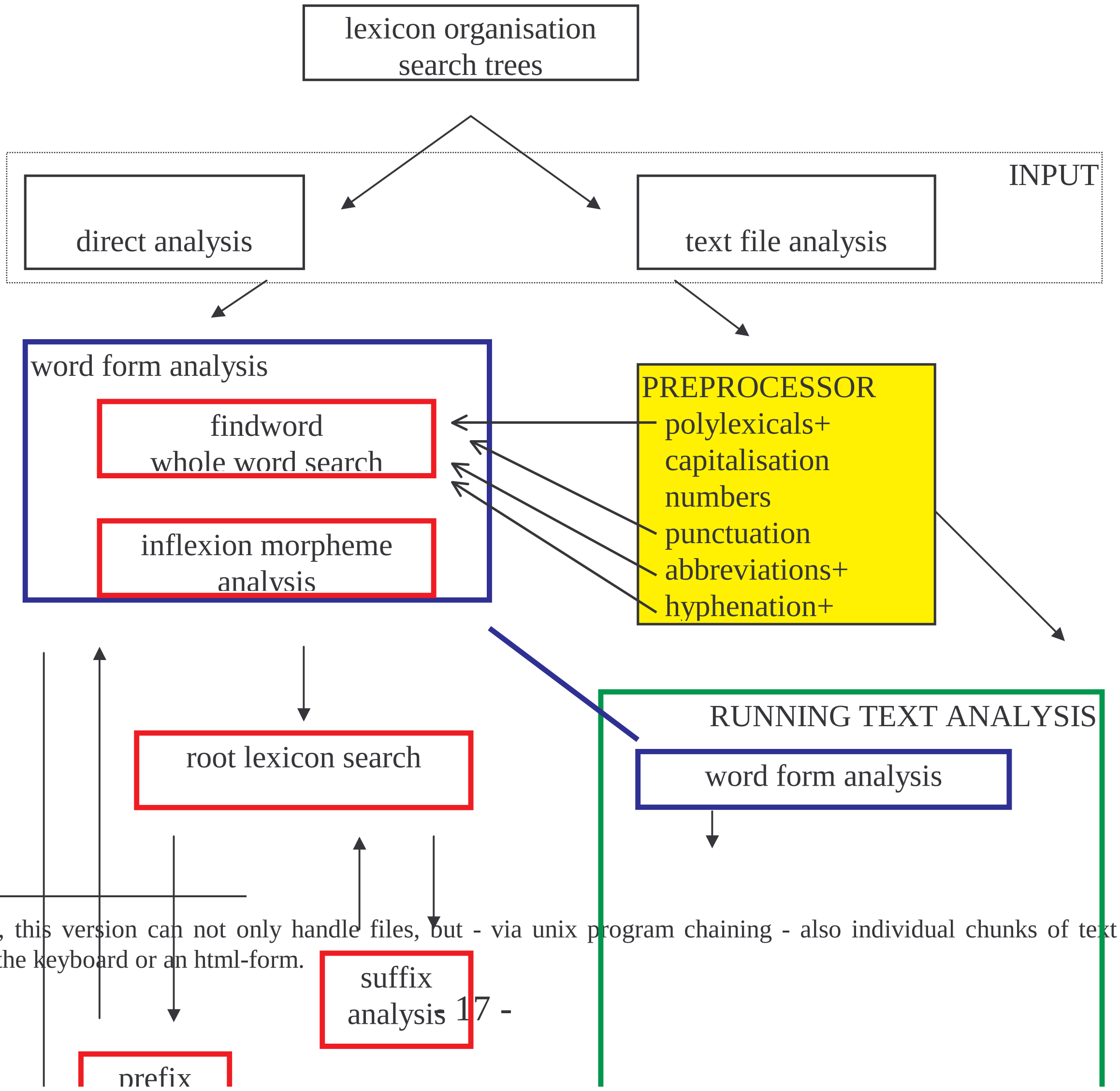
[7] Any orthographical changes introduced by the tagger's heuristics module - spelling/accent correction etc. - is marked with an ALT-tag after the original word form. The xxx in (ii) means a hypothesized root not found in the current PALAVRAS lexicon, or one normally disallowed by inflexional or word class - affix combination rules.

The core of PALMORF is written in C and runs on UNIX or MacOS platforms, tagging roughly 1000 words a second (preprocessing included). It consists of about 4000 lines of source code (+ most of the ANSI library), some 2000 lines of grammatical inflexion and derivation rules, and a 75.000 entry electronic lexicon. Due to the way the lexicon is organised at run time, the program requires some 8 MB of free RAM. For additional pre- and postprocessing, PALMORF is aided by a number of smaller filter programs written in Perl.

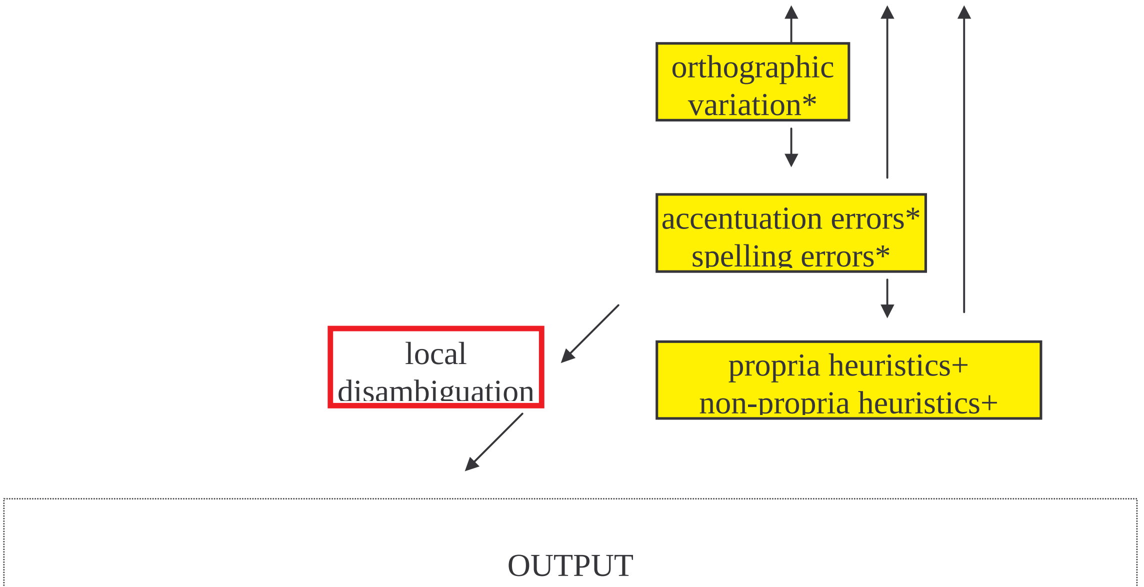
2.2.2 Program architecture

2.2.2.1 Program modules

Below, the basic "flow chart" structure of the PALMORF program is explained. Basically, there is a choice between one-word-only direct analysis and file-based[8] running text analysis, the latter featuring preprocessing and heuristics modules where also polylexicals, abbreviations, orthographic variation and sentence boundaries can be handled, as well as some simple context dependent heuristics. Both program paths make use of the same inflexion and derivation modules, that are applied recursively until an analysis is found, and hereafter, until *all* analyses of the same or lower derivational depth are found. A more detailed discussion of the program architecture of PALMORF can be found in the appendix section.



[8] Of course, this version can not only handle files, but - via unix program chaining - also individual chunks of text entered via the keyboard or an html-form.


orthographic variation*
accentuation errors* spelling errors*
local disambiguation
propria heuristics+ non-propria heuristics+
OUTPUT

As shown in the diagram (yellow boxes), PALMORF - or rather its preprocessor and heuristics modules - is quite capable of “meddling” with its data. Still, orthographic intervention as such (*) is used only heuristically, where no ordinary analysis has been found, and the altered word forms are marked 'ALT', so they can be identified later, for example for output statistics, and for the sake of general corpus fidelity. Affected areas are luso-brazilian orthographic variation (e.g. oi/ou digraphs, ct -> t, cp -> p), typographically based accentuation errors (e.g. 7-bit-ASCII vs. 8-bit-ASCII input) and some common spelling errors (e.g. cão -> ção, çao -> ção).

2.2.2.2 Preprocessing

Unlike post-analysis heuristics, preprocessor intervention (+) applies to all input, and is close to being a general parsing necessity. Among other things, a natural and unavoidable step in all NLP is the decision of *what* to tag. Obviously, in a word based tagger and a sentence based parser, this amounts to establishing word and sentence boundaries.

First, the preprocessor strives to establish what is *not* a word, and marks it by prefixing a $-sign: $. - $, - $(- $) - $% -$78.7 - $± - $” - $7:20 etc. Of these, some are later treated as words anyway. Thus, numbers will be assigned the word class NUM and a syntactic function, $% will be treated as a noun (N), $7:20 as a time adverbial. Punctuation is treated in four ways:

(a) as sentence delimiter. Ordinarily, it is the DELIMITERS list of the CG rule file that determines *which* punctuation marks are treated as sentence boundaries (e.g. $. and $:, but not $- and $,). However, the preprocessor can add sentence delimiters (¶) where it identifies sentence-final abbreviations, or - for instance - instead of double line feeds around punctuation-free headlines.

(b) as a regular non-word. Such punctuation is shown in the analysis file without a tag (e.g. $: or $!), but can still be referred to by CG-rules.

(c) as tag-bearing “words”. This is unusual in a Constraint Grammar, but $% (as a noun) is an example, and $, as a co-ordinator (like the conjunction ‘e’) is another one.

(d) as part of words. For instance, $” will become a <*1> tag (left quote border) if attached left of an alphanumeric string, and <*2> (right quote border) if attached right. Also, abbreviations often include punctuation (. , - /), which is especially problematic, since ambiguity with regard to sentence boundary punctuation arises. To solve the ambiguity, the preprocessor consults an abbreviation lexicon file and checks for typical sentence-initial/final context or typical context for individual abbreviations.

Second, the preprocessor separates what it thinks are words by line feeds. Here, the basic assumption of word-hood defines words as alphanumeric strings separated by blank spaces, hyphens, non-abbreviation-punctuation, line feeds or tabs. The reason for including hyphenation in the list is the need to morphologically analyse enclitic and mesoclitic pronouns (e.g. ‘dar-lhe-ei’), and to decrease the number of - lexiconwise - unknown words: The elements of hyphenated strings can thus be recognised and analysed individually by the PALMORF analyser, even if the compound as such does not figure in the lexicon. Thus, a word class and inflexional analysis can usually be provided and passed on to the syntactic and higher modules of the parser, even if only the last part of a hyphenated string is “analysable”.

Third, for pragmatic reasons, a number of *polylexicals* has been entered in the PALMORF lexicon, consisting of several space- or hyphen-separated units that would otherwise qualify as individual words (e.g. ‘guarda-chuva’, ‘em vez de’). These polylexicals have been defined ad hoc by parsing needs (e.g. complex prepositions), semantic considerations (machine translation) or dictionary tradition. Polylexicals are treated like ordinary words by the parser, i.e. assigned form and function tags etc., and can be addressed as individual contexts by Constraint Grammar rules. In the newest version of the parser, one type of polylexical is assembled independently of existing lexicon entries: Proper noun chains are fused into polylexical “words” if specified patterns of capital letters, non-Portuguese letter combinations and name chain particles (like ‘de’, ‘von’, ‘van’ etc.) are matched.

Criteria for the heuristic identification of non-Portuguese strings are, among others, letters like ‘y’ and ‘w’, gemination of letters other than ‘r’ and ‘s’, and word-final letters other than vowels, ‘r’, ‘s’ and ‘m’. Apart from name recognition, identification of non-Portuguese strings is useful in connection with hyphenated word chains - which will not be split if they contain at least one non-Portuguese element, in order to avoid “accidental” (i.e. affix or inflexion-heuristics based) assignment of non-noun word class[9].

2.2.2.3 Data bases and searching techniques

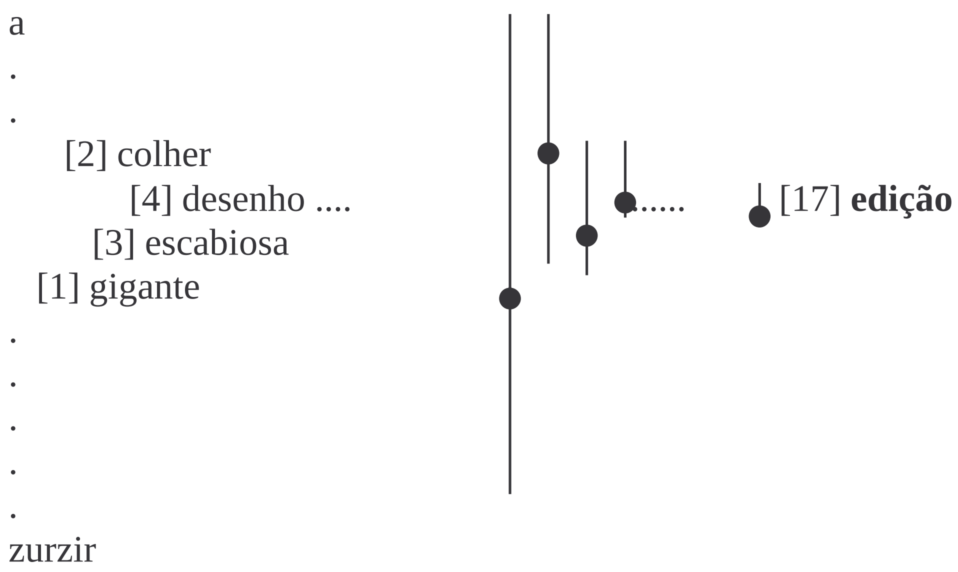
On start-up the program arranges its data-bases in a particular way in RAM:

a) the *grammatical lexicon* is organised alphabetically with grammatical information attached to the head word string. Each grammatical field has its own pointer. The

[9] N (noun) and PROP (proper noun) are the overwhelminly most common word classes for foreign language material in Portuguese.

alphabetical order allows the analyser to find word roots by binary search: 5 steps to search 16 words, 6 steps to search 32 words, 17 steps to search the whole lexicon (fig. 1). In analysing a particular word, multiple root searches are even faster: due to the fact that cutting various endings or suffixes off a word does not touch word initial letters, the remaining roots are alphabetically close to each other. So, having found the first root by cutting the lexicon in halves 17 times, one can get near the next root by a few "doubling up" steps from the first roots position. Normally this takes less than 5-6 steps.

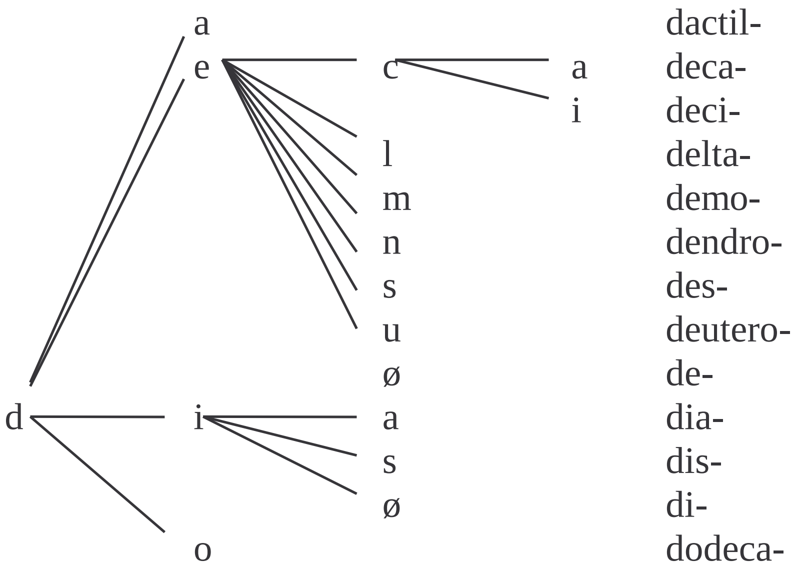
(1) binary search technique:



b) the *inflexion endings* are stored retrograde alphabetically in a sequential list, with combination rules, base conditions and tagging information attached in successive fields. For speedy access, position line numbers and block size for homonymous endings are stored separately. The first look-up of an ending controls the next, working backwards from the end of a word, thus minimising access time: in "comeis", for instance, -s is looked up first, then -is (in a list also featuring -as, -es, -is, -os etc.) and last -eis (in a list also containing -ais, -eis, -óis etc.); once "knowing" about the ending -s, the system does not have to compare for, say, -eio.

c) the *suffix* and *prefix* lexicons are both stored in the form of alphabetical pointer trees (fig. 2), with the suffixes inverted. To find, for instance, the prefix "dis-", the program looks under "d-", which points to a,e,i and o as second letter possibilities; "i" is selected, giving a choice between a and s ("dia-" and "dis-"). Finally we get d-i-s with a stop-symbol after the s. The last pointer gives access to the combination rules, base condition and tagging information concerning the chosen prefix. For suffixes the letter searching order is reversed: "-inho" is thus found as o-h-n-i. The pointers themselves are memory cells with C-style pointer addresses pointing to the next level row of letters each itself associated with a new pointer address, leading to ever finer branchlets of the letter-tree.

(2) pointer tree searching technique (d-segment of the prefix lexicon)



2.2.3 Data structures

2.2.3.1 Lexicon organisation

The electronic lexicon that PALMORF uses, is based on a paper version passive bilingual Portuguese-Danish dictionary (Bick, 1993, 1995, 1997) I have compiled in connection with my Masters thesis on lexicography (Bick, 1993), which is where information can be found about the lexicographic principles applied. The lexicon file now covers over 45.000 lexemes, 10.000 polylexicals and about 20.000 irregular inflexion forms. The present lexical content reflects the constant, circular interactivity of lexicon, parser and corpus. Over four years, every parse has - also - been a lexicon check.

Much of the information contained in the original dictionary had to be regularised and adapted for parsing purposes. Thus, many words had to have their valency spectrum widened for empirical reasons, and throughout the whole lexicon, a formal semantic classification was introduced, something a human reader of the paper dictionary would implicitly derive from the list of translation alternatives. Also, for use with regular inflexion rules, grammatical combinatorial subcategories (field 4 in table 2) had to be introduced for verbal (and some nominal) stems.

In (1), a number of authentic lexicon entries is listed, and table (2) summarises the kind of information that can be found in the different fields of a lexicon entry.

(1)

```
abalável#=#<amf>#TP######46
abalôo#2oar#<v.PR 1S>######52#
abana-moscas#=#<sfSP.il>###[ô]####57
acapachar-#1#<vt>#AaiD#####<vr>#412
```

```
acapitã#=#<sf.orn>####B(orn)###413
acara#acará#<sm.ich>#R###TU(ich)##414#
acaraje#acarajé#<sm.kul>#R###IO(kul)##415#
acarajé#=#<sm.kul>####IO(kul)###415
acertar-#1#<vt>#AaiD#<R[é]>####<vi>#481
acerto#=#<sm.am>###[ê]###<cP><tegn>#484
acervo#=#<sm.qus>###[ê/é]####486
aceráceas#=#<sfP.B>####(bo)###473
aceso#=#<adj>###[ê]####487
acessivel#acessível#<amf>#RTP#####490#
acetona#=#<sf.liqu>###[ô=]#(km, med)###498
alcatraz#=#<sm.orn>####AR(orn)#corvo-marinho##1741
algo#=#<SPEC M S>#######1924
algum#=#<DET M S.quant2>#<f:-a, P alguns/algumas>######1943
aliviar-#1#<vt>#AaiD#<R['i]>####<vi><vr>#2045
along-#alongar#<var>#B#####2133#
alongar-#1#<vt>#AaiD#<g/gu>####<vr>#2133
alongu-#alongar#<var>#Cc#####2133#
```

(2) **PALAVRAS lexicon fields**

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
word root	**base form**	**word class** (+ primary syntax or sem. class)	**combi-nation rules**	**gram. irregu-larities**	**phone-tics**	**etym. regist. region diachr. pragm.**	**syno-nyms**	**syntax &sem. classes** (also: ref. to identity number)	**ident. numb.**
alcatraz	=	<sm .orn>				AR (orn)	corvo-marinho		1741
alongar-	1	<vt>	AaiD	<g/gu>				<vr>	2133
along-	alongar	<var>	B					2133	
alongu-	alongar	<var>	Cc					2133	
aceso	=	<adj>			[ê]				487
abalável	=	<amf>	TP						46
acara	acará	<sm. ich>	R			TU (ich)		414	
abalôo	2oar	<v. PR 1S>						52	

Every lexicon entry consists of 10 fields (with translation information stored in separate lines ordered by semantic and valency-discriminators). Fields are separated by '#' and may be empty.

Word root is what the analysis program looks up after cutting inflexion endings and affixes off a word. A word root must be outward compatible with the word's other elements with regard to phonology, word class and combination rules.

Base form (and not word root) is what outputs as the base form of any derived reading. '=' means that it is identical to the word root, numbers mean removing the n last letters from the root form, letters are added to the root form. Thus '2oar' means: "cut 2 letters off 'abalôo', then add 'oar', in order to get the base form 'abaloar'.

Word class is used to determine outward compatibility, and is used to construe the output word classes N, V, ADJ, ADV from its first letters. For irregular word form entries, this field can contain inflexion information, e.g. 'abalôo': word class 'V' and inflexion state 'Present Tense 1st Person Singular'. Any syntactic or semantic information (like 't' for 'transitive' in 'vt', or 'prof' for 'profession') is not used on the tagger level. When used, at the disambiguation and syntactic levels, it is supplemented by the other possible syntactic or semantic classes (field 9).

Combination rules ("alternations") are idiosyncratic markings concerning outward compatibility with inflexion endings and the like. For instance, for verbs (which in Portuguese have hundreds of often superficially irregular inflexion forms) the following are used:

A combines with Infinitive (both non-personal and personal), Future and Conditional

a combines with Present Indicative forms with stressed inflexion ending (1. and 2. person plural), Imperative 2. Person Plural, and the regular participle endings.

i combines with "Past Tense" (Imperfeito)

D combines with "Present Perfect" (Perfeito simples), Past Perfect and Subjunctive Future Tense.

B combines with root-stressed forms where the initial inflexion ending letter is 'a' or 'o' (For the '-ar' conjugation Present Tense Indicative 1S, 2S, 3S, 3P and Imperative 2S, for the '-er' and '-ir' conjugation Present Tense Subjunctive 1S, 2S, 3S, 3P).

C combines with root-stressed forms where the initial inflexion ending letter is 'e' or 'i' (For the '-ar' conjugation Present Tense Subjunctive 1S, 2S, 3S, 3P, for the '-er' and '-ir' conjugation Present Tense Indicative 1S, 2S, 3S, 3P and Imperative 2S).

b combines with ending-stressed Present Tense Subjunctive forms (1P and 2P) of the '-er' and '-ir' conjugations.

c combines with ending-stressed Present Tense Subjunctive forms (1P and 2P) of the '-ar' conjugation.

Other word classes need fewer combination specifications, but an example is the TP for adjectives (meaning stress on the second last syllable, in opposition to TO for oxytonal stress), which for certain adjectives selects a particular plural ending ('-eis' for '-el' and '-il' adjectives).

Words with graphical accents often lose these in inflected or derived forms. They are therefore also alphabetised in the lexicon without accents, but combinationally marked R (prohibiting non-derived selection of the word root). This has also proved useful for correction of spelling, typing or ASCII errors in computerised texts, where accents may have been omitted or changed by either the author, typist or text transfer system.

Grammatical irregularities: This field contains information which has been used to design the irregular inflexion form entries in the lexicon, but since stem variations and irregular forms now all have their own entry, this field has been inactivated and is not read into active program memory on start up. Hard copy bilingual versions of the lexicon would, of course, make use of it.

Phonetics, too, are inactive in the PALMORF program. Any analytically relevant information from the field has been expressed as combination rules.

Field 7 contains so-called diasystematic information, lexicographically termed *diachronic* (e.g. archaisms or neologisms), *diatopic* (regional use), *diatechnical* (e.g. scientific or technical field), *diaevaluative* (pejorative or euphemistic) and *diaphatic* (formal, informal or slang). These diasystematic markers may be useful for disambiguation at a future stage, by means of selection restrictions and the like. Diaphatic speech level information, for instance, is being tentatively introduced: 'HV' (scientific "high level" term) can be used as an inward compatibility restriction for affixes; for instance, a Latin-Greek suffix like '-ologia' might be reserved for Latin-Greek word roots like 'cardio-' ("cardiology").

Synonyms are not used now, but might make selection restrictions "transferable" at a future stage.

Syntactic word class is specified throughout the lexicon, the main syntactic class being directly mapped from or incorporated into the primary (morphological) word class marking in field 3. Further classes eligible for the word root in question, are added here in field 9, as well as alternative semantic classes. Especially the valency structures and prepositional complementation of verbal roots generate many field 9 entries. Some examples are:

<vi>	intransitive verb
<vt>	monotransitive verb (with accusative object)
<PRP^vp>	transitive verb with preposition phrase argument (with the relevant preposition added as 'PRP^')
<x+GER>	auxiliary verb (with the non-finite verb form added, here '+Gerund')

Other word classes than verbs, too, can be marked for syntactic sub-class, for example:

<adj^+em> adjective that takes a prepositional complement headed by 'em'

Semantic subclassification is especially prominent for nouns:

<sm.orn> noun belonging to the 'bird' class of semantic prototypes

Identification number helps finding root entries, for example when cross-referencing to the translation file TRADLIST[10], or from an inflexion form entry to the relevant root entry. Only root entries have an identification number in this field, other entries have referring numbers in the second last field. The root word 'alongar-', for instance, has the identification number 2133 in field 10, and the word's other stem forms ('along-' and 'alongu-') refer to it in their number 9 fields.

[10] TRADLIST is compiled from the lexicon file, extracting all lines with translation equivalents, together with the relevant discriminators. At run time, TRADLIST is ordered by identification number.

2.2.3.2 The inflexional endings lexicon

(1)

<- - - - - INWARD COMPATIBILITY - - - - ->

1	**2**	**3**	**4**	**5**
inflexion ending	**base condition**	**word class condition**	**combination rules (alternation condition)**	**output**
iam	-	v	A	V COND 3P
iam	er-	v	i	V IMPF 3P IND
o	-	v	B	V PR 1S IND
as	o	a		ADJ F P
as	o	s	f:	N F P
eis	il	a	TP	ADJ M/F P

Inflexion ending is what the program cuts off the target word form, working backwards from the last letter.

Base condition is what the inflexion ending has to be substituted with before root search is undertaken. It is attached to the remaining word trunk, which then has to match one or more lexicon root forms.

Word class condition is then used to filter these possible root forms.

Combination rules are 1-letter-markings for verb stem class, stress pattern etc., that also appear with entries in the main lexicon. To match, the inflexion endings combination rule marker has to be part of the "allowing" string of combination rule markers in field 4 of the corresponding main lexicon root entry. E.g., the inflexion ending '-o' demands 'B' class of the combining verb root, and 'along-' allows it. Thus, 'alongo' is - correctly - analysed as 'V PR 1S IND', with the tag string taken from the field 5.

The **Output** field contains the tag string to be added to the active analysis line if a root is found that obeys all the relevant combination conditions. For non-verb word forms with a zero-morpheme-ending, the inflexion status is generated directly by the program, since checking for whole word lexeme entries constitutes the first step of inflexion analysis. Thus, if not marked otherwise, noun entries in the main lexicon are all classified 'singular'. Similarly, adjectives in root entry form are presented as 'male singular'.

In all, there are some 220 inflexion endings in the lexicon, differing very much in frequency. Some verbal endings (2. person plural) almost never occur in Brazilian Portuguese, and some irregular plural forms (like '-ães' for certain '-ão' nouns) are so rare, that it is a matter of lexicographer's choice whether to use individual inflexion form entries in the main lexicon instead, - both solutions are equally efficient.

There is quite a lot of homonymy among inflexion endings: '-a', for instance, occurs 8 times in the lexicon, covering 10 inflexional types. However, - due to different "inward compatibility" conditions - never more than two of these can attach to the same stem.

(2)

a,	-,	v,	D,	V MQP 1/3S IND VFIN,
a,	-,	v-,	B,	V PR 1/3S SUBJ VFIN,
a,	-,	var,	B,	V IMP 2S VFIN,
a,	-,	var,	B,	V PR 3S IND VFIN,
a,	e,	s,	f:-a,	N F S,
a,	o,	adj,	,	ADJ F S,
a,	o,	s,	f:-a,	N F S,
a,	o,	pc,	,	V PCP F S,

Empirically, one-root ambiguity is greatest for the unmodified infinitive ending 'r', where the number of competing readings, for most verbs, is brought up to 5 for one stem by the fact that the future subjunctive - in the 1. and 3. person singular - yields forms identical to the corresponding impersonal infinitive forms. Only some irregular verbs have different stems for the Infinitive (condition A) and the Future Subjunctive (condition D), respectively.

(3)

r,	r-,	v,	A,	V INF 0/1/3S,
r,	r-,	v,	D,	V FUT 1/3S SUBJ VFIN,

Note that the practical ambiguity handed down to the disambiguation module in the form of different tag lines, has been reduced both in (2) and (3) by the introduction of so-called Portmanteau-tags (1/3S and 0/1/3S)[11]. Since the subject in Portuguese clauses is optional or, rather, can be incorporated in the finite verb's inflexion ending, I have chosen to fuse the verbal 1. and 3. Person Singular where they can't be distinguished morphologically, i.e. for the Mais-que-perfeito tense, the Infinitive and Future Subjunctive, and, for the '-er'- and '-ir'-conjugations, also the Present Subjunctive. Another argument in favour of this choice is the fact that the 1. Person Singular is all but absent in many text types (typically those without speech quotes).

Of course, if a word form is ambiguous, and can also be derived from some other root by adding a non-zero-morpheme ending, this alternative will be found, too, - in the subsequent steps of the inflexion ending module.

[11] Portmanteau-tags are also used in the English Constraint Grammar of Karlsson et al. (1995). Here, the categories of person and number in verbs are untagged for most Past tense forms, and fused as -SG3 ('all but 3.Person Singular) or -SG1/3 ('all but 1. or 3.Person Singular) for most Present tense forms.

2.2.3.3 The suffix lexicon

(1)

<- INWARD COMPATIBILITY -> <-OUTWARD COMP.>

1	2	3	4	5	6	7
suffix form	**base** condition	**word class** condition	**combi-nation rules**	**output** derivation class and semantics	**suffix word class**	**suffix combina-tion rules**
-ista	V	asnb		DERS -ista	smf	
-ico	VTP	sn		DERS -ico [ATTR]	adj	
-otico	ose	s		DERS -ico [ATTR]	adj	R
-ei-	V	saA	a- anti- de- des- ...	DERS -ear [CAUSE]	var	BC

Suffix form is what the program's suffix module cuts off the word or word trunk it receives as input. One suffix can appear in the suffix lexicon in several disguises (for example '-inho' and '-zinho'), that are linked by the 'DERS' information in field 5 (it reads '-inho' even if the look up form is '-zinho'). Like inflexion endings, suffixes alphabetised in reverse order, because the search mechanism works backwards from the last letter.

Base condition is either a string, that is added before root search, or contains other orthographic combination information, like for example 'V', meaning that the suffix attaches to a root by substituting for any last letter vowel: thus various vowels are "tried out" when searching for a compatible root. The program must also provide for phonetic spelling changes at the root-suffix "interface". This is a very complex task, especially when a front vowel suffix (beginning with an 'e' or 'i' attaches to a root word ending in 'a', 'o' or 'u', or vice versa. Some neighbouring consonants will vary in these surroundings in order to keep their phonetic value:

(2)

spelling before back vowel	*spelling before front vowel*
ç	c
c	qu
g	gu
j	g

Also, diphthongs sometimes are substituted as one-vowel-units, sometimes they receive hiatus and accentuation of the second part, according to the stress pattern of that particular derivation or inflexion form (e.g. 'europeu' + '-izar' -> 'europeizar' [INF] and 'europeíza' [PR 3S]).

TP means paroxytony change: eligible roots need not have the same accentuation as the suffixed word form.

Word class condition is a list of all word classes eligible as roots for this suffix; each letter stands for a word class. Thus '-ico' attaches to nouns (s) and names (n).

Combination rules: Since the root-suffix interface is *inside* the word stem, the usual (inflexion based) combination class information in the main lexicon is of no great use. So far, I have only used few such rules (apart from word class and phonetic spelling, of course, which are treated in field 2 and 3). One rule says that certain short verbal suffixes may only be attached, if certain prefixes are present in the word form (cp. the '-ear' suffix in its '-ei' inflexion form, or '-ar' in 3b), in order to avoid over-generation.

```
(3a)   superamiga
         "amiga"  <title> <DERP super- [SUP]> N F S

(3b)   desamiga
         "amiga"  <DERP des-> N F S
         "amigo"  <DERP des-><DERS -ar [CAUSE]>  V IMP 2S VFIN
         "amigo"  <DERP des-><DERS -ar [CAUSE]>  V PR 3S IND VFIN
         "amiga"  <DERP des-><DERS -ar [CAUSE]>  V IMP 2S VFIN ###
         "amiga"  <DERP des-><DERS -ar [CAUSE]>  V PR 3S IND VFIN ###
```

While *desamiga* yields 4 verbal readings based on the “causative”[12] suffix '-ar', *superamiga* does not, because 'des-' is regarded as a "causative-combinable" prefix, and 'super-' is not. This is what I in the following will refer to as a "semantic circumfix condition". Note that 2 of the 4 verbal readings are marked for local disambiguation by ###, since there is no difference in tags, but only in base form ('amigo' vs. 'amiga').

Another possible field for suffix combination rules is register information like HV (high level scientific root).

Output is what appears in the analysis string: the derivative morpheme in its base form, often followed by some semantic class marker. Apart from 'CAUSE' for causative derivation in verbs, which is a combination condition for some prefixes ('a-' and 'es-'), these are not used by the program yet.

Suffix word class is the suffix' own inherent word class, which will be transferred to the word root it forms, and must be checked for outward compatibility with any "outer layer" suffixes or inflexion endings. The outermost suffix thus determines the final word class for the analysed text word.

[12] Meaning “cause to be”, “make”, “turn into”.

Suffix combination rules is also used for outward compatibility checks, especially before verbal inflexion endings (A,a,i,D,B,b,C,c). 'R' marks 'root only' forms, mostly in unaccentuated root variants of accent-bearing words (or suffixes, of course).

Below a commented list of suffixation examples is given. The suffixes in (4) are typical word class changing suffixes, changing a verbal root into a noun (4a), adjectives into nouns (4b) or place names into adjectives, while diminutives (DIM), augmentatives (AU) and superlative suffixes (SUP) are word class "transparent" (5). In (4d) the word class change is also inflexional, since deadjectival adverb derivation is treated in the inflexion lexicon. (5a) and (5b) are among the most productive suffixes in Portuguese. Loan words, of course, usually resist meaningful derivation by *Portuguese* morphological rules, but sometimes shared etymology of affixation elements allows derivative analyses even here. (5c) is such an example of a lucky hit where loan word structure and native derivative intuition coincide.

(4a) pesquisador
"pesquisar" <DERS -or [AGENT/INSTR/ACTLOC]> N M S
(4b) rotundidade
"rotunda" <DERS -idade [ABSTR]> N F S
(4c) pernambucano
"Pernambuco" <DERS -ano [PATR]> ADJ M S
(4d) temperamentalmente
"temperamento" <DERS -al [ATTR]> ADV

(5a) fetozinho
"feto" <DERS -inho [DIM]> N M S
(5b) rapidíssima
"rápido" <DERS -íssimo [SUP]> ADJ F S
(5c) disquete
"disco" <DERS -ete [DIM]> N M S

Of course, the international Latin-Greek "terminological" suffixes are productive in Portuguese, too, both in the scientific or pseudoscientific register (6a, 6b), and in everyday language, like in the political terms in (6c) and (6d).

Some suffixes, like '-ês' in (6e), seem, when used productively, to be characteristic of a certain genre, or usage, like for instance - in this case - "journalese" word games.

(6a) discografia
"disco" <DERS -grafia [HV]> N F S
(6b) jazzófilos
"jazz" <DERS -filo [DIM]> ADJ M P
(6c) presidencialista
"presidencial" <DERS -ista [ADEPTO]> N M/F S
(6d) federalização
"federal" <DERS -ização [CAUSE]> N F S

(6e) politiquês
"política" <DERS -ês [SPEECH]> N M S
"político" <DERS -ês [SPEECH]> N M S ###

The Portuguese successor of the Latin present participle ending, '-nte', does not have the broad open class productivity of an inflexion morpheme, and is therefore best termed a suffix in modern Portuguese. Also, '-nte' words have in many instances become lexicalised (i.e. dictionary-listed) nouns, suggesting that the original, attributive, participle reading is not really "alive" the same way, say, the past participle, '-do', is (with its full productivity for all verbs and its broad attributive potential). Thus, 'presidente', for example, must be regarded as a full noun, rather than a participle, since it can't even be used as an adjective any longer.

(7) galopante
" galopar" <DERS -ante [PART.PR]> ADJ M/F S
"galopar" <DERS -ante [AGENT]> N M/F S

Of course, more than one suffix may occur in the same word form:

(8a) halterofilistas
"haltere" <DERS -filia [HV]> <DERS -ista [ADEPTO]> N M/F P
(8b) peemedebistas
"P" <DERS M> <DERS D> <DERS B> <DERS -ista [ADEPTO]> N M/F P
(8c) percussionistas
"percussão" <DERS -ion [GEN]> <DERS -ista [ADEPTO]> N M/F P
"percussão" <DERS -ar [ACTION]> <DERS -ista [ADEPTO]> N M/F P
(8d) viabilizou
"viável" <DERS -bil> <DERS -izar [CAUSE]> V PS 3S IND VFIN

In (8b), multiple suffixation analysis is used as a technique to tackle productive derivation in abbreviations. The real root here is 'PMDB', a political party. The mechanism is described in detail in chapter 2.2.4.1.

Multiple suffixation analysis can also be a solution for "naturalised" loan words, like in (8c), or for capturing Latin-based etymological stem alternations, as in (8d).

2.2.3.4 The prefix lexicon

(1)

1	**2**	**3**	**4**	**5**
prefix form	**base condition**	**word class condition**	**combination rules**	**output** derivation class and semantics
a	C	asv	[CAUSE]	DERP a-[STATE]
a	C	a		DERP a- [ANT]

an	V	a		DERP a- [ANT]
brad	Vi	as	HV	DERP bradi-
psic	Vo	as		DERP psico-
psiqu	ei	as		DERP psico-
re		sv		DERP re [AG+]
mini		s		DERP mini- [DIM]

Prefix form is what the program's prefix module cuts off a word form it receives as input. Like suffixes, prefixes can come in several disguises, depending on the spelling context. Also, homonyms - with different combinatorial behaviour and semantics - do exist. Thus, 'a-' can be both an antonymy-prefix co-varying with 'an-' (one before vowels, the other before consonants) and a STATE-prefix, that combines in a kind of "circumfix" construction with causative suffixes like '-izar'.

Base condition controls which root-initial letters a prefix can combine with: it may read V for vowel, C for consonant, or just something like 'lmn' for the individual letters 'l', 'm' and 'n'. Any letter x after the 'V' can be added to the prefix form, when searching for roots without an initial vowel (x may be called the standard ligation vowel for this particular prefix). Thus, both 'psic|análise' and 'psico|drama' can be found. Root initial doubling of 's' and 'r' after a prefix-vowel (which preserves the [s]- and [R]- sounds) is not listed as a base condition in the prefix lexicon, but treated directly in the program's main module (inflexional analysis) when called from the prefix module: 'mini-ssaia' (root 'saia').

Word class condition must be matched by either the root's word class or - if any - by the words outermost suffix. Prefixes need only inward[13] compatibility conditions, since they do not by themselves influence a derived word's word class, so no information comparable to field 6 and 7 in the suffix lexicon is found here.

```
(2)                      _ _ _ _ _ _ _ _ ____________
                      |       |       |            |  (combinat. rules for inflexion endings)
                   #      #       #        #
        prefix ((( root ) suffix ) suffix ...) inflexion ending
           |______________|
           |______________________|
        (semantic "circumfix" conditions)
```

In the above expression, word class compatibility is checked along brackets, with "inward" and "outward" defined by the bracket's convexity orientation. Further combination rules apply between inflexion endings and the root or the last suffix.

[13] By *Inward compatibility* I understand word class or inflexion class compatibility with what the affix in question is attached to (i.e. a root or another affix closer to the root than itself), while *outward compatibility* is about what further/other affixes or endings may be attached to it, in the form of yet another onion layer - on top of the affix in question. This way, the use of a prefix may be conditioned not only by the root, but by another prefix, its inward neighbour in the affix segmentation chain, - and likewise, the use of a suffix may be conditioned by other (inwardly neighbouring) suffixes.

Phonetic-orthographic compatibility is checked at each derivation element border (marked #)

Combination rules for prefixes, apart from those mentioned in field 2 and 3, are few and semantic in nature. Examples are the 'es-' and 'a-' prefixes that demand causation suffixes, and register conditions for the root lexicon (like 'H' for 'high register language, not implemented in the present version of the analyser).

Output is what appears in the final analysis string, containing the standard form of the prefix (derivation class) and - so far unutilised - semantic information on that prefix, like 'DIM' (diminutive) or 'ANT' (antonym).

Most prefixes modify nominals (5 and 6), usually both adjectives and nouns, though some ('an-') prefer adjectives and some only attach to nouns ('mini-', 'maxi-', 'vice-'). With the possible exception of 'anti-' *(anticristo)* , none modifies proper nouns - unless these have been turned into ordinary nominals first, by '-ista'-suffixation, for instance. Pre-verbal prefixes (9) are often prepositional ('a-', 'des-', 'com-', 'sobre-', 'trans-'), denoting movement or change. The typical pattern is a circumfix-construction:

(4)

PRP +	adjective/noun +	CAUSATIVE
des-	sacral	-izar
con-	firm(e)	-ar

Of course, in many cases the causative is already lexicalised in a fixed way, and makes only etymological sense, like in (7), where a double analysis is found, one with the prepositional prefix frozen into the stem (the "participle" *compacto*), one with the causative suffix incorporated in the root ('pactuar'). In (9), both the analytical stem *(sacral)* and the lexicalised causative *(sacralizar)* are present in the lexicon.

Obviously, the root found in the lexicon may also be a nominalised form of the causative (for instance, *sacralização*), and therefore it is safest also to allow nominal stems for the prepositional prefixes.

The examples below are ordered by complexity. In (5) we find classical, syllabic prefixes, (5b) demonstrating the word class transparency of prefixes in general. The prefixes in (6) are semantically heavier, more words than syllables, typical of the international Esperanto of science where both prefixes, stems and suffixes are Latin-Greek elements, with word-like prefixes often substituting for root-compounding. The same element (for example *'gastr'* - "stomach") may appear in both root position (*'gastr-ite'* - "gastritis") and affix position *('gastro-grafia'* - "gastrography").

(5a) antimonogâmica
"monogâmico" <DERP anti- [ANT]> ADJ F S
(5b) arquiinimigos

"inimigo" <DERP arqui- [SUP]> ADJ M P
"inimigo" <DERP arqui- [SUP]> N M P

(6a) microprocessadores
"processador" <DERP micro- [DIM]> N M P
(6b) hidrelétrica
"elétrico" <DERP hidro-> ADJ F S
(6c) neuropsicóloga
"psicólogo" <DERP neuro-> N F S

In (7) both a prefixed and a suffixed analysis are found, and in (8) and (9) prefixes and suffixes are even present in the same reading. *Dessacralização* in (9) shows the phonetic interface rules at work, the s-doubling being necessary in order to retain the unvoiced [s] from the word-initial position in *sacral*. Also 'com-' in (7) exists in several phonetic variants (another is *con*-), 'com-' being used before 'p', 'b' and 'm'.

(7) compactuar
"compacto" <DERS -uar [CAUSE]> V INF 0/1/3S
"compacto" <DERS -uar [CAUSE]> V FUT 1/3S SUBJ
"pactuar" <DERP com-> V INF 0/1/3S
"pactuar" <DERP com-> V FUT 1/3S SUBJ

(8a) superfaturamento
"faturar" <DERP super- [SUP]> <DERS -mento [CAUS]> N M S
(8b) biodegradável
"degradar" <DERP bio-> <DERS -vel [POTENTIAL]> ADJ M/F S

(9) dessacralização
"sacralizar" <DERP de-> <DERS -ção [CAUSE]> N F S
"sacral" <DERP de-> <DERS -ização [CAUSE]> N F S

2.2.4 The dynamic lexicon

2.2.4.1 Polylexical expressions

It is useful to identify polylexical expressions of any frequency early in the analysis process, both in order to avoid unnecessary ambiguity of its element words and because the resulting complex word class may be better suited to a syntactic analysis than the individual word would.

Some structures are, of course, hyphenated and thus easily recognised. In the lexicon, these are tagged for word class and, if necessary, their complex inflexion patterns. <P12/P2>, for instance, means that a hyphenated noun or adjective with two elements, receives plural endings on both its elements, or, optionally, only on the second. Apart from pronominal and inflexional enclitics (cp. chapter 2.2.4.2), the elements of hyphenated word forms are first analysed individually by the tagger. This is necessary in order to recognise inflexion morphemes on the individual elements of a hyphenated word. In the next step, if the combined base forms are found in the lexicon, or if the preprocessor recognises the polylexical as foreign language material[14], the word is reassembled and passed on to the next parsing level as a whole (tagged with a summary word class tag, but also marked <_c> for "composite"). Otherwise, the hyphenated polylexical will be split into words bearing their own tag string and a <hyphen> tag. The parser will then assign a syntactic structure[15] to the word, like N @NPHR and - ADJ @N< for the first and second parts, respectively, of *corvo-marinho* (a bird species).

A special case are hyphenated prefixes, like in *anti-constitucional*. Here, *anti-*, since it isn't morphologically fused with the root, is not a "real" prefix in grapho-morphological terms, and can be assigned its own word class and function tags: EC @PREF ("elemento composto" functioning as prefix). In the newest version of the parser (1999), hyphenated prefixes are treated as individual (EC-) words on the morphological level only[16]. For syntactic analysis, EC-elements are re-fused onto their “head”, and the resulting compound treated as one syntactic unit. Since prefixes, unlike suffixes, do not usually have any influence on a word’s word-class, it makes sense to let EC-compounds inherit PoS and inflexion tags from the non-EC element. *Anti-constitucional* will thus become an adjective, *anti-soneto, contra-indicação, contra-reforma, contra-cheque* and *contra-revolução* will become nouns. Incidentally, *anti-* is the only “hyphenatable” prefix, where this word class inheritance strategy is not universally successful, as the following examples show, where ‘anti-’ prefixes a noun, but the resulting *function* is rather adjectival:

lei anti-resgate
comportamento anti-social
política anti-semita
protesto anti-racismo
sentido anti-horário

As a compromise solution, in these cases, the parser will still tag *form* as N (noun), but *function* as adjectival/attributive (@N<).

Non-hyphenated polylexicals are treated in the following way: non-varying expressions are marked in the lexicon by '=' between words, expressions containing

[14] Indications for foreign loan word status are, among other things, certain non-Portuguese letters or letter-combinations. In particular, Portuguese has no ‘y’ or ‘w’, does not allow gemination of letters other than ‘rr’ and ‘ss’, and is very restrictive as to which letters can *end* a word.

[15] Both here and in the case of hyphenated prefixes, one could argue that ("syntactic") function categories are introduced at the sub-word level. Since the distinctions made are subject to the same disambiguation procedures as word- or sentence-level analyses, this is yet another example of progressive level parsing, where the same tools are used on different levels, in order to incrementally achieve a more and more fine grained analysis.

[16] Since usage isn't stable with regard to hyphenation, it is paramount that the parser be able to assign meaningful analyses for both variants of the fusion/hyphenation dimorphia of prefixes, as well as handle inconstant hyphenation in "hyphenatable" polylexicals. A proposed orthographical reform in Brazil would abolish much hyphenisation, yet define a list of prefixes where hyphenation is mandatory, probably increasing the overall inter-individual variance of usage for a few decades ...The EC-tag is optional in the parser’s output, and presently (1999) remains invisible, since the hyphenated prefixes ar reattached by a filter-program at the syntactic output level.

variable word forms receive '_'- linking, with '%' after element words that can be inflected. In this last group (with variable non-hyphenated elements) one can find idiomatic expressions and even proverbs. Since these are mostly of semantic interest, the program - for now - ignores them, checking only for non-varying expressions, with the important exception of incorporating verb structures and plurals of complex nouns or adjectives. Also, with respect to proverbs and clause-level idioms, it seems to be more interesting for a parser to assign syntactic structure in an analytical way than to provide a summary treatment in a synthetic way.[17]

It is the preprocessor which has to identify and '=' - mark polylexical strings. Technically this is done by adding up running words to a potential polylexical string, until a maximum (at present: 4) is reached, or punctuation gets in the way, whichever happens sooner. This is more difficult than it sounds, - a *'.."WORD..'* -structure, e.g., breaks a running string, but is allowed string-initially, whereas *'..WORD,..'* becomes part of the string, but breaks it nevertheless, losing its *','* . When a string reaches maximum, the following happens:

a) polylexical search with negative result:

A group of 4 words is checked (in a left bounded fashion) first for long, then shorter polylexicals. If none is found, the 4-word window moves one word to the right, and the search process is repeated.

[17] Another matter is, of course, machine translation, where "synthetical treatment" is preferable and necessary for assigning an idiomatic translation.

```
        WORD1       WORD2       WORD3       WORD4       WORD5       ......
step 1  |______________________________________________|
step 2  |__________________________________|
step 3  |______________________|
step 4  *           |______________________________________________|
step 5              |__________________________________|
step 6              |______________________|
step 7              **          |____________________________________ _ _ _ _ _ _|
```

**   WORD1 is sent to single word processing.*
*** WORD2 is sent to single word processing.*

b) polylexical search with positive result (xxx):

If a polylexical is found, the 4-word window is reset with the new WORD1 immediately after the polylexical found.

```
        WORD1       WORD2       WORD3       WORD4       WORD5       ......
step 1  |______________________________________________|
step 2  |___________________________________|
step 3  |XXXXXXXXXXXXXXXX|
step 4                          |___________________________________ _ _ _ _ _ _|
step 5                          |__________________________________|
step 6                          |_______________________|
```

Broken strings are "finished" before progressing ...:

In this case, the flow of words is "broken" by punctuation, and a group of 4 words is isolated by a comma which can't be bridged by a polylexical string. So, all combinations up to the comma are tried before admitting WORD5 to the search string.

```
        WORD1       WORD2       WORD3       WORD4,      WORD5       ......
step 1  |_______________________________________________|
step 2  |___________________________________|
step 3  |_______________________|
step 4  *           |___________________________________|
step 5              |_______________________|
step 6              **          |_______________________|
step 7                          ***                     |_________________ _ _ _ _ _ _ _
|
```

**** WORD3 and WORD4 are sent to single word preprocessing.*

.... or, if a 2-word polylexical is found (xxx):

The remaining 2 words of the 4-word pre-comma group are checked first, before progressing, then, the search window is reset to after the break (comma).

```
        WORD1      WORD2      WORD3      WORD4,     WORD5      ......
step 1  |____________________________________________|
step 2  |_________________________________|
step 3  |XXXXXXXXXXXXXXX| |________________|
step 4                                               |________________ _ _ _ _ _ _
|
```

The overlapping search is clearly necessary to find all possible combinations: without punctuation breaks, n words may form n*(m-1) combinations of up to m elements. With a depth of 4 this amounts to 3000 possible polylexicals for a 1000 word text.

It is crucial to begin with the longest string and then work backwards, one might otherwise miss 3- or 4-word polylexicals, that "contain" smaller ones. E.g., in Portuguese, 'dentro=em' *(inside)* is a complex preposition, 'dentro=em=breve' *(before long)* a complex adverb. In searching from left to right one would miss out on the (longer) adverb reading, because 'dentro=em' is found first, and the search string reset to start from scratch at position 3.

2.2.4.2 Word or morpheme: enclitic pronouns

Generally, in inflecting languages like Portuguese, future tense endings are regarded as bound morphemes, whereas pronouns are classified as (free morpheme) words. However, making things less easy for the preprocessor, Portuguese allows both to appear as hyphenated "linked" morphemes, too. Consider the following examples:

(1a) O comprei amanhã. (I'll buy it tomorrow.)
(1b) Comprá-lo-ei amanhã.
(2a) Não o pode fazer. (He can't do it.)
(2b) Não pode fazê-lo.
(3a) O tinham visto. (They had seen him.)
(3b) Tinham-no visto.
(4) Chove. (It rains.)

In (1b) the direct object pronoun 'o'/'lo' is placed mesoclitically, *before* the future tense inflexion ending, which thus becomes enclitic. The preprocessor has to recognise this structure and transform it into a canonical form, which the word-based tagger can understand:

(1c) *Comprei- o amanhã.

As can be seen in (2) and (3) both the stem and the enclitic pronoun undergo phonetically motivated changes, the infinitive loosing its 'r' and receiving a stress accent, and the pronoun ‘o’ changing into 'lo' or 'no' depending on the preceding sound. While this is more difficult than the recognition of simple strings of adjacent words as polylexicals, it can become even more computing intensive to figure out whether the form 'xxxá-lo' has to be canonised into 'xxxar- o' (infinitive) or 'xxxaz- o' (irregular present tense 3rd person singular). The latter case is a morphological ambiguity, which can only be resolved by consulting the core lexicon - something a preprocessor isn't normally supposed to do.[18]

Another, more syntactic, puzzle in the cited examples - at least from an English point of view - is the missing subject. A strict generative rule for sentence analysis, like 'S -> NP VP', wouldn't work here. The subject is, in fact, represented by a bound morpheme: -'ei' (I), '-am' (they) or '-e' (he, it)[19]. This is one of the reasons why I prefer to analyse a Portuguese sentence not as a binary entity consisting of subject and predicate, or NP and VP, but as one big set of dependencies around a verbal nucleus, with the subject being read as a facultative (valency bound) argument of the verbal constituant. In (4) the subject argument, not being part of the verb's valency pattern, is altogether missing, - it can not be expressed as an independent word.[20]

2.2.4.3 The *petebista*-problem: productive abbreviations

Abbreviations have never been easy to recognize, neither for foreigners nor for parsers: there are new abbreviations all the time, names of organisations, products, new diseases, pharmaceuticals and others. Their morphology incorporates signs like '.', '-' and '/', making it difficult to decide what is a sentence delimiter and what is part of an abbreviation. Also, abbreviations can mimic other word classes, especially nouns, with gender category or even number inflexion.

But in (Brazilian) Portuguese newspaper and social science texts, they really come alive! For example, the names of political parties or interest groups, of which there are quite a few in Brazil, may have their abbreviations phoneticised letter by letter. Thus 'PTB' (a Brazilian Workers’ Party) reads 'pe-te-be', which becomes a new word root in its own right. Like many nouns and names, it may be suffixed with '-ista', '-ismo' and others. To make things even more complicated, letter names may

[18] In the PALAVRAS system, the preprocessor can access the main lexicon, both for this particular task and for others, - like polylexical identification, or for checking verbal incorporation patterns.

[19] This "subject pronoun inflexion morpheme" appears at the head verb of the sentence' verb chain, i.e. on the first auxiliary, if there is one, or else on the main verb. In Portuguese this holds even if this verb is not a finite form, but an infinitive. If the subject is (also) expressed as an independent word or group, there has to be agreement between the overt subject and the "enclitic inflexion ending subject".

[20] The above also precludes a view defining clauses as structures containing more than one word. Portuguese utterances like (4) are clearly sentences, and imperatives are an example that works for both Portuguese and English. Here, one must either accept one-word sentences or redefine the notion of 'word'. Is a word to be a blank space surrounded string, a hyphen/blank space surrounded string, or can it include even fused enclitics that are morphologically indistinguishable from inflexion endings (cp. chapter 2.2.2.2 and 2.2.4.2) ? Alternatively, one could emphasize the special (syntactic) status of a one-word “syntax-less” utterance like imperatives by calling it a sentence that is not a *clause* (unlike ordinary clauses that feature some kind of clausal nexus). For a more detailed discussion of word- and clause-hood, see also the VISL manual “Portuguese Syntax” (Bick, 1999).

appear truncated or not, depending on phonetic harmony and vowel distribution: 'N' may become both '-ene-' or 'en-'.

To solve this puzzle, I introduced all letter names in their various forms into the suffix lexicon, with combination restrictions saying that they belong to the word class 'b' (abbreviation) and have inward compatibility only with other elements of the same type. Certain suffixes (like '-ista'), then, allow for left hand combination with these letter elements. Since letter names also appear in the root form lexicon, the program can now analyse party member expressions as long derivation chains of abbreviation letters (which, formally, stand for the party name word elements). 'petebista' is thus recognised as a Portuguese word, and reads in the analysis file:

P <DERS T><DERS B><DERS -ista> N M/F S

In the same way, other productive expressions phonetically derived from abbreviations, can now be tagged.

2.2.4.4 Names: problems with an immigrant society

In my system, I define the word class of proper nouns (lexicon entry 'n', PoS tag 'PROP') as capitalised words distinguished from nouns and adjectives by featuring both number (S/P) and gender (M/F) as lexeme categories, not word form categories.

(1)

LEXICON ENTRY	TAG SEQUENCE
Filipinas <nfP>	PROP F P
Dardanelos <nmP>	PROP M P
Estados=Unidos <nmP>	PROP M P
Amado <nmS>	PROP M S
Berlim <nmS>	PROP M S
Andrómeda <nfS>	PROP F S
OMS <b-nfS>	PROP F S
PC <b-nmS>	PROP M S

Presently, there are about 1.300 names in the lexicon, consisting of single word proper nouns, or lexicalised name chains[21], about 8% being abbreviations, with a male/female ratio of roughly 4:3 (this being about the same as for ordinary nouns). Since proper nouns, like ordinary nouns, can trigger agreement in verb chains ('A OMS foi lançada ...') or modifiers ('o grande Amado'), lexicon information is quite important for disambiguation. The word 'a', which - among other things - can be either a preposition of movement or a feminine article, can be disambiguated with the help of the neighbouring noun's gender information in the following example.

[21] I define a name chain as consisting of at least one proper noun followed by any number of non-clausal dependents (with capitalised nouns and adjectives) and/or (possibly capitalised) distinctors (like jr., VI), and preceded by any number of capitalised prenominals and/or (possibly capitalised) pre-name nouns (titles etc.).

(2a) A mãe foi a Berlim. (The mother went to Berlin.)
(2b) A mãe foi a Maria. (The mother was Maria.)

However, about 1-2% of all word forms in running text are (lexically) *unknown* names[22]. This percentage is so high that even without the help of the lexicon, the parser has to recognise the word forms in question at least in terms of their word class. The obvious heuristics is, of course, treating capitalised words as names. On its own, capitalisation is not a sufficient criterion, but in combination with foreign word heuristics and some knowledge about typical in-name inter-capitalisation elements (‘de’, ‘von’, ‘of’), the preprocessor can filter out at least some lexicon-wise unknown names, and fuse them into PROP polylexicals.[23]

Since the morphological analyser program itself looks at one word at a time, analyses it, and then writes all possible readings to the output file, it can only look "backwards" (by storing information about the preceding word's analysis)[24]. Here four[25] cases can be distinguished, the probability for the word being a proper noun being highest in the first case, and lowest in the last:

- 1. A capitalised word in running text, preceded by a another name (heuristic or not), certain classes of pre-name nouns (<title>, e.g. 'senhor', <+n>, e.g. 'restaurante', 'rua', '-ista'-words and others) or the preposition 'de' after another name
- 2. A capitalised word in running text, preceded by some ordinary lower case word
- 3. A capitalised word in running text, preceded only by other capitalised words (The headline case)
- 4. A sentence initial capitalised word[26]

Another distinction made by the tagger is based upon whether or not the word in question can also be given some other (non-name) analysis, and upon how complex this analysis would be, in terms of derivational depth. The name reading is safest if no known root can be found, and least probable where an alternative analysis can be found without any derivation. Readings where the word's root part is short[27] in

[22] The numbers given are an average across different text types. In individual news magazine texts (like VEJA), name frequency can actually be much higher.

[23] This feature of the preprocessor was only activated recently (1999), and the statistics and examples in this chapter apply to corpus data analysed *without* preprocessor name recognition.

[24] Even this minimal context sensitiveness is worth mentioning - TWOL-analysers, for instance, never look back at the preceding word.

[25] In an earlier version, cases 1 and 2 were fused, resulting in a somewhat stronger "name bias": because ordinary lower case words would count as pre-name words, too, most upper case words in mid-sentence would get <HEUR> PROP as one of their tags.

[26] The tagger assumes "Sentence initiality", if the last "word" is either a question mark, exclamation mark or a full stop not integrated into an abbreviation or ordinal numeral.

[27] To avoid overgeneration, a number of very short lexemes, like the names of letters (tê, zê), have a <nd> (no derivation) tag in the lexicon. These lexemes are completely prohibited for ordinary derivation, - though some also exist in a special, for-derivation-only, orthographic variant, like letter-names (te, ze) that may combine with each other to form productive "phonetic" abbreviations.

comparison to the substring consisting of its derivational morphemes and inflexion endings, are also regarded as less probable.

The following table shows in which cases the tagger will choose a (derived) lexical analysis, a (heuristic) proper noun analysis, or both:

Table: Name heuristics - decision table

Preceding context **Competing analysis**	***sentence-initial***	***after only capitalised words: "headline"***	***after lower case word***	***after name or pre-name noun***
underived, pre-name class 'Senhor'	lexical	lexical	lexical	lexical
underived, not pre-name class 'Concordo'	lexical	lexical	lexical (older version: lexical/PROP)	lexical/PROP
long root, derivational 'Palestr-inha'	lexical	lexical/PROP	lexical/PROP	lexical/PROP
short root, derivational 'Cas-ina'	lexical/PROP	lexical/PROP	lexical/PROP	lexical/PROP
none	PROP	PROP	PROP	PROP

Originally I worked with a very "soft" definition of a pre-name context (all words that are not capitalised plus lexical pre-name expressions, even if they are capitalised), and most capitalised words would get both the lexical and the name-heuristic analysis. This kind of cautiousness is typical for the parsing system, and exploits its "progressive level" characteristics - ambiguity not resolved on one level, will be treated with better tools on the next. In this case, context sensitive Constraint Grammar rules would do the job.

There is, however, a reason for excluding ordinary lower case words from the pre-name context, at least where the competing analysis is non-derivational (i.e. inherently probable): Compound names retain more of their internal structure in the analysis, if compound initial (capitalised) adjectives or pre-name nouns (titles etc.) are tagged as ADJ or N (3b), respectively, than in an all-name chain analysis (3a):

(3a) Escola PROP @NPHR Santa PROP @N< Cecília PROP @N<
(3b) Escola N @NPHR Santa ADJ @>N Cecília PROP @N<

The price for the more fine grained analysis in (3b) is the risk of the tagger's not handing a PROP analysis at all to the CG-disambiguation module in the case of isolated upper case words that have a clear (non-derived) alternative analysis, like in *Bárbara* and *Xavier*, which both are simple adjectives in the lexicon (with the meaning of 'barbaric' and 'annoying', respectively.

My present linguistic solution[28] is to opt for the more analytic description of compound names and to tag some critical words as both PROP and ADJ or N <u>in the lexicon</u>. Since only *underived* competing analyses pose a problem (derivationals also in the new system still receive a tag for the PROP alternative), the list of these names is quite short - a check on a 1.5-million word chunk of corpus yielded less than 150 *different* cases (which isn't much compared to the 2% overall frequency of names).

In the appendix section, a list of context sensitive CG disambiguation rules is given for the disambiguation of words which the analyser has assigned other PoS tags alongside the proper noun tag. Apart from specific rules, which explicitly target proper nouns, many other rules may contribute to resolving the ambiguity in an indirect, cautious way - by eliminating competing PoS readings one by one, leaving only the desired one.

An important contribution to the proper noun sub-section of CG-rules is the structural information that follows from the recognition of certain types of name chains, typical of Portuguese text:

(4a) Felipe Cruz Guimarães
(4b) o presidente Fernando Collor de Mello
a carioca Maria dos Santos
o senhor Aurélio Buarque de Holanda Ferreira
(4c) Hamilton Mello jr.
(4d) o crítico de gastronomia Celso Nucci

(5a) a Guia Quatro Rodas
o Grupo Rui Barreto
(5b) o restaurante Arroz-de-Hausa
(5c) a Grande São Paulo
(5d) Europa Oriental

(6a) a Drake Beam Morin
(6b) o Instituto para Reprodução Humana de Roma

(7a) Massachusetts Institute of Technology
(7b) Guns 'n' Roses
(7c) Michael's Friends

The personal names in (4) can all be described by the pattern:

(4') *(N <title/prof/n>) PROP+ (de/do/da/dos/das PROP+) (jr./sr./I),*

[28] In standard mode, the parser today (1999) draws upon a special filter program written to capture likely name chains in a heuristic way *after* preprocessing and *before* morphological analysis, linking name chain elements in the same way recognized polylexicals are (by '=' signs). Thus, the capitalised parts of most name chains (plus interfering 'de', 'dos', 'von', 'van' etc.) become one-word units to the eyes of PALMORF, ensuring PROP analysis, but hiding most of the analytical structure of the name unit.

where brackets mean optional constituents, and the '+' means one or more constituents of the same type. In the PROP+ chains I have chosen a "left leaning" dependency analysis treating the *first* proper noun as the head and all others as postnominals: @NP-HEAD @N< @N< ... A strong argument for this choice is the fact that it is the first proper noun (usually a person's Christian name) that determines the gender of the whole PROP chain. The same argument may be used in deciding on a head for the name chain as a whole. Here, the leftward orientation continues, since - if there is one -, the leading pre-name noun (a title, for instance) will pass its gender and number features on to the name chain as a whole. Consider the agreement evidence in: *os senhores Smith são ric**os*** (plural), *a rainha Smith é ric**a*** (Queen Smith, feminine), or even (in a kindergarten role play) **a rainha George é bel**a*** (feminine?). Of course, in many cases title and name have the same gender and number anyway, or a gender ambiguous title like *presidente* may even draw its gender feature from the following name. In a constructed, conflicting case, however, the title "wins" the semantic struggle where surface marking is forced, like in the example of subject complement agreement (*rainha George é bel**a***) - though I must admit that I have yet to find a "real" corpus example.

Stress patterns in spoken Portuguese, English and Danish also support a "left leaning" analysis: One would expect the modifying ('special') piece of information to be stress-focused, as is indeed the case in "The White House", "Kennedy jr.", "King George", which implicitly answer the question "which house?", "which Kennedy?", "which king?". Finally, the modifier character of surnames is strengthened by the fact that surnames are often derived from patronyms, toponyms or profession terms, likewise specifying which of a number of bearers of the same Christian name is targeted: "Peter Johnson/Sørensen", "Peter Bloomfield/Sprogø", "Peter Miller/Møller".

In some languages, Portuguese included, PPs are used to form surnames (cp. 'de', 'of', 'von', 'zu', 'van' etc. in the European melting pot), clearly suggesting modifier etymology, and I will therefore treat recognisable prepositional groups in name chains accordingly - i.e. as postnominal modifiers (cp. 4') - adding more meat to the left leaning structural analysis. At the same time, the internal structure of the PP is retained, i.e. the (first) name inside the PP is tagged as argument of preposition (@P<).

Terms like 'jr.', 'sr.' and the Roman numerals, finally, are lexically marked as post-positioned attributives, which also translates into a @N< function. If there is a preceding pre-name expression, like a title ('senhor'), a professional function ('presidente') or an "ethnicity term" ('carioca'), then the whole name chain will be regarded as a postnominal itself, the first proper noun in the chain bearing the @N< tag that points to the pre-name noun. Sometimes the pre-name NP can be quite complex, too - cp. (4d), where the interfering postnominal PP 'de gastronomia' makes it difficult (in terms of rule number and complexity) for the CG-rules to "see" the link between 'crítico' and the name chain.

In some non-personal proper nouns, however, the pre-name term may be capitalised (5a), suggesting a PROP reading. The rules concerned had to learn the difference between pre-name terms that apply to persons (e.g. *senhor*, *carioca*, typically lower case) and those that don't (e.g. *Rua*, *Grupo*, often upper case), thus ignoring the upper case letter in the pre-name term and retaining its N reading, but still assigning the same overall function pattern (i.e. postnominal function for all the proper nouns in the chain, and nominal head function for the first word in the chain). With the new, "harder" (i.e. less ambiguous already at the analyser level), name tagging protocol, this case has become a lot easier, in terms of CG rule economy[29], since for simple (underived) name chain initial nouns no PROP reading is generated in the first place (i.e. on the lexical analyser's level).

In (5b) recognition of the pre-name term is easy (since it isn't capitalised), whereas the hyphens in the name term have to be recognised by the preprocessor as inter-word rather than intra-word, in order to make it possible for the parser to assign the correct structure (the same as in 4b).

(5c), finally, is different in that the first word of the expression is marked as part of the name structure by capitalisation, but could - internally - be described as a prenominal attributive. In the old version, the lexical analyser establishes a word class ambiguity between PROP and ADJ, which is then resolved in favour of the PROP reading by the CG-rules, sacrificing the attributive reading, but gaining name phrase continuity analogous to (4a)[30]. In the new version, in the case of a non-derivational (simple) ADJ reading, no PROP reading is added (and thus no disambiguation necessary). Here, the prenominal function will be recognised, but the name chain continuity (expressed by the capitalisation of the adjective) is less explicit.

Name chain *final* (capitalised) adjectives, as in (5d), are another matter - first, already on the tagger level, a *backward* look is possible, so (unlike in the chain initial adjective case in 5c) the tagger has a strong reason to make *'Oriental'* part of the name by adding a PROP tag, and, second, the postnominal @N< function tag works for both the PROP and ADJ classes, so it is not (as in the chain initial ADJ case) necessary to sacrifice the "part-of-the-name-ness" (expressed by the PROP tag) in order to achieve a structurally accurate description.

(6a) is the prototypical case of a (foreign) firm name - a colourful string of multinational names without immediately recognisable internal structure and usually without any lexicalised proper noun anywhere in the chain. Firm names are nearly

[29] meaning either fewer rules needed to achieve the same result, - or a better result achieved with the same number of rules.

[30] It is admittedly hard to make this choice. My general approach is to regard name chains as "leaning left", i.e. having their head in the leftmost capitalized word. This is why premodifiers of names must either (if lower case) stay outside the name chain proper (like the article in 'a Maria Moura') or (if upper case) become head of the name chain. Of course, <title> type nouns "tolerate" this treatment much better than adjectives, i.e. their chain internal function is described more adequately. On the other hand, it is very hard to ascertain how long an etymological adjective retains is adjectivity inside a name chain: Is 'pacific/atlantic' in 'The Pacific/Atlantic Ocean' still an adjective? Why, then, is it possible to substitute 'The Pacific/Atlantic' for the whole chain? Why does 'Ocean' get stress marking, and not the modifier 'Pacific/Atlantic'? My present choice is to treat some fixed expressions ("Pacific=Ocean") as single lexical units in the PALAVRAS lexicon, and to opt for the prenominal adjective reading in all the others.

always treated as feminine in Portuguese, which might be exploited heuristically in this case. Otherwise the chain is treated as in (4a), as a kind of analytical default. (6b) is an example for the prototypical institution name, which usually boasts much more internal structure. In fact, due to the scarceness of contiguous, "(6a) - type", potential nominal heads (which would favour the name reading over the noun reading, since the first allows for @NPHR @N< @N< ... chains[31]), the word class distinction between noun and name does not structurally make any difference in this case:

(8)	'o'	DET	@>N
	'Instituto'	**N/PROP**	@NPHR
	'para'	PRP	@N<
	'Reprodução'	**N/PROP**	@P<
	'Humana'	**ADJ/PROP**	@N<
	'de'	PRP	@N<
	'Roma'	**N/PROP**	@P<

The worst case scenario (7) are foreign language name chains containing syntactically important particles or content words with lower case first letters. As long as all words in the chain are capitalised, an approximate analysis can be obtained by assigning the PROP word class to all members of the chain allowing for a functional structure like in (4a). The examples in (7), however, contain the particles 'n', 'of' and the apostrophed inflexion morpheme "s' in lower case letters. The only easy solution to this problem is to enter the most frequent of those (English) particles into the (Portuguese!) lexicon. Thus, 'of' (as well as Dutch 'van' and German 'von') is listed as PRP, and -*'s* as <genitive> PROP M/F S. Thus, (7a) gets a fair internal analysis[32] (Massachusetts @NPHR Institute @N< of @N< Technology @P<), while (7c) has to live with a proper noun reading for the 's-morpheme - which at least guarantees name chain continuity (Michael @NPHR 's @N< Friends @N<). Only (7b) remains a total failure, the colloquial short form of the English co-ordinator not being listed in my Portuguese lexicon.

Are there alternatives to the semi-heuristic solution to the name chain problem proposed above? A short glance at the telephone directory of any larger town may convince even the most optimistic linguist of the futility of comprehensive dictionary cover for the whole word field. However, the real problem are not *all* the names that are treated heuristically, but only the ones that can also be assigned some convincing

[31] The only capitalized @N< word in the chain is 'Humana' where the competing reading is not N, but ADJ, which has no problem with being mapped as attributive postnominal (@N<).

[32] While one might regard 'Massachusetts' as a prenominal (@>N), from an English point of view, the principle of the 'left leaning name chain' demands the nominal head reading, which is also more appropriate from a Portuguese point of view, where names do not normally appear prenominally.

(i.e., not too complex) derivational analysis, one that escapes the heuristic filters described above. Research on large corpora can weed out the high frequency cases of these words, which can then be entered into the lexicon. Checking against a 35 million word corpus, where I filtered the output of the parser for derived and unknown words, I found only some hundred words (118 word form types) where a *derivative* analysis had - wrongly - been preferred over the proper noun analysis.[33] Many instances were syntactically isolated in one-word headlines or brackets. 10 lexemes accounted for half the cases. Quite a few of these words had been given a derivative analysis with very short or rare roots ('mar' for 'Maria', 'pá' for Paulo, 'the chemical element 'frâncio' for 'Francisco', 'tê', 'fê' and 'zê'). Since I have a tag in the lexicon (<nd>) for non-deriving lexemes, it was easy to prevent these roots from overgenerating. For others, like the group Cristiana, Cristiano, Cristina (root 'crista') entering the names into the lexicon may be the appropriate solution.

It was not quantitatively possible to inspect the large corpus (especially sentence initial words) for the opposite error, i.e. preferring a proper noun analysis over a lexical derivational analysis, but shorter samples suggest that sentence initial derived words are much less frequent than names. In mid-sentence, finally, the contextual constraints are quite effective and likely to make the right choices.

A final, though, quantification on 21.806 words from the Borba-Ramsey corpus, containing 452 (2.1%) of (real or supposed) name chains, yielded an error rate of 2% for the PROP class (positive and negative errors combined, shaded in table 9). This is higher than the parser's usual morphological/PoS error rate of under 1%, but one must take into consideration that all 11 errors occurred *heuristically*, mostly with lexically unknown words, of which half were spelled incorrectly.

(9) **Table: name frequency statistics**

correct analysis: ***chosen tag:***	**Proper noun**	**Other, simple**	**Other, derived**
PROP	79 (17.5%)	0	0
<HEUR> PROP	362 (80.1%)	2 (0.04%)	0
Other word classes	9 (2.0%)	-	-

The 2 cases of wrong positive choice were the sentence initial words *Lagartixou* (which should have been a verb, derived from *lagarto* - 'lizard'), and *Les* (misspelled for the verbal inflexion form *Lês* - of *ler* 'to read'). Of the 9 cases involving wrong negative choices, 4 were names spelled in lower case (*geraldinho, juraçy, sanhaço, playboy*), 2 were sentence initial words also occurring as common nouns (*nogueira* - nut tree, and *bezerra* - 'female calf'), one was a place name (*Santo Amaro*, read as a

[33] These statistics were done with an older version o the parser, which included ordinary lower case words in the pre-name context. With the up-to-date version, there is not such a strong bias in favour of PROP readings, and the percentage of false positive choices of a derivational reading might be expected to be somewhat higher.

common NP, 'bitter saint'), and the remaining 2 were a noun chain consisting of simple nouns and last, a lexicon error (*Nossa=Senhora* - only - as interjection).

2.2.4.5 Abbreviations and sentence boundary markers

PALMORF treats abbreviations more like a morphological feature than a word class: the tag <ABBR> is added to other - inflexionally defined - word classes. Logically, abbreviations mirror the inflexion categories (like gender and number) of their host classes:

(1)

PROP F S	VARIG (the national air carrier), ARENA (a party), Sudene (Superintendência do Desenvolvimento do Nordeste, a regional development institution), Mercosul (The South American Common Market)
PROP M S	AM (Amazonas, a federal state)
PROP F P	EUA or E.U.A. ('USA')
N M S	AI5 ("Ato Institucional 5", a decree), c.-el (coronel, a title)
N F S	Aids, aids, SIDA (3 variants of 'Aids')
N M P	bps (bauds per second)
ADJ M/F S/P	bras. (brasileiro, 'Brazilian', underspecified for number/gender)
ADV	c/c (conta corrente, 'a conto'), S.E. ('southeast')

An argument in favour of not regarding abbreviations as a separate word class is the fact that abbreviations tend to evolve into full words over time. For this there are both semantic indicators (people don't know any more what the abbreviation stands for, analytically, like in VARIG) and formal indicators, like productive derivation (cp. the discussion of *'petebista'* in 2.2.4.3) and lower case transformation of in-word capitals (*Sudene*). In the last case, the word will "feel" like a proper noun, the abbreviation status being based only on etymology. Finally, on a morphosyntactic level, it would seem counter-intuitive to relegate distinctions like "adjectivity" or "adverbiality" to secondary tags, and assign one homogeneous word class tag (for example, <adj> ABBR or <adv> ABBR) to words with a functional distribution of such diversity.

Abbreviations built from noun phrases may go all the way from a traditional capitalised abbreviation (SIDA), over the "name stage" (Aids, upper case) to a "common noun stage" (aids, lower case). The distinction between names and nouns is thus very fuzzy for abbreviations[34], and this fuzziness is visible in the tagger's lexicon, too. Where abbreviation noun phrases do not contain proper nouns, and not denote people, groups of people, institutions, parties or countries (entities that can function as human agents syntactically), I have assigned N rather than PROP class.

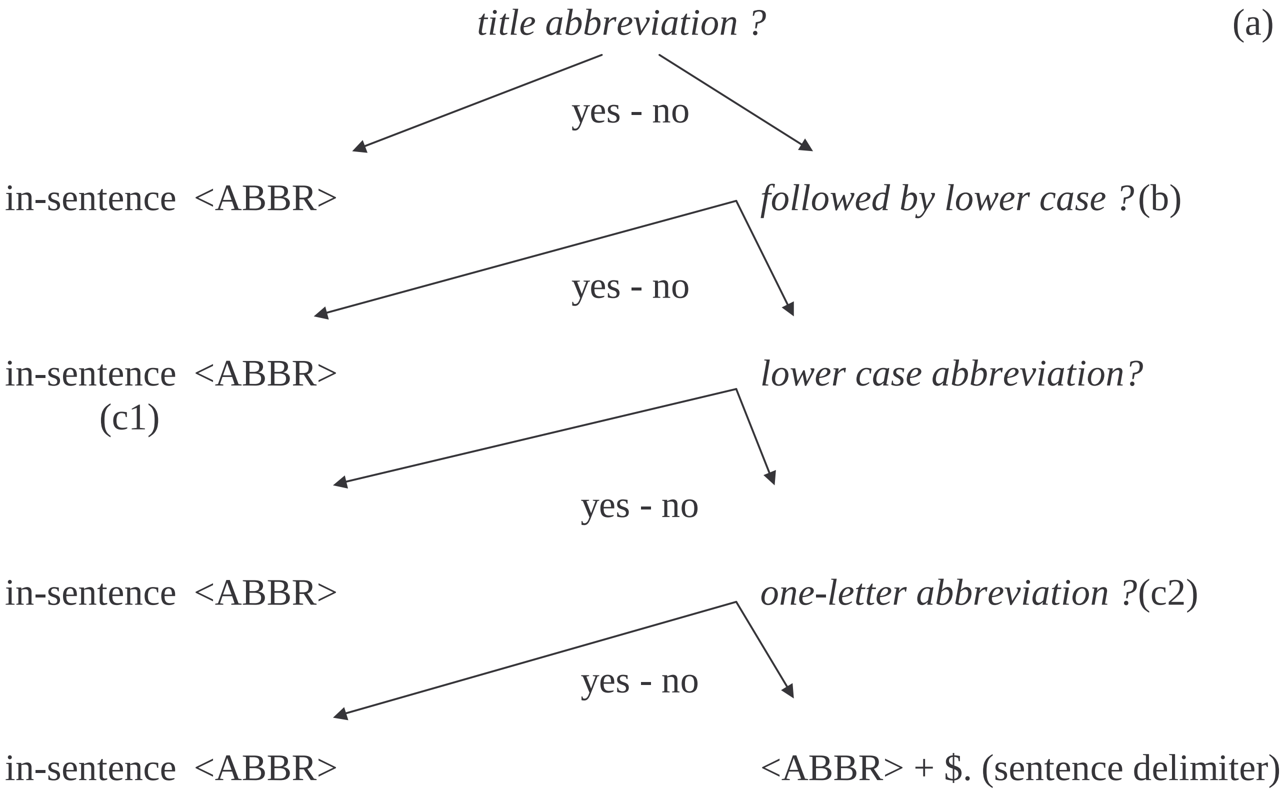
[34] The lexeme category test for number (otherwise used to distinguish between N and PROP), feels somewhat awkward in this case, too, since it would apply to an NP rather than a word (the prototypical inflexion bearer). *AI5* (Acto Institucional 5) as a whole cannot, of course, be pluralized, but *acto* (its head) or even the unnumbered *'acto institucional'* can.

Unlike other words, abbreviations may[35] contain

a) word internal capitalisation even in non-headline text (VARIG, eV)
b) punctuation and other non-letter characters, word internal or word final:
- full stop: *av.* (avenida)
- dash: *c.-a* (conpanhia)
- slash: *d/d* (dias de dato)

While the recognition of dashes and slashes as word internal is not a banality (one needs corresponding lexicon entries and a tagger with a "soft" notion of word delimiters), full stops are a particular nuisance. In order to weed out the alternative reading "sentence delimiter", it is necessary **(a)** to distinguish between those abbreviations that can appear in sentence-final position and those that can't (especially "title" abbreviations like cap., card., com., dr., fr., gen., gov., insp., l., maj., pres., prof., r., rev., s., sarg., sr., ten.), and **(b)** check the following word for potential "sentence-initiality" (i.e., upper case first letter). The last check **(c)** is for single capital letters, which may be part of a name chain when followed by an upper case word (e.g.: J.P.Jacobsen, where, incidentally, the 'J.' is *so* much part of the name, that its pronounciation, 'I', does not disturb any educated Dane).

(2) **Flow chart: abbreviation or clause boundary?**



[35] Since these traits are not universal, they can't be used by the tagger for defining abbreviations. Cp. the "ordinary looking" *Ag* (silver) and *cd* (the SI-unit candela) to the more distinctly "abbreviational" *ag.* (august) and *CD* (compact disk) or *Cd.* (cadmium).

Another case, where meaning bearing characters and punctuation can form a word together, are Arab numbers (tagged NUM <cif>).

(3a) *12.7*
1,000,000
12-7
3/4
3:4
(3b1) *o 10. mandamento*
(3b2) *Veja capítulo 7.*
(3b3) *Tomo VII.*
(3c) *a 2ª Guerra Mundial*
o 5º degrão
(3d) *A aula começa às 8h15*

Here, a dot, comma, slash, colon or dash flanked by numbers without spaces (3a), will be regarded as numeral-internal. Thus, both *12.7* and *1,000,000* will receive the tag chain '<cif> <card> NUM M/F P'. Also more complex expressions like *12-7, 3/4* or *3:4* will be recognised as numeral wholes.

If the dot is word final (3b), however, the word class '<ord><NUM> ADJ M/F S' is assigned, classifying ordinal numbers as a subclass of adjectives. Unless, that is, the number is an integer smaller than 100, and the tagger has classified the preceding word as a prenumeral (tag <+num>, e.g. *'Veja capítulo 7.'* - 'See chapter 7.'). In this case the dot is treated as sentence delimiter, and the numeral as '<cif><card> NUM @N<'. Roman numerals (3b3), by contrast, are lexically treated as a special class of (post-positioned) attributive adjectives, and their postnominal (@N<) function is assigned in the same way as for name chains (*Dom Pedro I* - Dom Pedro the first), the difference between (3b2) and (3b3) being that between a valency bound NP-constituent ('chapter 7' - <card> NUM, meaning "seven") and a modifier ('volume VII' - ADJ, meaning "seventh"). This way, consistency is maintained between the treatment of Roman numerals and other ordinal numbers (which in my system of morphologically motivated word class distinctions have to be tagged ADJ, because they inflexionally behave exactly like other, more "prototypical" adjectives).

Like in English '2nd', '3rd', Portuguese has letter markers for ordinal numbers (3c), *"ª"* (feminine singular) and *"º"* (masculine singular), the morphological tag string being either '<ord><NUM> ADJ F S' (*a 2ª Guerra Mundial* -The Second World War) or '<ord><NUM> ADJ M S' (*o 5º degrão* the 5th degree).

Worst is case (3d), where the letter 'h' (for *hora* - hour) intrudes into a string of numbers without blanks. Here, if the 'h' is in the 2nd or 3rd position, indicating a one or two digit number of hours[36], the tag will be <cif>**<temp>** ADV denoting a temporal adverb.

[36] The 'h' notation is not restricted to the 24-hour-clock, it also appears in connection with, for instance, sports results: 300 quilómetros em 30h35.

2.2.4.6 The human factor: variations and spelling errors

In spite of repeated joint Luso-Brazilian efforts to establish common norms for Portuguese orthography, the two language varieties, European and Brazilian Portuguese, are slowly drifting apart, first of all in terms of pronunciation, but more and more visibly also in spelling. Any tagger for Portuguese must take this into consideration[37].

- 1. 'c' and 'p' before 'c', 'ç' and 't' in etymologically Latin consonant clusters are often dropped in Brazilian pronunciation and spelling, but always preserved in Luso-Portuguese: *activo - ativo, nocturno - noturno, acção - ação*
- 2. Stressed vowels before 'n' and 'm' receive the circumflex in Brazil ("closed" nasalised pronunciation), but acute in Luso-Portuguese ("open" pronunciation): *anônimo - anónimo, convênio - convénio*
- 3. Luso-Portuguese 'mn' and 'nn' is in some words reduced to 'n' in Brazil: *conosco - connosco, indene - indemne*
- 4. the [kw] - or [gw] - pronunciation of 'qu' or 'gu' before light vowels is marked orthographically in Brazilian Portuguese with the umlaut sign: *agüentar - aguentar*.
- 5. The open pronunciation of 'e' in 'eia' and 'eico' is only marked by accentuation in Brazil: *idéia - ideia*
- 6. 'oi' alternates with 'ou', the last one being preferred in Brazil
- 7. There are a few differences in the accentuation of verbal endings, as in *caiu - caíu, amamos - amámos* (preterito perfeito tense) *perdôo - perdoo*

Only for a few cases (especially in group 1) these variations are listed in the main lexicon or the inflexion endings lexicon (most of group 7). In all other cases *PALMORF* "knows" the Brazilian form, and tests for variation possibilities if a first analysis fails. This test is based on simple string substitution, using the following pairs:

(1)

Brazil	c	ç	t	c	ç	t	ên	êm	ôn	ôm	ou	gü	qü	n	n	éia	éic
Europe	cc	cç	ct	pc	pç	pt	én	ém	ón	óm	oi	gu	qu	mn	nn	eia	eic

Some spelling variation in my corpus is due to "phonetic spelling", i.e. an author's attempt to "invent" a spelling variant for local-dialectal or colloquial-sociolectal differences in pronunciation, as in 'r-dropping', where the [H]-pronunciation of

[37] Many of the examples below would be made obsolete by a proposed orthographic reform abolishing many accents, but - first - the reform isn't very likely to pass all bureaucratic and ideological hurdles any time soon, and - second - texts would still mirror the older use in an unpredictable and personal way. Therefore, accent-adding or -removing heuristics may be the most sensible solution.

word-final 'r', especially in infinitives, approaches the zero-morpheme: **amá - amar* (to love), **agradecê - agradecer* (to thank), **mulhé - mulher* (woman).

The quantitatively most demanding problem, however, is faulty accentuation, due to typist errors when compiling the corpus (the Borba-Ramsey corpus on the European Corpus Initiative CD-ROM was not scanned or collected from pre-existing electronic text, but typed), or to 8th-bit-ASCII losses in traffic accidents on the information highway (where only English gets a safe ride). Removing or adding accentuation may, however, lead to mistakes, where both the accentuated and the un-accentuated word form represent perfectly normal lexical items, as in *maca* (hammock), *maça* (club) and *maçã* (apple). Also, there might be ambiguity as to <u>which</u> accent to add. Therefore, most of the accentuation heuristics module is only used on otherwise unanalysable words. Safe bets are the adding of the til in word final 'ao' and 'oes' (which are nearly unthinkable without the accent), yielding 'ão' and 'ões', whereas the change of 'c' into 'ç' before dark vowels is much more likely in the suffix '-ção' (plural '-ções') than, say, in word-initial position.

Of the non-nasal accents in Portuguese, the grave accent only appears when the preposition *a* is fused with pronouns whose first letter is 'a': *à (=a a), às (=a as), àquela (=a aquela)*. Since, on the tagging level, the parser has not yet enough contextual knowledge to disambiguate the isolated pronoun from the misspelled fused form *a + pronoun*, no accent-adding is attempted here.

The acute and circumflex accent spelling errors are handled by the tagger module in the following way:

If there is no prior analysis, and if:

(a1) the word contains no accent, and only 1 vowel
-> add an acute accent to the vowel
-> if the word is still unanalysable, add a circumflex instead

(a2) the word contains no accent, and more than 1 vowel
-> look the word's potential stems up as unaccented root-forms ("R-forms") in the lexicon.

Since the acute- and circumflex- accents in Portuguese - besides denoting vowel opening in 'e' and 'o' - are used as stress markers, and since stress can change in derivation, - accented potentially suffix-taking word stems (i.e. typically nouns and adjectives[38]) have "R-forms" (derivation root forms) entered in the lexicon, where the accent has been removed. Ordinarily, these are intended only to be used in combination with a stress-taking suffix, like *'-ável'* or *'-inho'*. In the spelling correction module, however, this condition is suspended, and "R-forms" may be used to recognise missing accent errors in suffix-less words. There <u>are</u> (acute- and

[38] Verbs, too, combine with a number of suffixes, but all verbs' base forms (infinitives) have oxitonous stress, and since accents in Portuguese are stress markers, no extra lexicon entry is necessary here.

circumflex-) accented words without R-forms, but these are typically one-syllable words with word-final vowel, where Portuguese orthographic convention adds accents with a phonetic distinctive value. These words are covered by procedure (a1).

-> if the word is still unanalysable, try the acute accent on one vowel after the other.

This procedure covers the few cases of multi-syllable function words (i.e., without R-forms) or missing accent errors in verb forms - other than monosyllabic (a1) - that are not covered by Luso-Brazilian variation rules.

(b1) the word does contain an accented vowel

-> remove the accent, unless it is located word-final

Word-final accents may be changed but not removed, because (a) this accent position is rarely chosen by error, and (b) word-final unaccented vowels, mimicking inflexion endings, bear a great risk of overgeneration, i.e. false positive analyses.

-> if the word is still unanalisable, exchange acute and circumflex instead

In the final analysis, in order to retain corpus fidelity[39], all changes - variation or spelling correction - are marked with the ALT-tag (='altered'), after the word form in question. The only exception are variations listed separately in the main or inflexion endings lexicon. These will sometimes be marked as rare (<Rare>), Brazilian (B) or European (L), but no canonical form will be given.

Below a short list of examples indicating the use of the ALT-tag:

(2)

```
        moiro ALT mouro
                "mouro"  ADJ M S
                "mouro"  N M S

(a1)    ve ALT vê
                "vê" <Rare> N M S
                "ver" V IMP 2S VFIN
                "ver" V PR 3S IND VFIN

(a2)    inaudivel ALT inaudível
                "inaudível"  ADJ M/F S

(b1)    francêsa ALT francesa
                "francês" N F S
                "francês" ADJ F S
```

[39] Ideally, any analysed corpus excerpt should allow the reconstruction of the original text. Therefore, all word form changes, typically introduced by the preprocessor, like splitting of fused preposition-determiner units (*da, nele* , marked by <sam>-tags), fusion into polylexicals (*em=vez=de*) or orthographical canonisation (the "ALT-case") must be marked on the altered form.

```
pôlos ALT pólos
        "pólo" N M S
```

Obviously, such changes may create unrealistic words with unwanted and improbable analyses. Thus the slang-decoder rule that changes word final '-ê' into an infinitive '-er' might, for the French-Portuguese word *ateliê*, permit an analysis like "a+tela+ia+er", by wrongly "recognising" an infinitive ending and reducing the word to the root 'tela' - while drawing upon both prefix- and suffix-lexica. Therefore, derivational depth is limited in these cases, to one prefix (and no suffix). Similar, but less rigorous, restrictions apply to Luso-Brazilian variation and spelling correction.

2.2.4.7 Heuristics: The last promille

About 0.05% - 0.2%[40] of lower case word forms in running text cannot be reduced to stems found in the PALMORF lexicon, even when using the derivational, variational or correctional modules described earlier. Name heuristics is not used on lower case word forms, exceptions like unknown pharmaceutical names being treated as common nouns.

Since the parser's higher levels (for example, syntax) need *some* reading for every word to work on, these unanalysable lower case word forms need to be given one or more heuristic readings with regard to word class and inflexion morphology. Three main groups may be distinguished, comprising of roughly one third of the cases each (Cp. the corresponding statistics table in 2.2.6 on recall figures):

a) orthographic errors not detected by the accent module
b) unknown and underivable Portuguese words or abbreviations
c) unknown foreign loan words

Sadly, for optimal performance, the three groups would require different strategies. Foreign words appearing in running Portuguese text are typically nouns or noun phrases, and trying to identify verbal elements only causes trouble. In "real" Portuguese words without spelling errors, structural clues - like inflexion endings and suffixes - should be emphasised. These will be meaningful in misspelled Portuguese words, too, but, in addition, specific rules about letter manipulation (doubling of letters, missing letters, letter inversion, missing blanks etc.) and even knowledge about keyboard characteristics might make a difference.

Motivated by a grammatical perspective rather than probabilistics, my approach has been to emphasise groups (a) and (b) and look for *Portuguese* morphological clues in words with unknown stems. Since prefixes have very little bearing on the probability of a word's word class or inflexional categories, only the inflexion endings and suffix lexica are used. As in ordinary analysis (chapter 2.2.3)

[40] These figures are heavily dependent on text type and corpus quality (i.e. number of orthographical errors). As I corrected and improved my lexica, the percentage of unanalysed word forms has fallen to well below 0.1% for "good quality" texts.

the tagger tries to identify a word from the right, i.e. backwards, cutting off potential endings or suffixes and checking for the remaining stem in the root lexicon (the main lexicon). Normally, using Karlsson's law (Karlsson 1992, 1995, discussed in chapter 3.4), the tagger would try to make the root as long as possible, and to use as few derivational layers as possible. For unanalysable words, however, I use the opposite strategy: Since I am looking for a hypothetical root, inflexion endings and suffixes are all I've got, and I try to make their half of the word (the right hand part) as large as possible.

Working with a minimal root length of 3 letters, and calling my hypothetical root 'xxx', I will start by replacing only the first 3 letters of the word in question by 'xxx' and try for an analysis, then I will replace the first 4 letters by 'xxx', and so on, until - if necessary - the whole word is replaced by 'xxx'.[41] For a word like *ontogeneticamente* the rewriting record will yield the chain below. Here, the full chain is given, with *all* readings encountered in the process. In the real case, however, the tagger - preferring long derivations/endings to short ones - would stop searching at the *xxxticamente* -level, where the first group of readings is found. In fact, the adverbial use of an adjectively suffixed word is much more likely than hitting upon, say, a "root-only" noun whose last 9 letters happen to include both the '-ico' and the '-mente' letter chains by chance.

In the readings lists, the tagger uses '###' to earmark lines it wants to discard because they only differ from the preceding one by deriving from a different base form, a distinction that is irrelevant with imaginary roots. Even in ordinary derivation (i.e., with real roots), the base form lexeme distinction for derivations is not upheld[42], because it has no significance for the word's word class and inflexion traits (which are based purely on the *last* suffix and its endings).

(1) **Flow chart: xxx-roots**

ontogeneticamente -> no analysis

*xxx**ogeneticamente***
*xxx**geneticamente***
*xxx**eneticamente***
*xxx**neticamente***
*xxx**eticamente***
*xxx**ticamente*** -> suffix '-ico' (variation '-tico') + adverbial ending '-mente'
"ontogene" <DERS -ico [ATTR]> <deadj> ADV

[41] A similar method of partial morphological recognition and circumstantial categorisation might be responsible for a human being's successful inflexional and syntactic treatment of unknown words in a known language; the Portuguese word games "collorido" (president Collor & *colorido* - 'coloured') and "tucanagem" (the party of the *tucanos* & *sacanagem* - 'dirty work'), for instance, will not be understood by a cultural novice in Brazil, even if he is a native speaker of European Portuguese - but he will still be able to identify both as singular, the first as a past participle ('-do') and the second as an abstract noun ('-agem') of the feminine gender.

[42] Of course, in words without derivational morphemes, the distinction of different stems is semantically important even in the case of identical word class and inflexion traits, and it will be upheld and later disambiguated contextually.

```
                        "ontogene"  <DERS -ico [ATTR]> <deadj> ADV ###[43]
xxxicamente          -> suffix '-ico' + adverbial ending '-mente'
                        "ontogene"  <DERS -ico [ATTR]> <deadj> ADV
                        "ontogene"  <DERS -ico [ATTR]> <deadj> ADV ###
                        "ontogenea"  <DERS -ico [ATTR]> <deadj> ADV ###
                        "ontogeneo"  <DERS -ico [ATTR]> <deadj> ADV ###
xxxcamente
xxxamente            -> adverbial ending '-mente' (variation '-amente')
                        "ontogenetico"  <xxxo>  <deadj> ADV
xxxmente
xxxente              -> "present participle"-suffix '-ente'
                        "ontogeneticamer"  <DERS -ente [PART.PR]> ADJ M/F S
                        "ontogeneticamer"  <DERS -ente [AGENT]> N M/F S
                        "ontogeneticamir"  <DERS -ente [PART.PR]> ADJ M/F S ###
                        "ontogeneticamir"  <DERS -ente [AGENT]> N M/F S ###
                     -> causative suffix '-entar'[44] + verbal inflexion ending '-e'
                        "ontogeneticam"  <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN
                        "ontogeneticam"  <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN   ###
                        "ontogeneticama"  <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN   ###
                        "ontogeneticamo"  <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN   ###
                        "ontogeneticamo"  <DERS -ar [CAUSE]> V PR 1/3S SUBJ VFIN   ###
xxxnte
xxxte
xxxe                 -> verbal inflexion ending '-e'
                        "ontogeneticamenter"  <xxxer>  V IMP 2S VFIN
                        "ontogeneticamentir"  <xxxir>  V IMP 2S VFIN ###
                        "ontogeneticamenter"  <xxxer>  V PR 3S IND VFIN
                        "ontogeneticamentir"  <xxxir>  V PR 3S IND VFIN ###
                        "ontogeneticamentar"  <xxxar>  V PR 1/3S SUBJ VFIN
xxx                  -> no derivation or inflexion
                        "ontogeneticamente"  <xxx> N F S
                        "ontogeneticamente"  <xxx> N M S
```

Roots with 'xxx' are present in the core lexicon alongside the "real" roots, including the necessary stem alternations for verbs:

(2) **xxx-roots: lexicon entries**

xxx#=#<sf.xxx>#######54572	feminine noun, typically foreign
xxx#=#<sm.xxx>#######54573	masculine noun, typically foreign
xxx-#xxxar#<var>#BbCc#####<xxxar>#54576	stem-stressed forms of '-ar'-verbs
xxx-#xxxer#<v-er>#BbCc#####<xxxer>#54574	stem-stressed forms of '-er'-verbs
xxx-#xxxir#<v-ir>#BbCc#####<xxxir>#54575	stem-stressed forms of '-ir'-verbs
xxxa#=#<sf.xxx>#######54577	feminine noun, typically Portuguese
xxxar#=#<amf>#######59547	Portuguese '-ar'-adjective*

[43] The only difference between this line and the preceding is the distinction between N and ADJ for the hypothetical roots 'ontogene', a difference not showing on the tag line, but immanent in the way the parser tests all root possibilities. In the development version of the parser, tags for root identity number do show the difference. As mentioned above, the '###'-mark means that the line is tag-wise superfluous and scheduled for deletion by local disambiguation.

[44] This suffix is regarded as a variant of '-ar', and therefore normalized in the DER-tag: <DERS -ar [CAUSE]>.

xxxar-#1#<vt>#AaiD#####<xxxar>#54578	endings-stressed forms of '-ar'-verbs
xxxer#=#<sm>#######54666	masculine noun, typically English*
xxxer-#1#<vt>#AaiD#####<xxxer>#54579	endings-stressed forms of '-er'-verbs
xxxia#=#<sf.xxx>#######54665	feminine noun, Latin-Portuguese*
xxxir-#1#<vt>#AaiD#####<xxxir>#54580	endings-stressed forms of '-ir'-verbs
xxxo#=#<adj.xxx>#######54581	ordinary Portuguese adjective
xxxo#=#<sm.xxx>#######54582	masculine noun, typically Portuguese

Besides the typical stems ending in '-o', '-a' and '-r', default stems consisting of a plain 'xxx' have been entered to accommodate for foreign nouns with "un-Portuguese" spelling. Like many other languages, Portuguese will force its own gender system even unto foreign loan words, so a masculine and a feminine case must be distinguished, for later use in the parser's disambiguation module.

Since the tagger's heuristics for unknown words prefers readings with endings (or suffixes) to those without, and longer ones to shorter ones, verbal readings (especially those with inflexion morphemes in 'r', 'a' or 'o') have a "natural" advantage over what really should be nouns or adjectives, especially when these appear in their uninflected singular base form. Lexicon-wise, this tendency is countered by adding three of the most commonly ignored nominal cases specifically into the lexicon: (a) English '-er' nouns otherwise only taken as Portuguese infinitives, (b) Latin-Portuguese '-ia' nouns otherwise only read as verbal forms in the imperfeito tense, and (c) '-ar' adjectives otherwise analysed only as infinitives.

Rule-wise, verbal readings alone are not allowed to stop the heuristics-machine, - it will proceed until it finds a reading with another word class on its way down the chain of hypothetical word forms with ever shorter suffix/endings-parts. In other words, the heuristics-machine will *record* verbal readings, but only stop if a noun, adjective or adverb reading is found in that level's cohort (list of readings). In this context, participles and gerunds - though verbal - are treated as "adjectives" and "adverbs", respectively, because they feature very characteristic endings ('-ado', '-ido', '-ando', '-endo', '-indo').

This raises the possibility of the heuristics-machine progressing from multi-derived analyses (with one or more suffixes) to simple analyses (without suffixes) before it encounters a non-verbal reading. In this case, the application of Karlsson's law does still make sense, and when the heuristics-machine hands its results over to the local disambiguation module, this will select the readings of lowest derivational complexity, weeding out all (read: verbal!) readings containing more (read: verbal!) suffixes than the group selected.

In the misspelled French word *'entaente'*, for example, the verbal reading

(3a) "enta" <DERS -(ent)ar [CAUSE]> V PR 1/3S SUBJ VFIN,

from the 'xxxaente'-level, is removed, leaving only underived verbal readings - form the 'xxxe'-level - along with the desired noun singular reading from the 'xxx'-level.

```
(3b)    entaente ALT xxxaente ALT xxxe ALT xxx
                "entaenter"  <xxxer>  V IMP 2S VFIN
                "entaentir"  <xxxir>  V IMP 2S VFIN ###
                "entaenter"  <xxxer>  V PR 3S IND VFIN
                "entaentir"  <xxxir>  V PR 3S IND VFIN ###
                "entaentar"  <xxxar>  V PR 1/3S SUBJ VFIN
                "entaente"  <xxx>  N F S
                "entaente"  <xxx>  N M S
```

Apart from word-internal disambiguation according to Karlsson's law (concerning minimal derivational complexity, cf. chapter 3.4), the lexical analyser module doesn't do any disambiguation - this is left to the CG-module. Thus, the word class choice between V and N will be contextual (and rule based), as well as the morphological sub-choice of mood and tense (IMP - PR IND) for the verb, and gender (M - F) for the noun. In the prototypical case of a preceding article, the verb reading is ruled out by

```
(4a)    REMOVE (V) IF (-1 ART)
```

and the gender choice is then taken by agreement rules such as

```
(4b)    REMOVE (N M) IF (- 1C DET) (NOT -1 M)
        REMOVE (N F) IF (- 1C DET) (NOT -1 F)
```

Consider the following examples of "unanalysable" words from real corpus sentences, where the final output, after morphological contextual disambiguation, is given:

```
(5a)    inventimanhas ALT xxxas           (also: one ADJ and three rare V-readings)
            "inventimanha" <xxx> N F P 'tricks'
        itamaroxia ALT xxxia              (also: V IMPF 1/3S IND VFIN)
            "itamaroxia" <xxx> N F S 'president Itamar + orthodoxy'

(5b)    corruptograma ALT xxxograma   (3 other NMS-readings removed by local disambiguation)
            "corrupt" <DERS -grama [HV]> N M S 'corruption diagram'
        araraquarenses ALT xxxenses      (3 other ADJ readings removed by local disambiguation)
            "araraquar" <DERS -ense [PATR]> <jh> <jn> ADJ M/F P 'from Araraquara'
        falocrática ALT xxxtica               (1 other AFS-reading removed by local disambiguation)
            "falocrá" <DERS -ico [ATTR]> ADJ F S 'phallocracy, reign of the phallos'
        ontogeneticamente ALT xxxticamente
            "ontogene" <DERS -ico [ATTR]> <deadj> ADV 'by ontogenesis'

 (5c)   sra ALT xxx                            (also: N M S)
            "sra" <xxx> N F S '=s.-ra - Mrs.'
        dra ALT xxx                            (also: N M S)
            "dra" <xxx> N F S '=d.-ra - Dr.'

(5d)    sombrancelhas ALT xxxas           (also: one ADJ and three rare V-readings)
```

```
          "sombrancelha" <xxx> N F P '=sobrancelhas - eye brows'
     balangou ALT xxxou                 (also: N M F and N M S)
          "balangar" <vt> <xxxar> V PS 3S IND VFIN '=balançou - balanced'
     linfadernite ALT xxxite            (3 other NFS-readings removed by local disambiguation)
          "linfadern" <DERS -ite [STATE]> N F S '=linfadenite - lymphadenoid inflammation)
     alfaltada ALT xxxada               (only reading)
          "alfaltar" <vt> <xxxar> V PCP F S '=asfaltado - paved'

(5e) cast ALT xxx                       (also: N F S)
          "cast" <*1> <*2> <xxx> N M S 'English: cast'
     gang ALT xxx                       (also: N M S)
          "gang" <*1> <*2> <xxx> N F S 'English: gang'
     tickets ALT xxxs                   (also: N F P)
          "ticket" <xxx> N M P 'English: tickets'
     hijos ALT xxxos                    (also: ADJ M P)
          "tierra" <xxx> N M P 'Spanish: sons'
```

In (5a) and (5b) the parser assigns correct readings to unknown, but well-formed Portuguese words. Since most ordinary words are already represented in the lexicon, or are at least derivable from lexicon words, unknown words will often come from the realms of word games ('itamaroxia', 'corruptograma'), names ('araraquarense') or science ('falocrática', 'ontogeneticamente'), usually involving productive affixes. Depending on the orthodoxy of the fusion process, these affixes may be recognised (5b), or not (5a). Correctly analysed suffixation greatly eases the burden of disambiguation: in all (5b) cases all members of a cohort have the same word class and morphology, making quick, local disambiguation possible. In (5a), where no suffixes are recognised, cohorts will typically cover several word classes, at least one nominal and one verbal. Still, for Portuguese words, inflexion endings and - in uninflected words - the word's last letter will almost guarantee that the correct reading is at least *part* of the cohort.

The parser proceeds much in the same way in (5d), with the lowest ambiguity occurring, where larger morphological chunks (morphemes) are recognised, as with the "inflammation-suffix" *'-ite'* and the past participle ending *'-ado'*, and the highest ambiguity where the analysis has to rely on inflexion endings alone ('sombrancelhas' and 'balangou', both with cross-word-class ambiguity). What is special about (5d), is the fact that all forms are misspellings, with (phonetically?) added ('so**m**brancelhas') or simply mistyped letters, as in 'a**l**faltada' where the typists right and left ring fingers have been confused on the keyboard. Even so, with the help of the surviving morphological clues and contextual disambiguation, the parser is able to assign the right analysis in most cases, especially if the words still look Portuguese. The examples seem to corroborate Constraint Grammar's claim that good morphology is the basis for any reasonable (syntactic) parse.[45]

[45] Cp. the following quote from *Constraint Grammar* (Karlsson et. al., 1995, p.37):
"The cornerstone of syntax is morphology, especially the language-particular systems of morphological features. Syntactic rules are generalisations telling (a) how word-forms, conceived as complexes of morphological features,

In (5c), 'dra' and 'sra' are not misspellings, but uncommon variants of the more canonical (and longer) title abbreviations 'd.-ra' (doutora) and 's.-ra' (senhora). There is no rule to describe this particular type of variation, so the word forms are treated as "unknown". With the possible exception of the '-a'-ending, both words don't look very Portuguese, and no structure can be found. Since verbs have the highest and nouns the lowest lexicon coverage[46], and since unknown Portuguese three-letter-verbs are virtually unthinkable, the standard analysis for very short words is N with regard to word class, leaving only gender to disambiguation. Here, a preceding feminine article or a following female name will help the CG rules.

(5e), finally, is the hard case - foreign loan words. English 'cast' and 'gang' do not fit with any Portuguese inflexion ending, therefore the default reading N is assigned, gender disambiguation relying on NP-context. In 'tickets' the nominal plural-morpheme is recognised, but the stem - 'ticket' - still lacks a Portuguesish last letter, so again, N is chosen for word class. Spanish loan words, being Romance themselves, fare somewhat better, and 'hijos' (an etymological variant of Portuguese 'filhos') qualifies for both plural nouns and adjectives. Of course, resemblances may be misleading, as in English "profession words" in '-er' ('runner', 'gambler') which mimic Portuguese infinitives. Since this kind of error is especially common within the very complex verbal paradigms, verbal readings - unlike noun readings (which are also favoured by statistics) - are never allowed to be the only ones, as described above. Thus, there is still a chance that contextual information will do the job in the disambiguation module.

In order to test the parser's performance and to identify the strengths and weaknesses of the heuristics strategy of the parser, I have manually inspected 757 "running" instances[47] of lower case word forms where the parser's disambiguation module received its input from the morphological analyser’s heuristics module. The first column shows the word class analysis chosen, and inside the three groups (errors, Portuguese, foreign) the left column gives the number of correct analyses, whereas the right column offers statistics about the mistakes, specifying - and quantifying - what the analysis *should* have been.

occur in particular word order configuations, and (b) what natural classes, "syntactic functions", can be isolated and inferred in such configurations."

[46] In the English CG-system described in (Karlsson et.al. 1995, p. 296), a similar claim is made: *"Because ENGTWOL [i.e. the morphological analyser] very seldom fails to recognize a verb, a verb reading is not assigned [heuristically] without a compelling reason. Word-final 'ed' is a good clue. ..."*.

For Portuguese, I have quantified the problem for a stretch of ca. 200.000 words (cp. table 7), showing that nouns account for 73.08% of unknown words (otherwise: 47.38%), and verbs for ca. 8% (otherwise: 38.5%). The bias against verbs is quite strong: Concluding from the above statistics, a Portuguese word unknown to the PALAVRAS lexical analyser is 9 times more likely to be a noun than a verb (and even if it isn't a noun, it's still three times as likely to be something else rather than a verb).

[47] The words comprise all "unanalysable" word forms in my corpus, that begin with the letters 'a' and 'b'. Since the relative distribution of foreing loan words and Portuguese words depends on which initial letters one works on ('a', for one, is over-representative of Portuguese words, whereas 'x'. 'w' and 'y' are English-only domains), no conclusions can be drawn about these two groups' relative percentages. Inside the Portuguese group, however, the distribution between real words and misspellings may be assumed to be fairly alphabet-independent. Any way, the sampling technique has no significance for error frequencies or distribution in relation to word class, which was the main objective in this case.

(6) Word class distribution and parser performance in "unanalysable" words

	A) orthographic errors		B) Portuguese words		C) foreign words[48]		all	
analysis	correct	other	correct	other	correct	other	correct	other
N	119	ADJ 8 ADV 8 VFIN 3 PRON 1 DET 1 PRP 1	212	ADJ 3	226	ADV 11 ADJ 3 PRON 2 PRP 2	557[49]	43
ADJ	25	N 8 GER 2	95	N 7	8	-	128	17
ADV	3	-	5	-	-	-	8	-
VFIN	13	N 4 PCP 1 ADV 1	9	N 4 ADJ 2	-	N 7 ADJ 1	22	20
PCP	10	-	16	-	-	-	26	-
GER	3	-	-	-	-	-	3	-
INF	9	-	4	-	-	N 4	13	4
	182	38	341	16	234	30	757	84
		(17.3%)		(4.5%)		(11.4%)		(10.0%)

The table shows that, when using lexical heuristics, the parser performs best - not entirely surprisingly - for well-formed Portuguese words (B). Of 323 nouns and adjectives in group B, only 16 (5%) were misanalysed as false positives or false negatives. The probability for an assigned N-tag being correct is as high as 98.6%, for the underrepresented adverb and non-finite verbal class even 100%. All false positive nominal readings (N and ADJ) are still in the nominal class, a fact that is quite favourable for later syntactic analysis.

Figures are lower for group C, unknown loan words, where the chance of an N-tag being correct is only 92.6%, even when allowing for a name-chain-like N-analysis of English adjectives integrated in noun clusters of the type 'big boss'. Finite verb readings, though rare (due to lacking inflexion indicators), are of course all

[48] Only individual words and short integrated groups are treated, foreign language sentences or syntactically complex quotations are treated as "corpus fall-out" in this table.

[49] This number contains all elements of English noun chains, i.e. the tag N is accepted for all elements in both *death star* and *dead star*, though the second contains what in an English analysis would be an adjective. However, since the English NP in the Portuguese sentence functions as one entity and no analytic Portuguese grammar rules apply inside the term, it seems fair to assign the N-tag to the whole *and* its parts, in the same way foreign name chains are treated as PROP PROP ..., even if one element happens etymologically to be an adjective, as in *United Nations.*

failures, and only the little adjective group was a hit, the few cases being triggered by morphologically "Portuguesish" Spanish or Italian words.

The results in group A (misspellings) resemble distributionally those of group B, with a good performance for classes with clear endings, i.e. non-finite verbs and '-mente'-adverbs, and a bad performance for finite verb forms. For the large nominal groups, figures are somewhat lower: 84.4% of N-tags, and only 71.4% of ADJ-tags are correct - though most false positive ADJ-tags are still within the nominal range. The lower figures can be partly explained by the fact that misspelled closed class words (adverbs, pronouns and the like) will get the (default, but wrong) noun reading - a technique that works somewhat better and more naturally for foreign loan words (C), which often are "terms" imported together with the thing or concept they stand for, or names. Also, the percentage of "simplex[50]" words without affixes is much higher among the misspellings in group A than in group B, where all simplex words - being spelled correctly - would have been recognised in the lexicon anyway, due to the good lexicon coverage *before* getting to the heuristics module. Therefore, nouns and adjectives in group A lack the structural information of suffixes that helps the parser in group B: 'xxxo' looks definitely less adjectival than 'xxxístico'. In particular, 'xxxo' invites the N/ADJ-confusion, whereas many suffixes are clearly N or ADJ. Thus, '-ístico' yields a safe adjective reading.

Is it possible, apart from morphological-structural clues, to use "probabilistics pure" for deciding on word class tags for "unanalysable" words? In order to answer this question, I will - in table 7 - rearrange information from table 6 and compare it to whole text data (in this case, from a 197.029 word stretch of corpus). Here, I will only be concerned with the open word classes, nominal, verbal and '-mente'-adverbial.

(7) Open word class frequency for "unanalysable" words as compared to whole text figures

	whole text	"unanalysable" words							
		orthographic errors		Portuguese words		foreign words		all heuristics	
analyses	%	cases	%	cases	%	cases	%	cases	%
N	47.38	131	63.59	232	63.39	237	95.18	600	73.08
ADJ	12.79	33	16.02	100	27.32	12	4.82	145	17.66
ADV[51]	1.26	3 (+9)	1.46	5	1.37	- (+11)	-	8	0.97

[50] "Simplex" words are here defined as words that can be found in the root lexicon without prior removal of prefixes or suffixes. Of course, the larger the lexicon the higher the likelihood of an (etymologically) affix-bearing word appearing in the lexicon, - and thus not needing "live" derivation from the parser.

[51] Only deadjectival '-mente'-adverbs can meaningfully be guessed at heuristically, and therefore only they should enter into the statistics for word class guessing. Also the base line figure of 1.26% for normal text is for '-mente'-adverbs only, the overall ADV frequency is nearly 12 times as high. Since non-'mente'-adverbs are a closed class in Portuguese, the latter will be absent from the heuristics class of wellformed unknown Portuguese words, but in the foreign loan word

VFIN	24.96	16	7.77	9	2.46	-	-	25	3.05
PCP	4.96	11	5.34	16	4.37	-	-	27	3.29
GER	2.47	3	1.46	-	-	-	-	3	0.37
INF	6.17	9	4.37	4	1.09	-	-	13	1.58
		206		366		249		821	

Among other things, the table shows that the noun bias in "unanalysable" words is much stronger than in Portuguese text as a whole, the difference being most marked in foreign loan words. The opposite is true of finite verbs which show a strong tendency to be analysable. Finite verbs are virtually absent from the unknown loan word group. For the non-finite verbal classes the distribution pattern is fairly uniform, again with the exception of foreign loan words.

As might be expected, among the unanalysable words, orthographic errors and correct Portuguese words show a remarkably similar word class distribution.

A lesson from the above findings might be to opt for noun readings and against finite verb readings in unanalysable words, when in doubt, especially where no Portuguese inflexion ending or suffix can be found, suggesting foreign material. As a matter of fact, this strategy has been implemented in the form of heuristical disambiguation rules, that discard VFIN readings and choose N readings for <MORF-HEUR> words, where lower level (i.e. safe) CG-rules haven't been able to decide the case contextually.

group and the orthographical error group they will appear in the false positive section of other word classes (numbers given here in parentheses). In the orthographical error group, both '-mente'-adverbs and closed class adverbs can occur, the first as correct ADV-hits, the other usually as false positive nouns (for instance, 'aimda').

2.2.5 Tagging categories: word classes and inflexion tags

2.2.5.1 Defining word classes morphologically

The parser's tag set contains 14 word class categories, that combine with 24 tags for inflexion categories, yielding several hundred distinct complex tag lines. Thus, in the tag-line 'V PR 3S IND VFIN', for example, the word class 'V' alternates with 12 other word classes, and within the V-class 'PR' (present tense) alternates with 5 other tenses, each of which comes in 6 different shades of person-number combinations, for both 'IND' (indicative) and 'SUBJ' (subjunctive). This way 6x6x2=72 finite verb forms can be described by using only 6+6+2=14 "partial" tags. This analytical character of the tag strings makes them more "transparent", and it also makes things easier for the disambiguation rules. In contrast to other systems (cp., for example, the CLAWS-system, as described in Leech, Garside, Bryant, 1994), a clear distinction is upheld in the tag string between base forms ("words"), word classes and inflexion categories.

Furthermore, word classes are almost exclusively defined in morphological terms, thus keeping them apart from the syntactic categories[52]. A noun (N), for instance, is defined paradigmatically as *that* word class, which features gender as (invariant) lexeme category and number as (variable) word form category. The opposite applies to numerals (NUM), while both gender and number are lexeme categories for proper nouns (PROP), and word form categories for adjectives (ADJ).

Pronouns can be classified along the same lines, yielding a determiner class (DET) with the same (variable) categories as adjectives, and a "specifier" class (SPEC) of "noun-like" pronouns featuring the same (invariant) categories as proper nouns. Personal pronouns (PERS), a third class, has 4 word form categories: number, gender, case and person. All three pronoun classes are distinguishable from the "real" nominal classes by the fact that they do not allow derivation (a typical characteristic of deictics).

Pronouns like 'o' and 'este', that can appear in both "adjectival" and "noun-like" position, are in my system unambiguous members of the DET-class, as judged by the exclusively morphological criterion of inflexional variability with regard to number and gender. The article class doesn't receive special treatment either: 'o' is always[53] DET, whether used as "article", "adjectival demonstrative" or "noun-like demonstrative". (Secondary) tags for <art> and <dem> do appear in the tag list, but they are *not* word class categories, and are therefore only disambiguated at a later stage (the valency level of CG), for use in the MT module.

Among participles, the word class world's enfants terribles, only the past (or perfective) participle (V PCP) is inflexionally productive in Portuguese, and I treat

[52] I owe the urge to define word classes as morphologically as possible to Hans Arndt, who advocates a strict distinction between decontextually defined (primary) tags and distributionally defined syntactic tags in corpus annotation, suggesting category inventory as a means of word class definition in Danish (Arndt, 1992).

[53] that is, if it is *not* the personal object pronoun 'o' or the letter name 'o', or the chemical abbreviation 'O'.

the present participle by derivational rules, permitting both a noun and an adjective reading. The past participle is morphologically marked ('-ido/-ado') and could thus be treated as an inflexional category of the verb, but outside the verb chain it assumes an adjective's word form categories (number and gender), and the analyser chooses in this case to "fuse" the PCP/ADJ ambiguity into a combination of a secondary and a primary tag: <ADJ> V PCP[54].

PALAVRAS' 14 word class tags are the following:

WORD CLASS TAGS

N	Nouns
PROP	Proper names
SPEC	Specifiers (defined as non-inflecting pronouns, that can't be used as prenominals): e.g. indefinite pronouns, nominal quantifiers, nominal relatives
DET	Determiners (defined as inflecting pronouns, that can be used as prenominals): e.g. articles, attributive quantifiers
PERS	Personal pronouns (defined as person-inflecting pronouns)
ADJ	Adjectives (including ordinals, excluding participles which are tagged V PCP)
ADV	Adverbs (both 'primary' adverbs and derived adverbs ending in 'mente')
V	Verbs (full verbs, auxiliaries)
NUM	Numerals (cardinals)
PRP	Prepositions
KS	Subordinating conjunctions
KC	Co-ordinating conjunctions
IN	Interjections
EC	Morphologically "visible" affixes (elemento composto, category not used on higher levels of analysis), e.g."anti-gás"

For (prepositional) polylexicals and incorporates[55] also the following higher level form categories may be used in the lexicon:

[54] The pure verbal reading is thus marked by the *absence* of the <ADJ> tag as well as by the syntactic tag @#ICL. One might argue that a M S tagging for masculine singular (the default) does not make sense in the pure verbal case of *active* participle after the auxiliary*'ter'*, where the participle *only* appears in this form, and is part of a tense construction. From this point of view, a NIL tag would be preferable. The distinction can be made by the parser by using syntactic information that is made available by the syntactic module at the next level. This kind of level-interaction is a positive side-effect of progressive level parsing. However, since filtering PCP M S @#ICL-AUX< into PCP NIL @#ICL-AUX< after *'ter'* doesn't increase the verb chain tags' information content, maybe this transformation is best regarded as a formality that can be left to the parser's user interface and its preference menu. An alternative approach for making the distinction would be a context dependent disambiguation of two secondary tags, <active> and <passive>, for verbal participles.

[55] Here defined as words or polylexicals that appear in incorporating verb constructions (described in 5.3.1), like: *fazer* ***boca-de-siri*** *sobre* ('keep s.th. secret'), *ser* ***batata*** ('to be o.k.'), *dar* ***bola*** *a* ('to court').

PP	Fused prepositional phrase (e.g. *de=graça* 'for free')
VNP	verb-incorporated noun phrase or nominal
VPP	verb-incorporated prepositional phrase
VADV	verb-incorporated adverbial phrase or adverb

Where secondary tags (shown as <...>) are retained in the analysis, adverbs (ADV) and the pronoun word classes (SPEC, DET, PERS) are further differentiated into subclasses, two of which (<rel> [relatives] and <interr> [interrogatives]) are functional features of a (shared) closed list of words so important for contextual disambiguation, that I have chosen to disambiguate them "early", i.e. on the morphological/PoS-level in spite of there not being any morphological or lexical basis to make the distinction (which is really syntactic). Other secondary tags (e.g. valency tags like <vt> for transitive verb) also help disambiguate the primary [morphological and, "later", syntactic] tags (of other words), but are not disambiguated themselves on the tagging or parsing levels. Like purely semantic tags (e.g. <prof> for profession) they may, however, be useful for resolving lexical polysemy on a higher (semantic) level of analysis.

INFLEXION TAGS

Gender: **M** (male), **F** (female), **M/F** [for: N', PROP', SPEC', DET, PERS, ADJ, V PCP, NUM]

Number: **S (singular), P (plural), S/P** [for: N, PROP', SPEC', DET, PERS, ADJ, V PCP, V VFIN, INF, NUM]

Case: **NOM** (nominative), **ACC** (accusative), **DAT** (dative), **PIV** (prepositive), **ACC/DAT, NOM/PIV** [for: PERS]

Person: **1** (first person), **2** (second person), **3** (third person), fused with number:
1S, 1P, 2S, 2P, 3S, 3P, 1/3S, 0/1/3S [for: PERS, V VFIN, V INF]

Tense: **PR** (present tense), **IMPF** (imperfeito), **PS** (perfeito simples),
MQP (mais-que-perfeito), **FUT** (futuro), **COND** (condicional) [for: V VFIN]

Mood: **IND** (indicative), **SUBJ** (subjunctive), **IMP** (imperative) [for: V VFIN]

Finiteness: **VFIN** (finite verb), **INF** (infinitive), **PCP** (participle), **GER** (gerund) [for: V]

(In this table, " ' " after a category means that the category in question for this word class is a lexeme category, and thus derived directly from the lexicon. No " ' " means that the category in question is a word form category for this word class, and thus expressed by inflexion.)

Inflexion tags combine with word classes as follows (* means 'lexeme category'):

word class	**gender**	**number**	**case**	**person**	**tense**	**mood**	***morph. marked***	***can derive***
N	+*	+						+
PROP	+*	+*					capitalisation	+
SPEC	+*	+*						
DET	+	+						
PERS	+	+	+	+				
ADJ	+	+						+
ADV							(-mente)	(+)
V		+		+	+	+		+
V PCP	+	+					-ad-/-id-	+
NUM	+	+*						
PRP								
KS								
KC								
IN							(!)	
EC							hyphen	

As can be seen from the above, it is possible to distinguish and define most classes by their word form and lexeme categories alone, e.g. the difference between nouns and adjectives would be, that in the former gender is a lexeme category, and in the latter it is not. Using these criteria alone, though, would leave PROP and SPEC in one class, as well as DET, ADJ and the subclass of V PCP. Further differentiation is possible by morphological markers and derivation paradigms: PROP is capitalised, SPEC is not. V PCP is marked 'ad'/'id' (on verbal roots), and DET can not be used as a derivational root.

Finally, only KS/KC, PRP, IN, and EC cannot be defined morphologically or paradigmatically, jointly forming a kind of (closed?) particle class. Conjunctions and prepositions are syntactically defined constituent “junctors” with much in common, and might be seen as subclasses of the same morphological class (for a discussion of conjunctional treatment of prepositions, see Bick, 1999).[56]

The EC class of affixes can be defined as a class of hyphenated bound morphemes (without inflexion categories) disjunct with all other PoS categories. The main reason for introducing the EC word class at all (and not as ordinary prefixes) was consistency with regard to the word boundary concept used elsewhere in the preprocessor and morphological analyser, defining a word as a text string limited by

[56] If it wasn't for the blanks surrounding them, prepositions might even be regarded not as words, but as structural morphemes attached to semantically heavier words, for example, as "case markers" for nouns.

blank spaces[57], hyphens[58] or - outside abbreviations - punctuation. In the higher levels of the newest version of the parser (1999), EC-affixes are rehyphenated and reattached to the main word body, losing their word class tag in the process.

[57] In order to establish closely knit syntactic or semantic units, that distributionally or translationally behave like words, the preprocessor can fuse fixed expressions into words by replacing spaces with equal-signs. The list of fused terms comprises some complex function words (e.g. a complex prepositions like*em=vez=de* - instead of) and verb incorporates (e.g. *dar* _*à=luz*_ - 'to give birth'), as well as names and terms that cannot be separated without destroying the basic meaning of the compound (e.g. *'Estados=Unidos'*).

[58] Hyphenated clitics are thus regarded as words, and this view is extended to the special case of European Portuguese mesoclitics with hyphens on _both_ sides, so *'comê-lo-ei'* is seen as _two_, not three words, with an object pronoun embedded between stem and inflexion ending.

2.2.5.2 The individual word classes and inflexional tag combinations

In this section the different word classes are presented together with their primary and secondary tags. The tables list primary tags in the first column, and secondary tags in the second column. Examples for the usage of the primary tags are given directly under the word class heading in the form of a simple (non-exhaustive) list, while examples for secondary tag usage are entered in the third column of the tables themselves, matching line by line the second column tags to be illustrated. An apostroph after an inflexion feature (e.g., M' in the noun section) means the category in question (gender, in the case of M') is a lexeme category (i.e. can not be freely inflected in that word class). If different, the corresponding lexicon entry for a word class and its lexeme features is given in square brackets.

N nouns (some abbreviations)

livro N M' S
árvore N F' S
leoa N F S
comunista N M'/F' S
xícaras N F P

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
gender	<+n> title	senhor Freire
M' (M) male [sm]	<qu> measure	uma garrafa de vinho
F' (F) female [sf]	<num+> unit	20 metros
M'/F' male/female [smf]	<+num> series	cap. 7
number	<cc><ac> countable	árvores
S (S') singular [smS, sfS]	<cm><am> mass noun	dinheiro
P (P') plural [smP, sfP]	<attr> attributive use likely	uma mulher comunista
(S'/P') [smSP, sfSP, smfSP]	<dur><quant> likely adv. object	durar anos
	<+PRP> PP-valency	
	<+de+INF>	uma discussão sobre
		a idéia de visitá-lo

PROP proper nouns (some abbreviations)

(o) Brasil PROP M' S'
(os) Apeninos PROP M' P'
(a) Funai <ABBR> PROP F' S'

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

gender M' male [nm] F' female [nf] *number* S' singular [unmarked] P' plural [nmP, nfP]	<u>' usually without article	Portugal

SPEC "specifiers": independent pronouns

indefinite pronouns (non-adjectival quantifiers)

tudo SPEC M' S'
isto SPEC M' S'
ninguém SPEC M' S'

non-adjectival relatives and interrogatives

quem SPEC M/F S/F
a=qual SPEC F S
que SPEC M/F S/P

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
gender M' absolute male ("neuter", "uter") *number* S' absolute singular	*saturated NP* <dem> demonstratives <quant0> non-inflecting quantifier <enum> enumeratives <enum><hum> +HUM enumerative	tudo acabou.. aquilo, isto tudo, nada,um=pouco cada=um cada=qual alguém, ninguém
relatives/interrogatives also: *gender:* M male, F female, M/F *number:* S singular, P plural, S/P *(both categories more "anaphorical" than morphol.)*	<rel> relatives <interr> interrogatives <rel><hum> +HUM relatives <interr><hum> +HUM interrogatives	a janela que quebrei o=qual, os=quais quem foram os outros?

Traditionally, 'o=que' is regarded as a (pronoun) unit, and as such should be included in the SPEC list. However, due to the ambiguity between synthetic reading ('o=que') and analytical reading ('o que'), this is problematic in a word based grammar like CG, and individual word class tagging is chosen in the parser's CG proper, i.e. on the disambiguation stage, with 'o' functioning as modifier in the synthetic, and as head in the analytical case. Compare:

O @>N que quer? - Um bolo. (synthethical)

[What would you like? - a cake.]

Não quero este, quero o @<ACC que vi ontem. (analytical)

[I don't want this (one), I want the one (that) I saw yesterday.]

At a later stage, an automatic post-editing program reassembles 'o=que' in those cases where 'o' has received a @>N tag.

DET determiners, articles, attributive quantifiers

este DET M S
esta DET F S
estes DET M P
estas DET F P
cujos DET M P

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
gender	*DET (part of NP)*	algumas grandes empresas
M male	**DETA**: <quant1> quantifier type 1	
F female	<enum> enumerative	todo, todos, ambos
M/F	**DETB**: <dem> demonstrative	estes, essa, aquele
number	<KOMP><igual> equalitative	tal
S singular	<art> article	a, o, as, os
P plural	**DETAB**: <quant2> quantifier type 2	cada, nenhum, alguns, um, qualquer, uns, uns=quantos, certo, um=certo, uma=certa vários DET, diversos DET
	<KOMP><igual> equalitative	tantos
	<integr> integrative	todo=o
	<enum> enumerative	todos=os
	<interr> interrogative	quais, que
	<komp><igual> equalitative	quantos, qual
	DETC: <poss> possessive	meus, seu, nossos
	DETABC: <rel.poss> relative possessive	cujo
	DETD: <diff> differentiator	outros, mesmos
	<ident> identifier	próprio DET
	DETE: <quant3> quantifier type 3	poucos, muitos,
	<KOMP><corr> correlative	mais, menos
	(NUM and some ADJ)	quatro, mil, inúmeros ADJ
	[QUAL: ADJ (also <num>)]	[novo, duplo, último, terceiro, meio, tal, próprio ADJ, outros]
	POST: <poss> possessive	uma carta sua
	<post-det> post-determiner	mesmo, qualquer, tal, todo <integr>, próprio DET/ADJ
	[<post-attr>]	[diverso ADJ, vário ADJ]

PERS personal pronouns

eu PERS M/F 1S NOM
os PERS M 3P ACC
lhes PERS M/F 3P DAT
mim PERS M/F 1S PIV
ela PERS F 3S NOM/PIV
nos PERS M/F 1P DAT/ACC

WORD FORM and **LEXEME CATEGORIES**	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
gender: M F M/F(jf. SPEC) *number* S P (jf. SPEC) *person:* 1,2,3 first-second-third *case* NOM nominative (reto) ACC accusative (obliquo átono) DAT dative (obliquo átono) PIV prepositive (obliquo tónico) NOM/PIV, DAT/ACC	saturated NP' (allows only DETA in left and "próprio"/"mesmo" in right position, and only for NOM/PIV)	pobre de mim, todos nós ele próprio, eu mesmo

ADJ adjectives

cheio ADJ M S
nova ADJ F S
exterior ADJ M/F S
pretos ADJ M P
azul-celeste ADJ M/F S/P

melhor <KOMP> <SUP> <corr> ADJ M/F S
raríssimos <DERS -íssimo [SUP]> ADJ M P

mutualmente <deadj> ADV

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
gender M male [adj unmarked] F female [af] M/F (amf) *number*	<post-attr> only posterior	assim, bastante, certo ADJ, diferente ADJ, diverso ADJ, azul-celeste, (among others, all hyphenated polylexical adjectives)
S singular [adj unmarked]	<ante-attr> often anterior position	alto, pequeno, grande
P plural [amP, afP]	<ante-attr><NUM> only anteriorly	meio, último
S/P [amfSP]	<n> "national", (quite) likely NP-head	dinamarquês
comparison	<DERS -oso> (less) likely NP-head	preguiçoso
COMP comparative [few] DER:-íssimo [SUP] *adverbialisation* (suffix "-mente")	<+PRP> takes valency bound PP-argument	rico em ouro

ADV <deadj>/<lex> derived adverbs in '-mente'

regular ADJ -> regularmente ADV
muito ADV -> muitíssimo ADV SUP
devagar ADV -> devagarinho ADV
de=novo ADV

underived adverbs

hoje ADV
menos ADV

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
deadjectival derivation	**AD-VP**:	
(suffix "-mente")	*<mod> modal adverbs	devagarinho, a=fundo
derivation like adjectives	*comparison*	
e.g.:	mais/menos + ADV	menos devagar
DER:-íssimo [SUP]	**AD-S**:	
DER:-inho [DIM]	<setop><ameta> meta-operator	obviamente, infelizmente,
fixed syntagms with PRP as	*no comparison*	simplesmente, obviamente
first part and N , ADJ,	*clause-initial or clause-final, with*	
SPEC as second	*comma*	
	AD-ADJ/ADV:	
	<quant> intensifier, quantifier	
	no comparison	imensamente rico
underived adverbs,	**AD-ADJ/ADV**:	
no inflexion	<quant> quantifying adverbs	só, bastante, muito, pouco
nonproductive in derivation,	<KOMP><corr> correlative hook	mais1, menos
no graphical markers	<KOMP><igual> equalitative hook	tanto, tão
(closed class)	<komp><igual> equalitative header	quanto, quão, como
	<det> "determiner" subclass	algo, meio, metade, nada,
		quase, que, todo, um=pouco,
		um=tanto
	<post-adv> post-adverb	demais, mais/menos, mesmo
	AD-S/N:	
	<setop><aset> set operator	apenas, até, nem, não, senão,
		sequer, sobretudo, só,
		somente, também, tampouco
	AD-S/PRED:	
	<setop><atemp> time operator	ainda, de=novo, em=breve,
		enfim, já, sempre, mais2, mal
	<dei> deictic adverbs	
	(proforms)	aqui, aí, alí
	<atemp> TIME-adverb	hoje, ontem, depois, nunca
	<aloc> PLACE-adverb	nenhures
	<interr> interrogatives	onde?, por=que?
	<rel> relatives (proforms)	onde, quando, como
	(subordinating)	conforme, segundo,
		assim=como
	<+de>	antes de, depois de
	AD-S-S:	
	<k><kc> conjunctional adverb	pois, por=conseqüência
	clause-initial without comma or	
	<post> with comma	

V verbs

compro	V PR 1S IND
levasse	V IMPF 3S SUBJ
comei	V IMP 2P
comerem	V INF 3P
chamando	V GER

V PCP participles

(morphological definition: '-ado/-ido' on verbal stem)

compradas	PCP F P (regular)
entregues	PCP M/F P (irregular)
comprado	PCP M S (passive/active) or (passive) PCP NIL (active)

WORD FORM and <u>LEXEME</u> (..') CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

mood	<vi> SV intransitive	dormir
VFIN finite	<vt> SVO monotransitive, direct object	comer pão
IND indicative	<vq> mostly cognitive	duvidar que, achar que
SUBJ subjunctive	<va> SVADV monotr., adverbial object	
IMP imperative	<va+LOC> place	morar {em Londres}
INF infinitive	<va+DIR> direction	ir {para a casa}
PCP (past) participle	<va+QUAL> quality	ir {bem}
active: no inflexion	<vt+DIST> distance	caminhar {7 kilômetros}
passive: ->"ADJ"	<vt+TEMP> time	durar {sete meses}
M,F,S,P	<vt+QUANT> quantity	custar {muito dinheiro}
GER gerund	<vta> transobjective, adv. complement	
	<vta+LOC> place	pôr {na mesa}
tense	<vta+DIR> direction	carregar {ao porto}
PR present tense	<PRP^vp> SVP monotr., prep. object	acreditar em
IMPF imperfeito	<vdt>/<a^vtp> SVOO ditransitive, dative	dar ac. a alg. = dar-lhe ac.
PS perfeito simples	<PRP^vtp> SVOP ditransitive, prep.obj.	habituar alg. a ac.
MQP mais-que-perfeito	also: semantic transobjective	chamar alg. de ac.
FUT futurum	<vK> SVC, copula	estar doente, ser presidente
COND condicional	<vtK> SVOC transobjective	deixar alg perturbado
	<vr> SVR reflexive	lavar-se
person	<vrK> SVRC reflexive copula	achar-se um grande escritor
1 first person	<PRP^vrp> SVRP reflexive, prep. obj.	acostumar-se a
2 second person	<vUi> V impersonal	chover
3 third person	<vUK> VO impersonal transitive	faz frio, há muitos países
	<x> X-I auxiliary	poder
number	<x+PCP> X-PCP auxiliary with participle	ter +PCP (present perfect)
S singular	<x+GER> X-GER auxiliary with gerund	estar +GER
P plural	<xt> XO-I transitive auxiliary	fazer alg. lavar-se (make do)
	<PRP^xp> X-PI auxiliary with prepositional particle	ficar a ser mais barato
	<PRP^xtp> XO-PI transitive auxiliary with prepositional particle	encomendar alg. a trabalhar
V PCP	<active> no inflexion, regular one of double participles	temos aceitado a proposta.
gender M F M/F NIL		
number S P NIL	<passive> gender, number irregular one of double participles	elas não foram aceitas na turma.

NUM NUM numerals

(only cardinal numerals are regarded as real NUM, since they can be defined separately: as having number as lexeme category and gender as word form category)

duzentas	NUM DET F P'
uma	NUM DET F S'
três	NUM DET M/F P'

<NUM> tag for other word classes

terceira	<NUM-ord> ADJ F S
triplo	<NUM-mult> ADJ M S
primeiramente	<deadj>ADV
terço	<NUM-fract> N M' S
centenas	<NUM-qu> N F' P

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
number	NUM <card> cardinal	cinco
P' plural (2,3,4,...)	NUM <cif> <card> arab cardinal	17, 1997
S'..singular (1)	ADJ <NUM-ord> ordinal	quinto/-a(s), oitavo/-a(s)
	ADJ <cif> <NUM> arab ordinal	1., 3.
<NUM-ord> *ordinal,*	ADJ <post-attr> post-attributive	século XX
<NUM.mult> *multiple*	ADJ <NUM.mult> multiple	triplo
-> ADJ *gender number*	N <NUM-fract> fraction	oitavo(s)
adverbialisation	N <num+>	bilhão
<NUM-fract> *fraction*	N <NUM-qu> measure noun	centenas
<NUM-qu> *measure*	ADV <cif> <temp>	7h30
-> N *gender* , *number*		

PRP prepositions

sem	PRP

complex prepositions

graças=a	PRP
além=de	PRP

WOR DFORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

prefixal derivation in verbs		*desde* a semana passada
complex adverb formation	<+INF> with infinitive	*para* viver com o seu amigo
graph. def. characteristics:	<+que> with que-clause	*antes* que, *depois* que
no capitalisation in headlines	<komp><corr> correlative header	*de*

KS subordinating conjunctions

que	KS
a=fim=de=que	KS

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
graph. def. characteristics: associated with punctuation no capitalisation in headlines	<+IND> with indicative <+SUBJ> with subjunctive <komp><corr> correlative header	se, de=tal=modo=que a=não=ser=que que, do=que

KC co-ordinating conjunctions

e	KC
ou	KC
mas	KC

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
graph. def. characteristics: also without punctuation no capitalisation in headlines	ADV <kc><k> conjunctional adverbials <post> may be in postposition	pois, por=conseguinte porém, no=entanto

IN interjections

oh	IN

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**
graph. def. characteristics: often followed by '!'	none	adeus, ai, alo

Many interjections are really other word classes in morphological terms, especially adverbs (*não!* - 'no!') or even imperatives (*agarra!* - 'stop him!'), used in an "interjectional way", i.e. in one-word exclamatory sentences or sentence-initial followed by comma and a name ("vocative"). Because of empirical problems with especially the IN - ADJ/ADV disambiguation, I have much reduced the IN class in comparison to what one would find in a traditional paper dictionary.

EC prefixes, in isolation

anti-	EC

WORD FORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

graph. def. characteristics: hyphenation	none	anti-, vice-

The EC class honours the fact that some prefixes occur not fused, but hyphenated in front of their root, thus providing them with a certain "wordiness". Since the phenomenon is productive, and therefore evades lexicalisation, the preprocessor will prefer to pass such words - like other hyphenated words - on to the tagger in bits and pieces for separate analysis of the hyphen-isolated word-parts. None of the parts in isolation, however, will be recognized by the tagger's prefix module, since it is looking for *fused* prefixes. This is why a full word analysis (EC) is easiest to handle at that level. Of course, in the vein of progressive level parsing, one can reattach such prefixes before the syntactic stage, by means of an "inter-processor" (as is done in the 1999 version of the parser).

V...		verb-incorporates[59]
	VNP	nouns, adjectives, nouns phrases
	VPP	prepositional group
	VADV	adverbs, adberbial group
	VFS	finite subclause
	VKS	subordinating conjunction

(chorar) baldes	VNP	'weep extensively'
(ser) batata	VNP	'to be o.k.'
(ter) cabelo=na=venta	VNP_PP	'to be a bitch'
(voar) baixinho	VADV	'to keep a low profile'
(crescer) como=cogumelo	VADV	'grow fast'
(sair) da=linha	VPP	'go too far'
(saber) onde=tem=as=ventas	VFS	'be competent'
(fazer) com=que	VKS	'to pretend'

WORDFORM and LEXEME CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

[59] Though not a morphologically definable class, one can design syntactico-semantic tests for incorporation constructions, like substitution by a simplex verb and co-ordination restrictions for incorporates. However, in practice, the V... class is assigned for reasons of easy syntactic management (in fact, here the preprocessor can do the work of higher parsing levels) and translation quality. Lexicon implementation, though extensive, is still somewhat patchy and inconsistent. For a detailed discussion of incorporating constructions, see chapter 5.3.1).

incorporate class membership is created artificially by the preprocessor *if the incorporate is a polylexical itself, there is '=' ligation* *form categories:* VNP noun phrase or noun VPP preposition phrase VADV adverb or adverb phrase VFS finite subclause *sequences:* VNP_PP, VNP_ADV, VPP_PP, VADV_ADV	*internal syntactic function:* <acc> becomes @<ACC <piv> becomes @<PIV <sc> becomes @<SC <oc> becomes @<OC <advo> becomes @<ADV <adv> becomes @<ADVL <+PRP-piv> takes prepos. obj. <fs-acc> @SUB @#FS-<ACC *sequences:* <acc_oc> @<ACC_<OC <acc_piv> @<ACC_<PIV <acc_advo> @<ACC_<ADV <adv_adv> @<ADVL_<ADVL <piv_piv> @<PIV_<PIV	 ter barbas, saber onde=tem=as=ventas ser batata tremer que=nem=varas=verdes dar bola a ('to court') fazer com=que ter a=alma=em=frangalhos passar das=palavras=aos=fatos

PP prepositional group

de=aluguel	<adj> PP
ao=mesmo=tempo	<adv> PP

WORDFORM and LEXEME (..') CATEGORIES	**SYNTACTIC SUB-CLASSES (secondary tags)**	**TEXT EXAMPLES**

graph. def. characteristics:
word-initial PRP+ '=' (preporcessor)
	<PP.adv> complex adverb
	<PP.adj> complex adjective
		com=a=mão=do=gato
		de=alto=coturno

The PP word class allows both "adjectival" and "adverbial" usage, the subclasses <adj> or <adv> only indicate what's more likely. PPs that *only* appear as either one or the other, are assigned a real ADJ or ADV tag, like in *das=arábias ADJ* ('expert') and *de=novo ADV* ('again') .

Classes N, ADJ, irregular PCP, and V have root- or base-form entries in the lexicon, with only irregular inflexion forms being listed separately. For the SPEC, DET and PERS classes all individual word forms appear in the lexicon with their inflexion tags. Numerals are treated according to their subclasses NUM <card> (cardinals, individual word form entries), ADJ <NUM><ord> (ordinals, base form entries) and N <NUM> (fractions and multiples, base form entries). While numerals only in part mimic other word classes, the abbreviations class (<ABBR>) is completely "parasitic", and for this reason it isn't treated as a word class at all, but tagged as a *morphological feature* along with other such tags (like <*> for capitalisation, <*1> and <*2> for left and right quotes).

q.v. <ABBR> V IMP 2S
c.-tes <ABBR> ADJ M/F P
fig. <ABBR> N F' S
E.U.A <ABBR> PROPR M' P'

ADV, PRP, IN and the K classes have no inflexion forms, and words from these classes get one lexicon entry per word (some words, though, like 'como', can belong to several classes).

In a way, even punctuation marks might be called a morphological (word?) class. Unlike words, punctuation marks are pre-tagged - by prefixing the '$' sign (at the preprocessor level):

$.	full stop
$,	comma
$-	hyphen

This graphical marker ($) tells the tagger what not to analyse with ordinary word form analysis tools, and makes it easier to run word-based statistics on the parser's output.

2.2.5.3 Portuguese particles

By ‘particles’ I will here define those “function word subclasses” of ADV, KS, KC, PRP, SPEC, DET that help glue the sentence together, like relatives, interrogatives, conjunctions and conjunctional adverbs, quantifiers and operator adverbs. These adverbs constitute more or less closed lists (with the possible exception of polylexicals, which are included in the analyser’s lexicon for practical reasons).

The word lists of this chapter have been included mainly in order to illustrate the kind of lexical information that the parsing modules can draw upon. A more detailed discussion of the syntactic function of these word classes can be found in chapter 4.5.4.

Relative adverbs ADV <rel>

CONJUNCTIONAL FUNCTION **<ks>** a=proporção=que, ainda=quando, ao=passo=que, ao=tempo=que, apenas, aquando, assim=como, assim=que, bem=como, cada=vez=que, conforme, consoante, da=mesma=maneira=que, enquanto, logo=que, na=medida=em=que, onde, qual [Rare], quando, segundo, sempre=que, senão=quando, tal=como, toda=a=vez=que, todas=as=vezes=que, tão=como, tão=logo, à=maneira=que, à=medida=que, à=proporção=que
PREPOSITIONAL FUNCTION **<prp>** conforme, consoante, qual, segundo, tão=como
COMPARATIVE FUNCTION **<prp> <komp><igual>** como, quanto, que=nem, quão

The word class of relative adverbs is not universally recognised, often one finds most or all of its members referred to as conjunctions ('venha quando quiser') or prepositions ('grande como um urso'). However, since a "conjunctional" adverb like 'como' in 'não sei como funciona' is morphologically indistinguishable from the prepositional 'como' or the "pure" adverbial variant in, for example, an interrogative sentence like 'como se chama?', I prefer to call them all adverbs and distinguish between semantico-syntactic *sub*-classes in order to prepare for syntactic disambiguation.

A strong argument in favour of the “existence” of relative adverbs in Portuguese is the use of the future subjunctive tense in "temporally relative" finite subclauses like 'me avisem *quando* ele vier!', in much the same way as in postnominal (attributive) or absolute nominal relative subclauses: 'Seja *quem* for', 'Podem comprar os livros *que* acharem interessantes'.

Interrogative adverbs ADV <interr>

a=que=propósito, aonde, como, donde, há=quanto=tempo, onde, para=onde, por=que, por=quê, quando
QUANTIFIERS (INTENSIFIERS) **<quant>** quanto, *only as pre-adjects (@>A):* quão, que

Not all adverbs can appear in all adverbial slots of the Portuguese sentence, and lexical knowledge about which adverbs are allowed where, can be of great use to the

CG-rules at the disambiguation level (cp. 4.5). Interrogative adverbs, for instance, can head infinitive subclauses but not future subjunctive subclauses while the opposite is true of relative adverbs, an important piece of contextual information for the morphological disambiguation of verbs.[60]

Operator adverbs ADV

SET OPERATOR **<setop>/<aset>** apenas, até, não, nem, senão [‘only’], sequer, somente, só, sobretudo, principalmente, também, tampouco, mais +NUM, meno +NUM, mesmo, inclusive
POST-ADJECT (@A<) **<post-adv>** mais, menos, demais *(um bolo mais)*
BEFORE NUMERALS **<+num>** mais *(comeu mais dois bolos)*
TIME OPERATOR **<atemp>** ainda, de=novo, em=breve, enfim, já, já=não, mais *(não mais)*, mal
META OPERATOR **<ameta>** absolutamente, certamente, simplesmente, obviamente, possivelmente, provavelmente, realmente, talvez

Operator adverb distribution is discussed in detail in chapter 4.5.4.5.

Quantifying adverbs (intensifiers) ADV <quant>

assaz, bastante, bem, cada=vez=mais, eminentemente, extremamente, igualmente, imensamente, incrivelmente, mais=ou=menos, mui, muito, muitíssimo, particularmente, pelo=menos, pouco, pouquíssimo, quanto=mais, sobremaneira, sobremodo, terrivelmente, totalmente, tremendamente, vagamente
POST-ADJECTS (@A<) **<post-adv>** demais, paca, por=demais, por=demasiado *(devagar demais)*
MORPHOLOGICAL PRONOUNS **<det>** algo, meio, nada, que, todo, um=tanto, um=pouco
CORRELATIVE COMPARATIVES **<KOMP><corr>** mais, menos, mesmo
EQUALITATIVE COMPARATIVES **<KOMP><igual>** tanto, tão

Syntactically, intensifiers can be defined as items that can[61] appear in adject position (@>A, @A<), modifying adjectives or adverbs where these semantically permit quantifying. Traditionally, it is this syntactic distribution that makes the above words adverbs, even where one morphologically might argue that many really are pronouns, *used* with intensifier *function*. I have chosen to follow the traditional distinction, retaining only a secondary pronoun tag <det>[62].

Deictic adverbs ADV <dei>

TIME: agora, amanhã, então, hoje, nunca, ontem, sempre
PLACE and DIRECTION: alhures, ali, alí, aqui, aí, daqui, lá, nenhures, praqui
MANNER: assim

[60] In Portuguese, the infinitive and future subjunctive forms are identical in all regular verbs.

[61] In which case their secondary tag of 'intensifier' (<quant>) will be "instantiated", i.e. not be removed by CG-rules on the valency/semantic level.

[62] The tag was originally introduced for the sake of ‘todo’ which can sometimes inflect (!) even when used as an adverbial adject (cp. 4.5.3). Since the tag is also used for adverbial ‘nada’ and ‘algo’, it should eventually be supplemented by a <spec> tag, or changed into <pron>.

Deictic adverbs refer to discourse place and time, using "real" and not text context. Like operator adverbs, deictic adverbs can neither be pre-modified by intensifier adjects nor post-modified by PPs. They can, however, appear as "subjects", this being cited as a distinctive trait in some grammars: *hoje e ontem tem sido dias muito agradáveis*[63].

More ordinary functional roles for deictic adverbs are those of adverbial adjunct and adverbial complement where they replace ADVPs in a "pronominal" fashion. From a disambiguation perspective the non-modifiability allows CG-rules like:

```
REMOVE (@>A) (0 <intensifier>) (1 <dei>)
      Discard the preadjectal reading for intensifiers if they are followed by a deictic
REMOVE (@A<) (0 PRP) (-1 <dei>)
      Discard the postadjectal reading for prepositions if they are preceded by a deictic
```

Conjunctional adverbs ADV <k...>

CO-ORDINATING **<kc>** agora, ainda=por=cima, apesar=disso, assim, conseqüentemente, de=contrário, eis=porque, já, ainda=assim, ainda=menos, assim=mesmo, haja=vista, mesmo=assim, nada=menos, nada=obstante, no=mais, ora, ora=pois, ou=seja, pois, pois=bem, pois=então, portanto, quando=muito, quando=não, quer=dizer, senão [‘otherwise’], só=que
POST-POSITIONED **<kc><post>** contudo, entanto, entretanto, nem=por=isso, no=entanto, no=entretanto, porém, todavia
SUBORDINATING **<ks>** [cp. the ADV <rel> list above]
others: nem ... nem, não ... nem, ora ... ora *(all treated analytically at present)*

Conjunctional[64] adverbs are adverbs, that introduce a proposition in much the same way co-ordinating conjunctions do, and bind it to a preceding sentence, establishing a consecutive (*assim, conseqüentemente*), or - more typically - a concessive or adversative relation (*ainda=assim, senão, nada=obstante*). When in sentence-initial position, the former can be syntactically replaced by *'e'*, the latter by *'mas*'.

A number of adversatives may, however, be postpositioned to the right of the focus, too (*porém, todavia, entretanto*):

a	[a] <art> DET F @>N ‘the’
dúvida	[dúvida] <p> N F S @SUBJ> ‘doubt’
$,	
porém	[porim] **<kc> <post> ADV @ADVL>** ‘however’
$,	
persiste	[persistir] <vi> <sN> V PR 3S IND VFIN S:2207 @FMV ‘remains’

[63] In my parser, I retain an adverbial analysis in these cases, both because some deictics (directives like *praqui* and *daqui*) do not seem to have the subject option, and because there is an alternative impersonal predicative analysis, as the viability of the truncated sentence shows *tem side dias muito agradáveis,* as well as the asterisc-icity of subject-pronominalisation: **Eles tem sido dias muito agradáveis.*

[64] In Portuguese grammars, expressions like *connective adverbs, referential adverbs* and *anaphorical adverbs* cover more or less the same concept.

Also others (e.g. *assim*, *pois*, *ora*) may appear in other but sentence-initial position, this being in argument against fusing the classes of conjunctional adverbs and co-ordinating conjunctions. Another reason for maintaining the distinction is the fact that real KC and ADV <kc> can be juxtaposed, while two KC are mutually exclusive:

mas	[mas] **KC** @CO 'but'
ainda=assim	[ainda=assim] **<kc> ADV** @ADVL> 'still'
estará	[estar] <x+GER> V FUT 3S IND VFIN @FAUX 'will be'
faltando	[faltar] <vi> <sN> V GER @IMV @#ICL-AUX< 'missing'
um	[um] <quant2> <arti> DET M S @>N 'a'
componente	[componente] <cc> N M S @<SUBJ 'component'
vital	[vital] <n> ADJ M/F S @N< 'vital'

Subordinating conjunctions

como=que, do=que, entrementes=que, entretanto=que, se, sendo=que, tanto=mais=que
<+SUBJ> a=fim=de=que, a=menos=que, a=modo=que, a=não=ser=que, ainda=que, a=pesar=de=que, bem=que, caso, como=quer=que, conquanto, contanto=que, dado=o=caso=que, dado=que, embora, exceto=se, nem=que, no=caso=que, por=maior=que, por=mais=que, por=menor=que, por=menos=que, por=modo=que, por=muito=que, por=pouco=que, posto=que, quando=mesmo, salvo=se, se=bem=que, seja=que, suposto=que
<+IND/SUBJ> a=ponto=que, a=tal=ponto=que, a=termo=que, como, de=feição=que, de=forma=que, de=jeito=que, de=maneira=que, de=modo=que, de=sorte=que, de=tal=forma=que, de=tal=maneira=que, de=tal=modo=que, de=tal=sorte=que, que, tanto=assim=que
<+IND> ao=passo=que, ca, desde=que, enquanto=que, já=que, pois=que, por=isso=mesmo=que, por=isso=que, porquanto, porque, uma=vez=que, visto=como, visto=que
<+SUBJ_PR> desde, primeiro=do=que, primeiro=que, onde=quer=que, quando=quer=que,
<+SUBJ_FUT>
<+IND/FUT_PR> mal
COMPARATIVE **<komp><corr>** que, do=que

Co-ordinating conjunctions

e, mas, ou, ou=antes, quer, senão ['but', 'but only'], senão=que ['but rather']; *Latin:* et, i.e.
<parkc-1><parkc-2> nem ... nem, ou ... ou, quer ... quer
<others> quanto ... tanto, seja ... seja, seja ... ou, tanto ... como *(all treated analytically)*

Disjunct co-ordinators like *'ou ... ou'* cannot, of course, be morphologically tagged as "one" in a word-based notation scheme, but any *'ou'*, *'nem'* or *'quer'* is tagged for both first and second part function (<parkc-1><parkc-2>) which is then disambiguated by a syntactic grammar module. The two introducing parts of a comparative equalitative construction ('quanto ... tanto', 'tanto ... como') are not treated as co-ordinations at all. Rather, the second part is seen as a postadject argument of the first (cp. chapter 4.5.2).

Independent quantifier pronouns SPEC <quant0>

algo, algum=tanto, nada, nadinha?, nem=um=nem=outro, neres=de=neres, neres=de=pitibiriba, o=demais, outro=tanto, seja=o=que=for, tudinho, tudo, tudo=isso, tudo=isto, tudo=junto, tudo=o=mais, um=isto, um=pequenote, um=pouco, um=tanto, um=tique, um=tiquinho, um=tudo=nada
RELATIVES **<rel>** quanto=mais, todo=quanto, tudo=o=que, tudo=quanto

Independent +HUM pronouns SPEC <hum>

ENUMERATIVE **<enum><hum>** alguém, cada=um, ninguém, toda=a=gente
RELATIVE **<rel><hum>** quem
INTERROGATIVE **<interr><hum>** quem

Interrogative pronouns DET/SPEC <interr>

DET qual/quais, que, que=espécie=de; QUANTIFIER **<quant2>** quanto/-a/-os/-as
SPEC que, quê, quem, quem=mais; QUANTIFIER **<quant0>** quanto=mais

Relative pronouns DET/SPEC <rel>

DET o=qual, a=qual, os=quais, as=quais; QUANTIFIER **<quant2>** quanto/-a/-os/-as
SPEC que, quem; QUANTIFIER **<quant0>** quanto=mais, todo=quanto, tudo=o=que, tudo=quanto

Quantifying determiners DET

<quant1> a=generalidade=de, ambos/-as, dezenas=de, o=comum=de, todo/-a [<integr><post-det>], todo/-a/-os/-as [<enum>]
<quant2> algum/-a/-ns/-as, cada, certo/-a/-os/-as [visse], muito/-a/-os/-as, nenhum/-a/-ns/-as, qualquer, um/-a/-ns/-as, vários/-as, quanto/-a/-os/-as [<interr>], tanto/-a/-os/-as [<KOMP><igual>], mais [<KOMP><corr>], menos, [<KOMP><corr>],
<quant3> diferentes, diversos/-as], muito/-a/-os/-as, vários/-as, bastante/-s, diferentes, diversos/-as, pouco/-a/-os/-as,
<quant2><quant3> muito/-a/-os/-as, vários/-as
<KOMP><corr> mais, menos, mesmo
<KOMP><igual> tanto, tal
<komp><igual> quanto [<quant2>], tal, qual

Prenominal adjectives ADJ <ante-attr>

bom, crítico, futuro, fêmeo, grande, jovem, lindo, livre, mau/má, médio, novo, pequeno, pio, pobre, rico, velho, demasiado, meio, mero

Postnominal adjectives ADJ <post-attr>

assim, bastante, certo [‘certain-safe’], diferente, diverso, I II III IV ...,

Postnominal adverbs ADV <post-attr>

demais, mais [‘yet’], menos, aí, mesmo, paca, pra=chuchu

Simplex prepositions

a, ad, afora, ante, apesar, após, até, com, contra, de [<komp><corr>], diante, durante, em [<+GER>], embora, entre, exceto, fora, malgrado, mediante, per [A], pera [A], perante, por [<+INF>], pra, pós, qua, salvante, senão [where translated as ‘but’], sob, sobre, trás, versus, via
<+que> antes, depois, para, sem
<num+><+num> mais, menos
<komp><corr> de

Specifying word class internal particle ambiguity

Generally the morphological module only disambiguates primary (morphologically and paradigmatically defined) word classes. However, some adverbs and a few pronouns are so important for structuring the sentence and, thus, important for the disambiguation of other words, that some of the secondary (<>) features have been subjected to disambiguation themselves. The main case is the ‘relative’ - ‘interrogative’ ambiguity of *como, onde, quando, quanto, quão, que* and *quem.* Apart from these, only the ‘set operator’ - ‘intensifier’ ambiguity of *mais/menos* and the ‘quantifier’ - ‘conjunctional adverb’ ambiguity in *senão*, that have been treated in a similar way.

Puristically, no harm is done to the integrity of primary word classes, since these subdivisions could easily be fused again afterwards, removing all secondary <>-tags.

Apart from syntactic function, there are semantic reasons for introducing the subdivisions in the disambiguation scheme: as shown in the table below, nearly all subclass distinctions in the adverb group coincide with the necessity of using different translation equivalents in Portuguese-Danish word pairs. Early disambiguation - a hallmark of CG - would make things easier for a bilingually oriented semantics module later on. For the word *mais* even more subtle distinctions might be appropriate (‘temporal’ *não mais - not any more*).

Table: particle subclasses

	ADV	**ADV**	**ADV**	**ADV**	**ADV**	**KS**	**KC**	**PRP**
	<rel>	***<interr>***	***<setop>***	***<quant>***	***<kc>***			
como	som	hvordan				da, fordi[65]		
onde	hvor	hvor						

[65] The distinction between *como <rel> ADV* and *como KS* is really a semantic one - one I wished to make early, i.e. in the morphological/word class module of the parser. Another possibility would be to fuse the two classes, and replacing the KS tag by ADV plus a secondary tag like <cause>. After all, it is not uncommon for adverbs or comparators to share lexical shape with what otherwise might be called a causal conjunction, like the Romance 'por que' (Spanish 'porque', Italian 'perche'), the English 'as' and the - spoken - German 'wo'.

quando	når, da	hvornår						
quanto	[QU+] som	hvor meget						
quão	som	hvor [+A]						
mais			endnu [x]	mere/mest				plus
<post-adv>			[x] til					
<temp>			længere					
menos			[x] mindre	mindre/-st				minus
senão			kun		ellers		men	end

2.2.6 Recall: Quantifying the problems

In this chapter, I will attempt a quantitative evaluation of the performance of the morphological analyser module, i.e. that part of the system most prominent in section 2. Obviously, performance at this level only reflects lexical/morphological coverage and the efficiency of the morphological heuristics and derivation modules. For an evaluation of the performance of the parsing system as a whole, see chapters 3.9 and 8.1.

Since it does not include contextual disambiguation, the morphological analyser will have very low precision figures, directly reflecting the inherent morphological ambiguity of the Portuguese language. Thus, on average, every word form is assigned 2 morphological/PoS readings (cp. chapter 3.2). However, recall[66] is high at the morphological analyser level, and can be measured in a meaningful way. Assuming a reasonably good lexicon coverage and quantification *before* analytical heuristics[67] (as explained in 2.2.4.7), most cases of recall failure for a given word form will be cases of "no reading" rather than "wrong" reading, and the statistics below will be based upon the assumption that, *if* non-heuristic readings are found, the correct one will be among them. With this slight simplification it becomes possible to judge recall by quantifying "unanalysable", no-reading words.[68] This allows automatic extraction of the problematic words for closer inspection, reducing inspection work load from 100% to about 0.5%.

The sample in (1) consists of a 131.981 word corpus of literature and commentaries, containing 604 unanalysable words. For comparison, language specified percentages for loan word frequency in a larger sample (629.364 words, 2599 unanalysables) from the mixed Borba-Ramsey corpus of Brazilian Portuguese are given in parentheses.

[66] Basically, disambiguation improves precision and reduces recall, ultimately - at 100% precision -, recall will become "correctness", i.e. the percentage of correct readings.

[67] The only kind of heuristics that does have a bearing on the numbers below, are some rules for orthographical regional variation, but the respective figures are included in table (1). Capitalized names are here not regarded as "unanalysable", and not included. Name heuristics, involving up to 2% of word forms in running text, is described in detail in 2.2.4.4.

[68] The management of “unanalysable” words is discussed in detail in chapter 2.2.4.7.

(1) Language distribution and error type in unanalysable words

DOMAIN	NUMBER OF TOKENS	PERCENTAGE	
English	77	12.8	(9.3)
French	78	12.9	(3.7)
Italian	10	1.7	(1.5)
Spanish	28	4.6	(0.6)
German	15	2.5	(0.2)
Latin	24	4.0	(2.7)
orthographic variation (European/accentuation)	125	20.7	*Correctables*
other port. orthographic	74	12.3	*Misspellings*
non-capitalised names and abbreviations	37	6.1	***Encyclopaedic lexicon failures***
names and name roots	18	3.0	
abbreviations	19	3.1	
root not found in lexicon	119	19.7	***Core lexicon failures***
found in Aurelio[69]	91	15.1	
not found in Aurelio	28	4.6	
derivation/flexion problem	15	2.5	***Affix lexicon failures***
suffix	8	1.3	
prefix	3	0.5	
inflexion ending	2	0.3	
alternation information	2	0.3	
other	2	0.3	
SUM	604	100.0	

The table shows a roughly equal distribution of unanalysable words between three main groups, (a) foreign loan words, (b) spelling problems (shaded), and (c) lexicon failures (including abbreviations and name derived words). Of course, the spelling problem group will vary greatly in size depending on corpus quality and provenience. Also, the one-register corpus above is not typical with regard to loan word distribution. Ordinarily - as the numbers from the Borba-Ramsey corpus show, English has a larger and French a smaller share in the loan word pool. And while nearly non-existent in the literature corpus, scientific domain words can be quite prominent. Cp. the following percentages from the Borba-Ramsey corpus:

(2)

domain	number	percentage	
medical terms	129	5.0%	(of all unanalysable words)
botanical terms	45	1.7%	(of all unanalysable words)
pharmaceutical names	102	3.9%	(of all unanalysable words)

[69] Aurélio Buarque de Holanda Ferreira, *"Novo Dicionário Aurelio"*, second edition, Rio de Janeiro 1986

The overall frequency of unanalysable words, however, is quite stable: For both the literature and the Borba-Ramsey corpora, as well as for VEJA news magazine texts, the figure is roughly 0.4%.

3

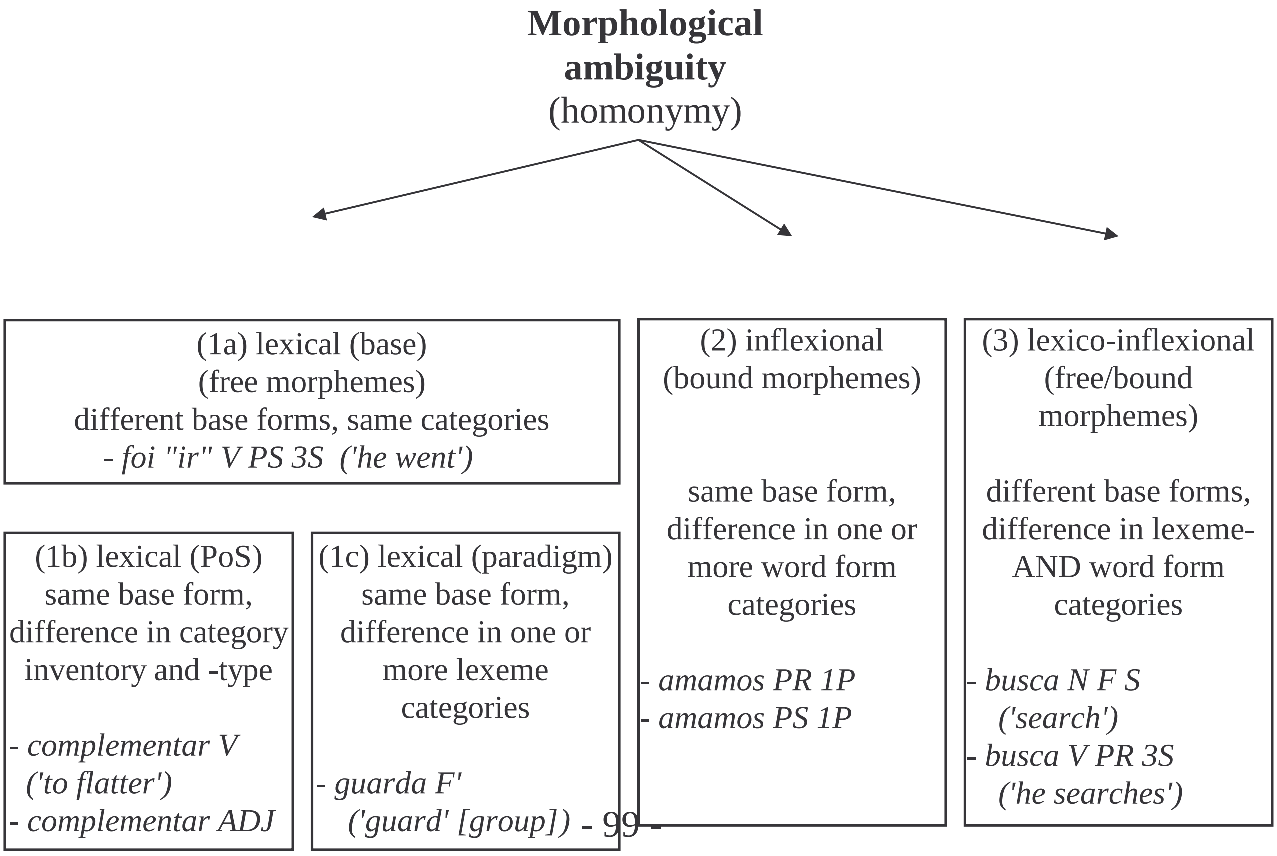
Morphosyntactic disambiguation: The holographic picture

3.1 The problem of ambiguity: A pragmatic approach

3.1.1 Relevant ambiguity

Handling and resolving ambiguity is the central mechanism in the Constraint Grammar formalism. So, in a way, it seems reasonable to assume that which types of ambiguity are specified will condition the kind (and quality) of the grammatical description to be achieved. In a modular, level based, parsing approach, it makes sense to classify ambiguities according to the morphological, syntactic, semantic (and possibly, pragmatic) levels, and then to address the problem of ambiguity with similar tools at ever higher levels, while exploiting different levels and increasing amounts of lexical and contextual information. In this process, any advance in disambiguation on one level would improve the informational leverage for the next level.

In this vein of layered analyses, homonymy can be defined as morphological ambiguity, involving (1) free morphemes (lexical ambiguity), (2) bound morphemes (inflexional ambiguity), or (3) both (lexico-inflexional ambiguity).


Morphological ambiguity (homonymy)
(1a) lexical (base) (free morphemes) different base forms, same categories - foi "ir" V PS 3S ('he went')
(1b) lexical (PoS) same base form, difference in category inventory and -type - complementar V ('to flatter') - complementar ADJ
(1c) lexical (paradigm) same base form, difference in one or more lexeme categories - guarda F' ('guard' [group])
(2) inflexional (bound morphemes) same base form, difference in one or more word form categories - amamos PR 1P - amamos PS 1P
(3) lexico-inflexional (free/bound morphemes) different base forms, difference in lexeme- AND word form categories - busca N F S ('search') - busca V PR 3S ('he searches')

While inflexional ambiguity (2) concerns differences in word form categories, lexical ambiguity can be subclassified according to whether it involves base forms (1a), category inventory (i.e. morphological word class as I define it, 1b) or differences in lexeme category (1c).

Representing the lowest level of context dependent rule based disambiguation, morphological ambiguity (PoS, inflexion) needs special attention, since rules can here only draw upon lexical/morphological context and on each other, *not* on implicit information from any earlier levels of disambiguation. One by one rules will then improve the quality ("unambiguity") of the context clues and thus make life easier for each other.

Many morphological ambiguities can be resolved by using local context and immediate group neighbourhood only (i.e. without global, unbounded rules). Thus rules based on agreement are more prominent on this level than valency based rules[70]. The more fundamental the ambiguity, the more profound a given reading's impact on its surroundings will be, which is why PoS-ambiguity (1b, 3) is more "important" for later stages of analysis than purely paradigmatic (1c) or inflexional (2) ambiguity. In my Portuguese test text data, each PoS-error will on average cause 1-2 syntactic errors around it. Consequently, it is more permissible to use portmanteau-tags for inflexion than for PoS, this being my choice in a few cases of type (2) ambiguity, where many members of a word class lack a certain categorical distinction, like gender in '-ar' -adjectives (M/F) and the present and perfeito simples tense distinction in the 1.person plural of regular verbs (PR/PS). An especially recalcitrant problem is (1a): In the example, all PoS and inflexion tags are the same, but a difference in base form forces the parser to make a semantic lexeme distinction that would otherwise belong to a much higher level of analysis (cp. chapter 3.7.2.1). For the 'ir' - 'ser' pair, 35 rules are needed, many using higher level information, like the copula-valency of 'ser' or the membership of 'ir' in the MOVE-class of intransitive verbs.

Some traditional word class distinctions do not really belong on the morphological level, but are rather syntactic classes derivable not from the words morphological category inventory, but its syntactic uses:

(1)	-ista	noun or adjective ?
(2)	"que"	conjunction or pronoun or determiner ?
(3)	"o"	article or demonstrative pronoun ?

[70] Though valency based rules may become necessary where everything else (short of semantics proper) proves inapplicable, as in (1a).

(4) "quando" conjunction or adverb or WH-word ?

Example (1) refers to an *open, productive* (!) class of "attributive" nouns, many of which are lexicographically registered as adjectives, too. What this *really* means is just that they can appear in syntactic places where one would normally expect an adjective:

(5a) Conhece @FMV muitos @>N **comunistas @<ACC**.
(5b) Leu @FMV vários @>N manifestos @<ACC **comunistas @N<.**

As can be seen from the "noun-function" @<ACC (direct object) and the "adjective-function" @N< (post-nominal modifier adject), the syntactic level has to make the distinction anyway, and adding the word class tags N and ADJ, respectively, is quite redundant. However, though virtually all '-ista' nouns in Portuguese *can* appear postnominally, traditional lexicographic treatment as ADJ is still an indicator of the frequency of this usage for a particular word, and disambiguation on the word class level may be a way of providing early (and easy) syntactic "bootstrapping" information for the next round of (syntactic) CG-rules.

The same holds for the 3 uses of "que" mentioned in (2):

(6a) Sei @FMV **que @SUB** @#FS-<ACC era @FMV comunista @<SC.
(6b) **Que @ACC>** quer @FMV ?
(6c) **Que @>N** carro @ACC> quer @FMV ?

Here, @SUB (subordinator) translates into conjunction word class (KS), @ACC> implies pronoun class (SPEC) and @>N (pre-nominal modifier) the PoS class of determiner (DET). Still, "que" is so central to clause structure, that early disambiguation (i.e. on the morphological, or rather, PoS/inflexion level) is desirable.

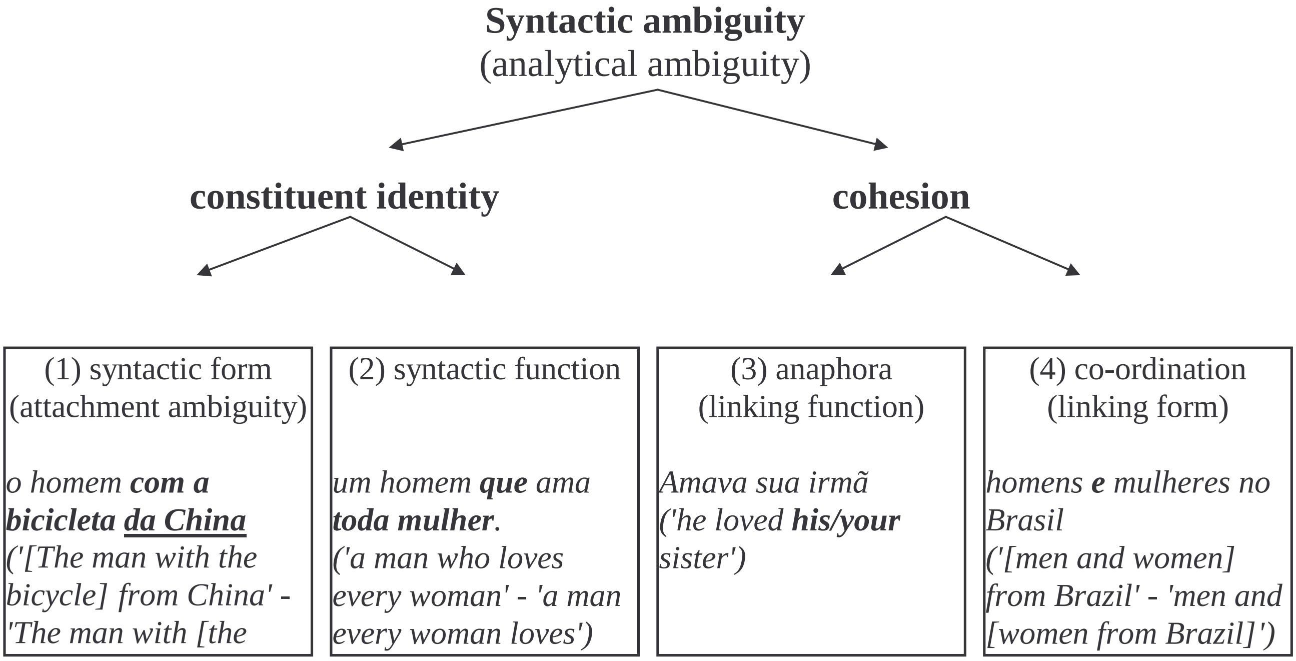
For (3) and (4), I have chosen a slightly different path, opting for one word class (DET for "o" and ADV for "quando"), but adding secondary tags. For "o", <art> matches @>N use, and <dem> pronominal use (@NPHR), but the secondary tags are not disambiguated in the morphological module, the reason being, that the distinction is a *class* feature of the whole determiner class, most of whose members can (also) be used nominally in Portuguese. QU-adverbs like "quando" receive the secondary tags <interr> and <rel>, the latter implying "conjunctional" use (7c), the former covering the traditional (interrogative) adverb reading (7a). These word class boundaries are, however, difficult to maintain. (7d) forces a 'relative' reading not traditionally compatible with the conjunction class, and (7b) places the ADV-"quando" in complementiser[71] position otherwise typical of conjunctions or relatives.

[71] In this text, “coplementiser position” is the clause header field which is obligatory in Portuguese finite and averbal subclauses, and optional in non-finite subclauses. “Complementisers” are the items able to fill this position, subordinating conjunctions, relative adverbs and interrogative adverbs. Contrary to some Portuguese grammar traditions, tbe notion of complementiser is *not* restricted to “completive” (substantival) finite subclauses typically

I would therefore argue that the distinction be handled not as one of word class (ADV/KS) but of semantico-syntactic subclass (<interr> and <rel>) and syntactic function (complementiser or not).

(7a) Quando <interr> @ADVL> vem?
(7b) Não sei quando <interr> @ADVL> @#FS-<ACC vem.
(7c) Venha quando <rel> @ADVL> @#FS-<ADVL quiser!
(7d) Aconteceu no dia quando <rel> @ADVL> @#FS-N< nasceu.

On the syntactic level, in a CG system, one must distinguish between the structural ambiguity encountered in the text itself, and the multiply mapped dependency and functional ambiguity introduced and then disambiguated as a natural and usually unavoidable intermediate step in CG-based parsing. While the ambiguity involved in the latter is "temporary" and designed for disambiguation, the former type of - text immanent - ambiguity is "true ambiguity" from a purely syntactic point of view, - and much harder to resolve. Small and lexically idiosyncratic clues have to be exploited, and world knowledge as well as a larger-than-a-sentence text window may well be necessary for full resolution:



Syntactic ambiguity
(analytical ambiguity)

constituent identity

cohesion

(1) syntactic form
(attachment ambiguity)

o homem ***com a bicicleta*** ***<u>da China</u>***
('[The man with the bicycle] from China' - 'The man with [the

(2) syntactic function

um homem ***que*** *ama* ***toda mulher****.*
('a man who loves every woman' - 'a man every woman loves')

(3) anaphora
(linking function)

Amava sua irmã
('he loved ***his/your*** *sister')*

(4) co-ordination
(linking form)

homens ***e*** *mulheres no Brasil*
('[men and women] from Brazil' - 'men and [women from Brazil]')

Where at all resolvable within the universe of one isolated sentence, ambiguities on the syntactic level are typically addressed by exploiting lexical information about valency patterns, basic word order probabilities and chunking information from punctuation, complementiser words or - less important - agreement links. This holds

headed by a conjunctional 'que'. Rather, the concept extends to relative ("attributive") and adverbial subclauses, as well as to averbal subclauses.

both for the simple ambiguities introduced by the CG mapping module (not focused upon in this section) and the more difficult text immanent ambiguities:

(8a) Falava VFIN com PRP o DET homem N que <rel> SPEC ama VFIN.
(8b) Falava VFIN com PRP o DET homem N que <rel> SPEC ama VFIN outra ADJ mulher N.
(8c) Falava VFIN com PRP o DET homem N que <rel> SPEC outra ADJ mulher N ama VFIN.
(8d) Falava VFIN com PRP os DET homens N que <rel> SPEC outra ADJ mulher N amam VFIN.

In the example sentence (8a), in its input form to the syntactic module, *que* has already been identified as relative pronoun by the *morphological* CG-module, and can therefore be used for chunking - it identifies an important piece of syntactic form, the break between main clause and subclause, and - being a relative - even suggests the subclause's function: postnominal modifier (@#FS-N<). But, since subjects are optional in Portuguese, *que* is still ambiguous *clause-internally,* between subject (@SUBJ) and direct object (@ACC). Knowing from the lexicon that 'amar' is preferably monotransitive, the parser opts for the @ACC reading. In (8b), valency is not enough to make the choice, since two NPs are present. For non-ergative words, however, Portuguese prefers preverbal subject position (cp. 8c), so *outra mulher* is ruled out as subject in (8b), and the uniqueness principle makes *que* the subject of the subclause. Another word order rule states that non-pronominal objects do not normally precede the verb, which is why *outra mulher* in (8c) is read as @SUBJ. Still, the @ACC-reading can be forced by interference from an agreement rule, like that of number-agreement between subject and finite verb (8d).

Ideally, for the sake of notational consistency, both form and function of syntactic constituents should be disambiguated in all cases. Sometimes, however, it is advantageous to underspecify one of the two. The syntactic form ambiguities of hierarchical (example 1) and co-ordinated (example 4) postnominal attachment, for instance, can be tackled by agreement rules in the case of adjectival modifiers, but not for PP-modifiers. It does not make sense to introduce ambiguity on a level of analysis where it cannot be resolved, and CG's flat dependency grammar provides a kind of "structural portmanteau"-solution, providing an unambiguous function reading (postnominal) in combination with an ambiguous attachment reading (left attachment, <):

(9a) o homem com @N< a bicicleta da @N< China.
(9b) homens e @CO mulheres no @N< Brasil.
(9a') o homem com @N< [a bicicleta de @N< alumínio].
(9b') homens e @CO [mulheres na @N< menopause].
(9c) um homem com @N< [formação em @N< direito].

Sometimes, valency (9c) or semantics (9a', 9b') can provide a clue, that could be exploited by a tree transformation application. For certain other applications, like machine translation from Portuguese into English or Danish, this particular kind of *dependency* underspecification does not pose problems.

Underspecification of syntactic *function,* too, is very common in all parsers, simply because there is no obvious limit to the degree of "delicateness of syntax" one might want to introduce. In general, Constraint Grammar adopts the pragmatic and methodologically logical solution of viewing ambiguities as strict surface phenomena (Karlsson et. al. 1995:22). In my parser, for instance, the distinction between complement and modifier postnominal PPs is not made explicit in the PP's syntactic tag (which is @N< in any case), but only inferable from the preceding noun's valency tag. However, the line between surface and deep structure, as between syntax and semantics, is not an easy one to draw - and one may well end up defending a “deep/semantic” distinction in the name of surface syntax in one place, and omitting it in another.

The reason why the syntactic underspecification problem needs mentioning, is obvious from my annotation scheme: On clause level, both dependency and syntactic function are specified (e.g. @<SUBJ, @SUBJ>), while on group level dependency takes over, and head-dependent relation (with the head marked at the tip of the dependency marker arrow) is the only function there is (e.g. @>N, @N<, @>A, @A<, where the function marker's place at the base of the dependency marker arrow is left empty). Yet, in a language without case marking of nouns and without fixed word order, even the subject-object ambiguity is not entirely a surface syntactic problem, and rules have to exploit both valency potential and semantic distinctions like ±HUM.

With clause functions in mind, one could, therefore, argue that *genitivus subjectivus* (“a promessa da mãe”) and *genitivus objectivus* (“a promessa de ajuda”) are syntactic categories (to be marked functionally, @N<SUBJ, @N<ACC) rather than (not to be marked) semantic ones (cp. thematic roles in the next section).

Even more than genitivus subjectivus and genitivus objectivus, postnominal participle-clauses (cp. chapter 4.4.4.2), ablativus absolutus (cp. chapter 4.4.4.1) and NP-AP nexus structures after 'com/sem' (cp. chapter 4.4.2, example (5)) are examples where the otherwise clear distinction between the clause and group levels becomes fuzzy, which is why I have here opted for the more specific notation and introduced - minimal - function markers (e.g. @A<PIV, @A<ADV, @N<PRED) instead of the “naked” dependency markers (@A<, @N<).

Another case of syntactic underspecification concerns determiner pronouns (like in *amava* ***sua*** *irmã*), which in principle have <u>two</u> syntactic links, both cemented by agreement rules[72], one being that of prenominal (@>N), the other that of "possessive", which in the case of the third person possessive may attach reflexively to the subject (for clause constituents) or to the NP-head (for group constituents), or to some referent outside the sentence[73]. The parser specifies the second link only indirectly by means of a simple secondary tag (<1S poss> or <3S/P poss>). In the

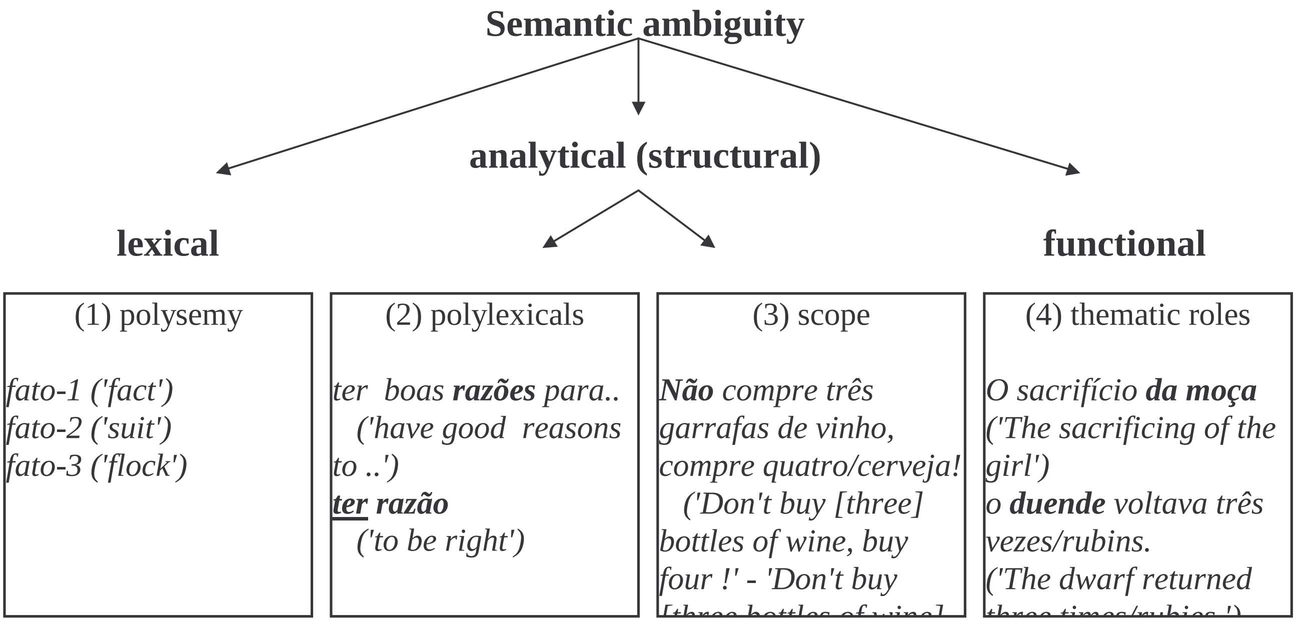
[72] In Portuguese, the categories gender and number have agreement with the modifier head, and the category of person with the possessor.

[73] Always evolving, Brazilian Portuguese does now have a colloquial language alternative for this case: the terms *dele, dela, deles, delas* (literally: "of him, of her, of them") in postnominal position. *Seu, sua, seus, suas* can then be reserved for the reflexive case ('his/her/their own') and the polite 3.person addressing pronouns (= 'de você' [your]).

case of a 3.person possessor, the underspecification with relation to gender and number can prove a problem when trying to translate into languages where possessives do have gender/number agreement with the possessor. The same problem arises for subject-less Portuguese clauses with a finite verb in the 3.person. In both cases, the translation module runs a "subject gender/number" counter, that helps resolve the ambiguity of possessives or verb-incorporated personal pronouns in subclauses or subsequent main clauses[74].

Working upwards, the next disambiguational distinction to make is the semantic one. Since most lower level ambiguities have semantic consequences, and semantics needs a textual vehicle anyway (either lexical, inflexional or syntactic), I will discuss semantic ambiguity along lexical, analytical and functional lines:

[74] This is the only case, where the parser uses an analysis window larger than one sentence (which is the default window for all CG-based modules).



fato-1 ('fact')
fato-2 ('suit')
fato-3 ('flock')

ter boas ***razões*** *para..*
('have good reasons to ..')
ter razão
('to be right')

Não *compre três garrafas de vinho, compre quatro/cerveja!*
('Don't buy [three] bottles of wine, buy four !' - 'Don't buy [three bottles of wine]

O sacrifício ***da moça***
('The sacrificing of the girl')
o ***duende*** *voltava três vezes/rubins.*
('The dwarf returned three times/rubies.')

The above sequence of semantic ambiguity types mirrors and supplants, in a way, what has been said about ambiguity on lower levels: Thus, the lexical level of polysemy corresponds to homonymy, and more specifically, to lexeme category ambiguity (type 1 of morphological ambiguity), while the analytical and functional types mirror the corresponding syntactic ambiguity classes of syntactic form and syntactic function. Of course, lower level distinctions can imply higher level ones (this upward implication is one important aspect of progressive level parsing, downward application of lexical categories being another one), as shown for the intermediate level *syntactic word classes*. Thus, treating the semantic ambiguity types 2-4 as syntactically inspired, one might call thematic roles semantic arguments, polylexical meaning could be regarded as a side effect of a very closely knitted syntactic relation, and scope could be described as the semantic result of operator attachment. Conversely, the cohesion section of syntactic ambiguity (anaphora and co-ordination) might be seen as a syntactic description of semantic structure.

In the three diagrams above, the difference between the semantic ambiguity level and the two lower levels is that, for polysemy, polylexical meaning, scope and thematic roles, none of the existing morphological or syntactic tags can capture the ambiguity in a principled way on the word itself, since both (semantic) readings will receive the same *lower level* analysis:

(10)
type 1: **fato** N M S
type 2: ter **razão** @<ACC
ter boas **razões** @<ACC para ..
type 3: **não** @ADVL> compre três garrafas de vinho, compre quatro/cerveja !
type 4: o sacrifício **da** @N< moça
o **duende** @SUBJ> voltava três vezes/rubins

The fact that these semantic distinctions are not "taggable" on the lower levels, does not, however, mean that lower level tag *context* is without disambiguational power. While the polysemy of 'fato' has to be resolved solely by using semantic discriminators (abstract countable <ac>, clothing <tøj>, group of animals <AA>) and semantic context (cp. 6.3.1), other polysemous words can be disambiguated by morpho-syntactic means:

(11a)	mais **ar**	<cm> (concrete mass noun)	'air'
(11b)	um **ar** de	<anfeat> (anatomical feature) <+de> (@N< argument)	'flair'
(11c)	boas **ar**es	<ac> (abstract countable) <smP> (plural noun)	'climate'

In (11), the prenominal context answers the question of the word's countability, *mais* ('much') in the negative, the numeral *um* ('one') and the plural *boas* ('good' P) in the positive, matching the mass noun tag <am> and the countable tags <anfeat> and <ac>, respectively. In (11c), a morphological feature of the word itself (plural: P) accomplishes the same thing. Finally, for differentiation between (11b) and (11c), nominal valency (<+de>) is used, matching the postnominal PP context in *um ar* ***de santo*** ('an air of holiness').

In the second type 4 example, the thematic role of 'duende' (AG or PAT) could be deduced either from the verb's valency instantiation (transitive or ergative) or the @ACC/@ADVL function of 'rubins' /'vezes', respectively. Of course, to provide this kind of valency or argument information, other semantic information may be necessary, like - in this case - knowledge about the time-class membership of 'vezes', and the concrete object feature of 'rubins' (which, in fact, both happen to be marked in the system's lexicon);

For the incorporating verb example (type 2) it is important, that the inflexional form of incorporated nouns is lexically fixed, and directly adjacent to the incorporating verb, not allowing for adnominal modifiers or arguments. This is lexicalised by different lexicon entries for *razão:*

(12)

razão#=#<VNP.acc>######<dar+><+a-piv>#42712
__ give {ngn} ret ('to concede that sb is right')
razão#=#<VNP.acc>######<ter+>#42722
__ have ret ('to be right')
razão#=#<sf.cause>######<am><ak><+para><+para+INF>#42707
__ <cause><+para><+para+INF> grund, årsag ('reason, cause')
__ <am> fornuft ('reason')
__ <ak> (ma) forhold, proportion ('proportion')

Here, the distinction is made by assigning a hybrid PoS to the incorporated noun: VNP, and tagging it for its incorporating verb (<ter+>). Disambiguation of the full noun in *ter boas razões para* relies on the word's inflexion (plural P), its -1 context (not 'ter') and its right hand argument context ('para'), all of which interfere with the VNP reading.

Scope ambiguities, like the above type 3 example, are structural to such a degree that they are very hard to tag on any word, even when using semantic tags. Only global utterance context and full “knowledge of the world” allow to decide whether the negation should be applied to the number of bottles, *três*, or the type of beverage, *vinho*. Luckily, scope operators are often placed directly before the entity they operate on:

(13a) Compre [**ao menos @>A** três] garrafas de vinho!
(13a) Compre [**ao menos @>N** [três garrafas]] de vinho!

Thus, in (13), *ao=menos* ('at least') can apply to either the numeral *três*, or the NP *três garrafas*, and in either instance attachment structure would optimally have to be tagged syntactically (@>A and @>N, respectively), mirroring semantic scope structure.

In contrast to scope relations, the thematic role ambiguity of type 4 could easily be explicited by word-based tags, - in the postnominal case by adding "clause function" in the same way used for participle clauses, as in the following example of “true” ambiguity (14):

(14a)	o sacrifício da @N<SUBJ moça	(genitivus subjectivus)
(14b)	o sacrifício da @N<ACC moça	(genitivus objectivus)

Alternatively, one might argue that a modifier/argument distinction (e.g. @N< vs. @N<ARG) would be enough, since the meaning of the postnominal *de* in (14a) is similar to a kind of default possessive meaning of *de*, which is compatible with almost any head noun, while the "object" meaning of *de* in (14b) asks for the right valency potential on the part of preceding noun.

In the case of thematic role marking, semantic function tags like the following could be used:

(15a) o duende @SUBJ> **@*PAT** voltava <ve> três vezes @<ADVL. (patient)
‘The goblin returned three times.’

(15b) o duende @SUBJ> **@*AG** voltava <vt> três rubins @<ACC. (agent)
‘The goblin returned three rubies.’

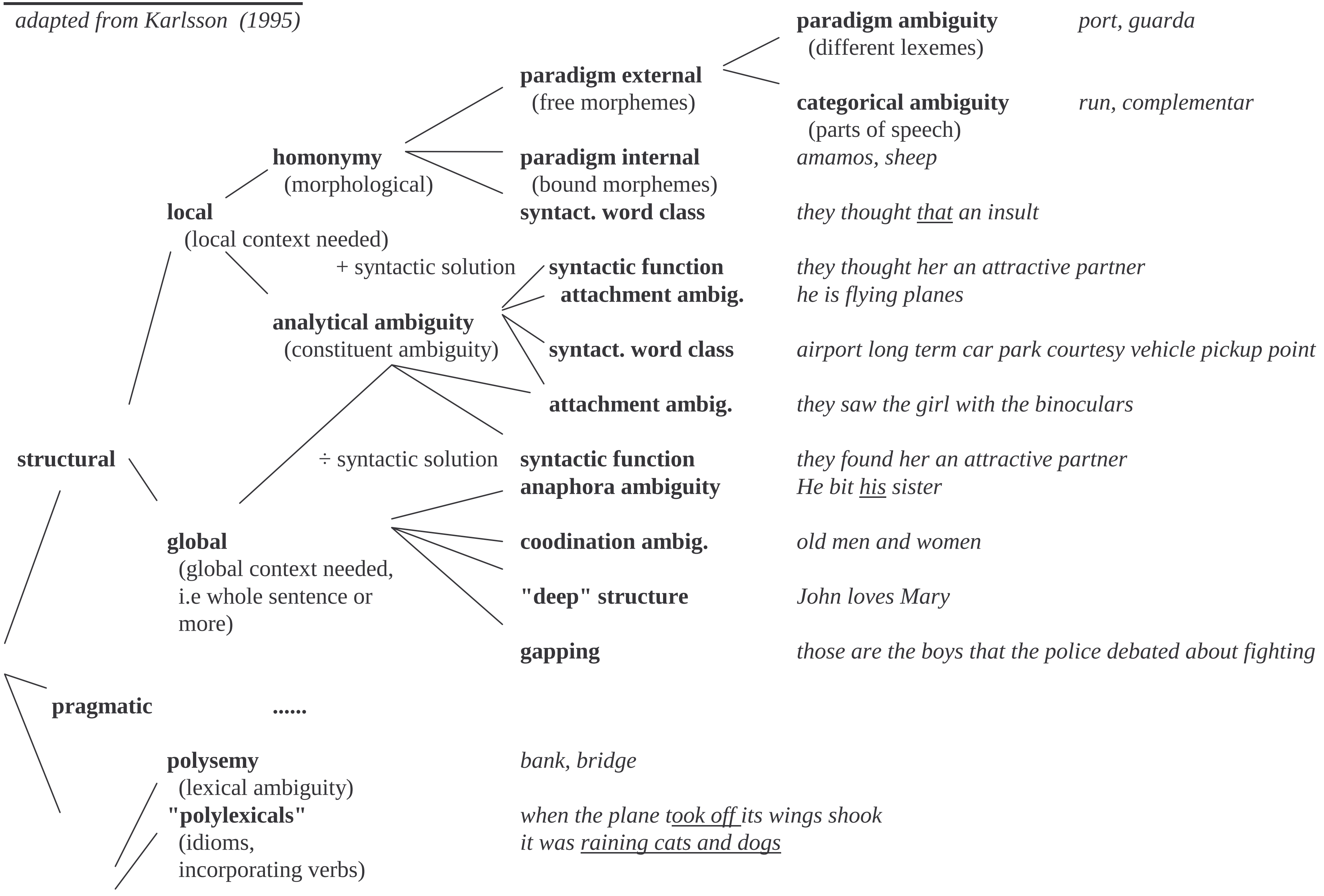
The @*-tags in (15) could, in principle, be mapped and disambiguated just like syntactic tags, profiting from a new round of CG-rules constituting a new (intermediate) level of syntactico-semantic analysis. The necessary information is already available in the present parser: In (15a) the main verb valency is instantiated as <ve> (ergative), because *vezes* prefers an adverbial reading over the direct object reading, while the main verb in (15b) is disambiguated as <vt> (monotransitive) with *rubins* being its direct object. Since ergative verbs have *patient* subjects, and transitive verbs have *agent* subjects, the correct thematic role tags can now be easily inferred.

In (Karlsson et. al, 1995:19ff) an interesting disambiguation oriented (and thus CG-near) view on ambiguity is presented: Ambiguity is here classified according to how much context is needed in order to resolve it (i.e. with CG type rules). The resolvability criterion is applied to structural ambiguities in particular (rather than meaning or pragmatic ambiguity), yielding local ambiguities on the one side, which can be addressed by drawing only upon local sentence context, and global ambiguity on the other side, where sentence-transcending context would be needed for full disambiguation.

Analytical (syntactic) ambiguities can be found in both groups (cp. the "resolvable" *they thought her an attractive partner* to the "unresolvable" *they found her an attractive partner*), whereas homonymy (morphological ambiguity) belongs almost entirely in the realm of locally resolvable ambiguity.

TYPES OF AMBIGUITY

adapted from Karlsson (1995)



paradigm ambiguity (different lexemes) *port, guarda*

paradigm external (free morphemes)

categorical ambiguity (parts of speech) *run, complementar*

homonymy (morphological)

paradigm internal (bound morphemes) *amamos, sheep*

local (local context needed)

syntact. word class *they thought that an insult*

+ syntactic solution

syntactic function *they thought her an attractive partner*

attachment ambig. *he is flying planes*

analytical ambiguity (constituent ambiguity)

syntact. word class *airport long term car park courtesy vehicle pickup point*

attachment ambig. *they saw the girl with the binoculars*

structural

÷ syntactic solution

syntactic function *they found her an attractive partner*

anaphora ambiguity *He bit his sister*

global (global context needed, i.e whole sentence or more)

coodination ambig. *old men and women*

"deep" structure *John loves Mary*

gapping *those are the boys that the police debated about fighting*

pragmatic **......**

polysemy (lexical ambiguity) *bank, bridge*

"polylexicals" (idioms, incorporating verbs)

when the plane took off its wings shook

it was raining cats and dogs

meaning-

scope (quantifier, negatives)	*someone loves everybody*
thematic roles (e.g. genitives)	*the door opened a few inches* *the shooting of the hunters*

An important aspect of the local-global resolvability distinction is that it can be seen as dynamic: some PP attachment ambiguities may be moved from global (i.e. unresolvable) to local (i.e. resolvable) by using better rules and more lexical information, for instance, nominal valency classes.

As a matter of fact, as long as enough lexico-semantic information is provided for the CG-rule to work on, I even believe that the local-global distinction can be applied to *meaning* ambiguity as well. Thus, in my parser, I have been able to resolve certain types of polysemy and verbal incorporation locally, i.e. through sentence context alone (cp. chapter 6). Some other meaning ambiguities, like thematic roles, might prove local, once they are introduced into the tagging scheme, and can be addressed by contextual rules.

3.1.2 Why tags? - The advantages of the tagging notation

All Constraint Grammar (to date) is implicitly tag-based. In fact, by extending the use of tags to the realm of syntax, Constraint Grammar has effectively widened the horizon of what traditionally (in HMM-analysers) was understood as tagging. Specifically, the term ‘tag’ in grammatical analysis will here be used to designate any word based (word-attached) alphanumeric string bearing meta-information about the word’s form and function. Tag notation is *not* some kind of necessary evil stemming solely from the CG-formalism’s needs, but has a number of important advantages in its own right:

- 1. Information from all levels (morphology, syntax, semantics etc.), both form and function, can be represented in the same formalism, and interact in the disambiguation process.
- 2. Tags can be combined/juxtaposed graphically as a text line after the word form, without confusing parenthesis hierarchies or the like, while also being easier to manipulate in a data-linguistic context (especially after the text has been "verticalised").
- 3. Tags make it easier to express ambiguity without graphically or structurally breaking the sentence context in an analysis. Thus, in an alternative sentence reading, it is not necessary to repeat those parts of the sentence that are *not* ambiguous. The longer the sentence, and the less restrictive the grammar, the bigger the advantage will be.
- 4. Disambiguation is not an "either-or"-process, and can be accomplished gradually by eliminating incorrect tags. This way the process has a high tolerance of both incomplete grammars and incomplete (or grammatically wrong) sentences, making the system a very robust one. Output like "no parse" or "time out", as known from classical generative grammar, is virtually unthinkable.

- 5. Tags can be integrated as meta-information in running text. This is an important advantage for the user-friendliness of tagged corpora and the versatility of searching tools.
- 6. Tags are easy to evaluate statistically and facilitate lexicographic corpus research.
- 7. Word based information is pedagogically transparent, and anticipates especially children's intuition about grammatical structure. In principle, tags can be presented not only as attached text, but also as colour-markings, underline highlights, subscripts etc. (cp. chapter 7.2).
- 8. Grammatical information in tag-notation can easily be filtered into different annotation schemes by standard text processing tools. It is easily accessible to secondary application programs.

Traditionally, due to these special aspects, the tagging notation has appealed to only a certain section of the linguistic community. The table shows typical target areas and the role of tagging in a number of corpus annotation projects.

Users and non-users of word based tagging

user	*non-user or partial user*
traditionally used for morphology: lexicon -> morphology -> word class	traditionally used for syntax word class -> syntax
analytical applications	generative applications
Many big corpora (BNC, LOB, Bank of English) Corpus-researchers (searching tools) Hidden Markov Models TWOL (two level morphology) Constraint Grammar Probabilistic parsers (e.g. CLAWS, PARTS, de Marcken)	Some hand checked corpora (Suzanne, Penn Tree Bank) Generative linguists DCG, PSG, GPSG, HPSG - parsers (e.g. ALVEY, TOSCA)

3.2 Morphological ambiguity in Portuguese

3.2.1 Overall morphological ambiguity

In order to quantitatively assess the ambiguity problem in Portuguese, before writing disambiguation rules, I ran the morphological tagger on two larger chunks of corpus, accessible to me at the time:

(a) a 630.000-word ECI-excerpt from the Borba-Ramsey corpus of written Brazilian Portuguese
(b) a 132.000 word corpus derived from the on-line data base of Brazilian literature in São Paulo (Rede Nacional de Pesquisa)

Table (1) shows the number and percentage of word form tokens with 0, 1, 2 ... 20 readings. The 1-readings row contains the figures for unambiguous cases, the 0-readings row covers recall failures.

(1) **Table: morphological ambiguity in Portuguese**

Number of readings	**Number of word form tokens**		**%**		**cumulative %**	
	mixed	**literature**	**mixed**	**literature**	**mixed**	**literature**
0	2108	479	0.3	0.4	0.3	0.4
1	290131	62527	46.1	47.4	46.4	47.7
2	149148	30860	23.7	23.4	70.1	71.1
3	74142	15075	11.8	11.4	81.9	82.5
4	81732	17126	13.0	13.0	94.9	95.5
5	23837	4209	3.8	3.2	98.7	98.7
6	6582	1437	1.0	1.1	99.7	99.8
7	1043	159	0.2	0.1	99.9	99.9
8	520	79	0.1	0.1	100.0	100.0
9	9	1	-	-	-	-
10	37	15	-	-	-	-
11	16	2	-	-	-	-
12	23	6	-	-	-	-
13	4	0	-	-	-	-
14	5	0	-	-	-	-
15	6	3	-	-	-	-
16	5	1	-	-	-	-
17	1	1	-	-	-	-
18	1	0	-	-	-	-

19	0	0	-	-	-	-
>= 20	15	1	-	-	-	-
total	**629364**	**131981**	100.0	100.0	100.0	100.0
propria-heuristics	10372	2112	1.6	1.6	-	-
orthographic intervention	491	125	0.1	0.1	-	-

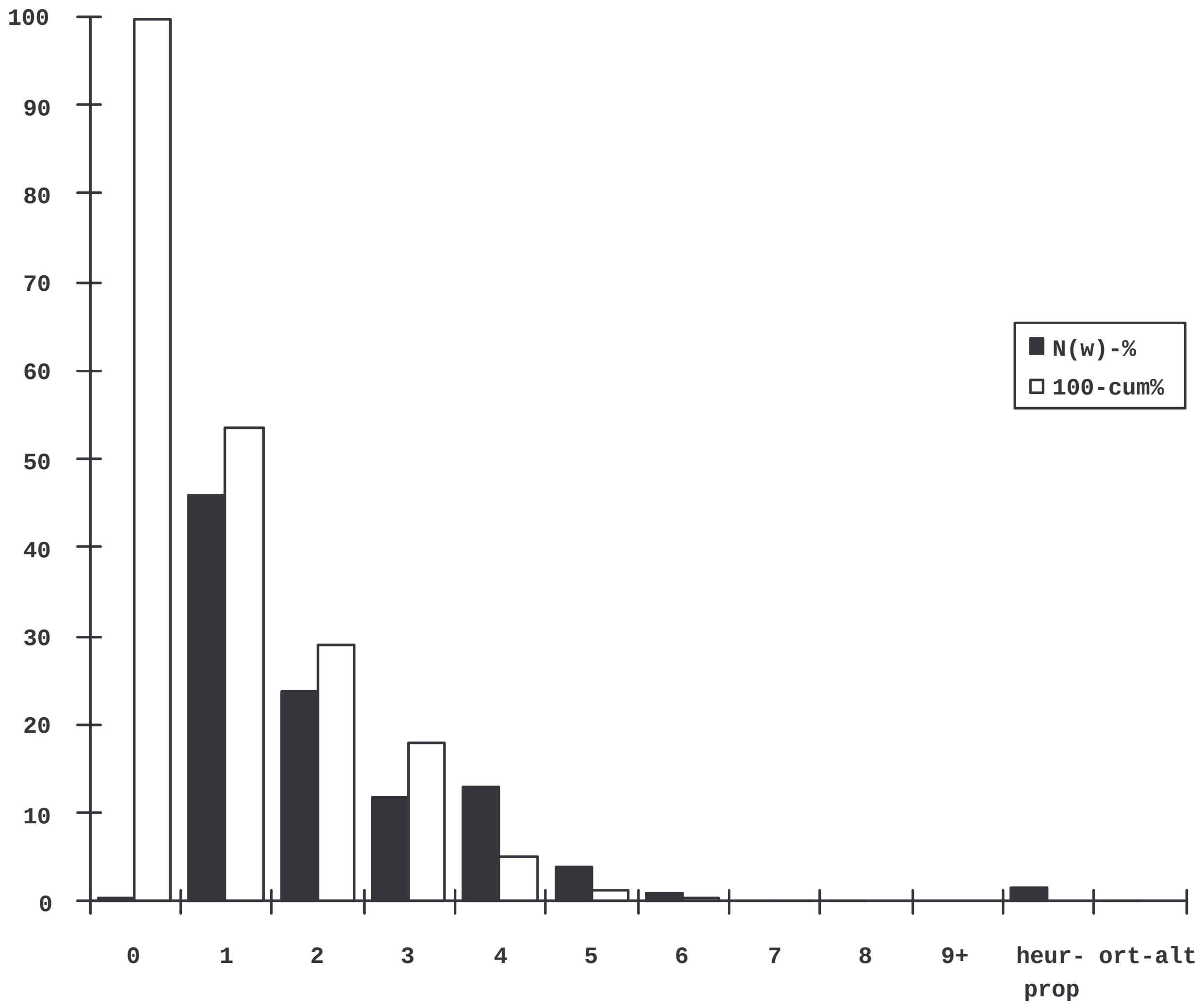
Though almost all words can be analysed, more than half the word form tokens get more than one reading, the average being 2.0 - 2.1[75].

The cumulative percentage column shows the proportion of word forms having n *or fewer* readings. The graphical representation in (2) maps *inverse* cumulation, showing the proportion of word form tokens with n *or more* readings.

(2) **Morphological ambiguity in Portuguese**
(in an excerpt from the Borba-Ramsey Corpus)

[75] These are the figures without the use of portmanteau tags. Later, portmanteau tags were introduced for 3 cases of verbal inflexion: the 0/1/3S tag for infinitives, the 1/3S tag for some present subjunctive cases, and the PS/MQP tense tag for some cases of 1.person and 3.person plural endings. The figure for the new version is, accordingly, lower: 1.7 readings per word form.

The missing portmanteau-tags and the ensuing close lumping of certain tags is also the reason for the strange relative ambiguity peak at the 4-readings mark.



% word form tokens with n readings: N(w)-%
% word form tokens more than n-ways ambiguous: 100-cum%

Highly ambiguous words with more than 5 readings are very rare, cumulating to roughly 1%. Very high ambiguity is usually a symptom of derivational complexity, where every word class or inflexion reading can again be ambiguous with regard to the derivational path assumed (prefix & suffix or 2 suffixes?, noun or adjective root?).

Of the 0.3-0.4 % words lacking analysis, most are misspellings, quotations or loan words from other languages (mainly English, but also French, German and Latin), and "names" without capitalisation, e.g. pharmaceutical drug names (cp. chapter 2.2.6).

The RNP corpus contains both literature, secondary literature and a considerable portion of bibliographical information. Considering that the latter accounts for some text passages in English, French and Spanish as well as foreign language book titles, bibliographical abbreviations etc., a recall failure of 0.4% must be regarded as quite low, - and only one forth of this (0.1% or 134 tokens) consists of unanalysable

Portuguese non-name words. Apart from that, the ambiguity distribution is almost the same as in the mixed Borba-Ramsey Corpus (where the portion of unanalysed Portuguese words is higher, 0.2-0.3%, due at least in part to scientific and dialectal text contributions).

In 1.6% of all cases, a PROP tag was applied heuristically, - to capitalised words that could not be given another analysis without orthographical change (in mid sentence), or even after orthographical alteration (sentence initially).

3.2.2 Word class specific morphological ambiguity

In order to know where the CG-rules could be made to be most effective, or, in other words, for which cases it was worth the trouble to write a lot of rules, I was interested in getting a more detailed picture of Portuguese morphological ambiguity. For the closed word classes (PRP, KS, KC, IN, DET, SPEC, PERS, NUM) ambiguity classes can be taken directly from the lexicon, and it would in principle be possible to write rules for every single word. For the open word classes (N, ADJ, PROP, V, ADV[76]), however, a statistical approach seemed appropriate to assess the magnitude of the problem.

Table (1) shows the numbers for a 170.666 word VEJA newspaper corpus, containing 121.170 words (71%) that are assigned at least one open word class reading. The basis for measuring ambiguity was a version of the parser that uses certain 3 verbal portmanteau tags not used in 3.5.1, as well as some word internal disambiguation (cp. 3.4). The resulting reduction in overall ambiguity from 2.0 to 1.7 has to be borne in mind when comparing the word class specific figures below with the findings in 3.5.1.

I have split up the V class into finite verbs (VFIN) and three non-finite subclasses, INF, GER and PCP, both because they show a syntactically completely different behaviour, and because the non-finite classes with their well-defined ending ('ar/er/ir' for INF, 'ando/indo' for GER and ado/ido' for PCP) can be expected to show their own, narrow ambiguity pattern. That the latter is quite distinct from that of finite verbs, can be seen form the low numbers for VFIN-INF, VFIN-GER and VFIN-PCP ("verb internal") ambiguity, respectively. The somewhat higher figure for VFIN-INF is due to the fact that the Portuguese infinitive *can* be inflected - yielding ambiguity with future subjunctive readings.

(1) **Table: PoS-ambiguity class frequencies**

	N	**ADJ**	**VFIN**	**INF**	**GER**	**PCP**	**ADV**	**PROP**	**all ambiguous PoS pairs**
N	2188	9273	10959	766	6	2197	2057	1940	29386
ADJ		241	2369	113	9	2334	1168	916	16423

[76] This class does contain both a kind of "closed subclass" and the open class in '-mente', but is here treated as one.

VFIN			9185	3748	19	375	1079	540	28274
INF				11	0	0	0	26	4664
GER					0	0	0	1	35
PCP						88	16	23	5033
ADV							2670	33	7023
PROP								283	3762
all words:	69603	17950	30619	4970	903	5335	13938	11704	54633 121170

Since this statistical analysis ignores closed class, the overall ambiguity figures will obviously be lower than what is found for the language as a whole (about 1.7 readings pr. word form when using portmanteau tags, 2.0 when not). When also ignoring word class internal inflexion and subclass ambiguity (shaded), the 121.170 potential open class words get 155.022 *different* word class readings (about 1,28 pr. potential open class word form). In all, the text contains 170.998 open class readings (about 1,41 pr. potential open class word form). The remaining 0,3 readings pr. word form (to reach 1,7) can be accounted for as the sum of cross-group ambiguity between the closed and open word class groups, plus closed-class internal ambiguity.

As can be seen, the most common ambiguity is the N-VFIN class, followed closely by N-ADJ and VFIN-VFIN internal ambiguity. Of these, the first is syntactically most important, since an error here will cause additional errors in the syntactic tags. The risk of such error spreading is smaller for N-ADJ and very small for word class internal ambiguities like VFIN-VFIN.

Apart from sheer number, the importance of an ambiguity class must, however, be measured against the size of the word classes in question. Thus, N is a very large word class, so maybe this explains its ambiguity rating in absolute terms, - but how large is the ambiguity risk for, say, a noun in relative terms?

(2) **Table: relative frequencies for word class ambiguity**

WC2 **WC1**	**N** (%)	**ADJ** (%)	**VFIN** (%)	**INF** (%)	**GER** (%)	**PCP** (%)	**ADV** (%)	**PROP** (%)	**ambiguity index**
N	3.1	13.3	15.7	1.1	0.0	3.1	3.0	2.8	42.2
ADJ	51.6	1.3	13.2	0.6	0.0	13.0	6.5	5.1	91.5
VFIN	35.8	7.7	30.0	12.2	0.1	1.2	3.5	1.8	96.0
INF	15.4	2.3	75.4	0.2	0.0	0.0	0.0	0.5	93.8
GER	0.7	1.0	2.1	0.0	0.0	0.0	0.0	0.1	3.9
PCP	41.2	43.7	7.0	0.0	0.0	1.6	0.3	0.4	94.3
ADV	14.8	8.3	7.7	0.0	0.0	0.1	19.2	0.2	50.4
PROP	16.6	7.8	4.6	0.2	0.0	0.2	0.3	2.4	32.2
all									45.1

words:	40.8	10.5	17.9	2.9	0.5	3.1	8.2	6.9	71.0

Table (2) shows the relative risks of a word class WC1 word form to be WC1-WC2 ambiguous. The percentage given is the ratio between the frequency of this ambiguity class and the frequency of words with at least one WC1 reading: WC1&WC2/WC1. For example, 15.7% of all words with N readings are ambiguous with at least one VFIN reading. The isolated word class frequencies for the undisambiguated text are given in the last row (shaded, e.g. for N, 40.8%).

My ambiguity index is *not* a percentage, but the sum of all instances of different ambiguity pairs for a word class WC1 (given in the last column in table 1, i.e. for VFIN, 28.274, the sum of VFIN-N, VFIN-ADJ, VFIN-VFIN, VFIN-INF and so on), divided by the number of all VFIN candidate word forms (30.619). The resulting figure *looks* like a percentage, in fact, it is the sum of all percentages in one row, yet due to the fact that many word forms host several WC ambiguity pairs, this "sum" is somewhat higher than what would be the "real" percentage of ambiguous instances for that word class. The overall ambiguity index for open word class ambiguity (45.1) is calculated as the ratio between the sum of all WC ambiguity instances (equalling half the sum of the last column in table 1, minus the shaded boxes), divided by the number of open word class candidates.

Maybe the most striking result is the fact that nouns appear to be frequent but harmless, while adjectives and participles are rarer, but very likely to belong too another nominal[77] class, too. The reason for the latter is a semantico-etymological one - many participles tend to be treated lexicographically as adjectives, and many adjectives function as nouns, too. Since lexicography is often bilingually motivated, and word classes often defined functionally, adjectives like *dinamarquês* ('Danish') are also listed as nouns ('Dane'), though there is no morphological reason for this - even the lexeme category test fails, since these nouns often - atypically - possess gender inflexion like their adjective counterpart. In the case of ADJ-PCP ambiguity, the parser is set to routinely discard the ADJ reading, and only "remember" it for later translational purposes, by adding an <ADJ> tag. However, this is done *after* the tagging stage, though the full ambiguity is preserved in table (2).

The most dangerous case, however, are VFIN readings. Because of finite verbs' crucial role in syntactic mapping, the nearly 50% chance of VFIN-*nominal* ambiguity (N, ADJ, PCP, PROP combined) is disconcerting, which is why I will provide a short assessment of this particular disambiguation task *ante temporem*. Several morphologically different endings cases can be distinguished:

(3)

[77] 'Nominal' is here used as an umbrella term for the open word classes defined by number and gender (N, ADJ, PCP and, where relevant, PROP)

		VFIN	nominal group
3a)	'-o'	1S	M S
3b)	'-a'	3S, 1/3S	F S
3c)	'-s'	2S	P
3d)	'-ar'	1/3S FUT SUBJ	ADJ M/F S

The solution for these cases is text dependent. Many text types do not contain 1.person verb forms, and here, VFIN could be routinely discarded. However, my parser is meant to be able to handle *any* written text, so more complex disambiguation is appropriate, involving nominal group agreement and checking for personal pronouns.

In (3c) there is a tendency towards avoiding 2.person verb forms which have become all but non-existent in Brazilian Portuguese. (3d) is comparably rare, but difficult to tackle. (3b), finally, is the most problematic, since both the verbal and the nominal reading are very common. Worse, while a feminine article *"a"*, preceding the word form, might be a way to recognise NP-agreement, it is not in this case, since the article itself is multi-ambiguous, one reading being that of object pronoun, which in Portuguese is very common in front of finite verbs.

The 12% chance of confusing VFIN with INF is problematic for syntactic reasons, too. It involves the future subjunctive readings that are often crucial for the recognition of relative subclauses, a typical corollary error being the a wrong choice in the pronoun-conjunction ambiguity of *"se"*. The inverse case, INF vs. VFIN, is - quantitatively - even worse: 75% of all infinitive readings (virtually all regular infinitives) can also be read as finite future subjunctives.

The friendly cases are gerunds, which are both rare and morphologically well defined by the nearly unmimickable ending '-ndo', and proper nouns, that have the advantage of capitalisation marking, and only in sentence initial position pose certain, limited problems. In fact, part of the disambiguation load for the PROP class resides in the morphological analyser (tag assignment level), i.e. *before* the level table (2) is concerned with (cp. section 2.2.4.4).

As could be expected, the word class internal ambiguity is highest in finite verbs, due to the rich inflexional possibilities and stem variations.

(4) revista

"revestir" <vt> <de^vrp> <de^vtp> V PR 1/3S SUBJ VFIN ‘to cover’
"revistar" <vt> V IMP 2S VFIN ‘to review’
"revistar" <vt> V PR 3S IND VFIN ‘to review’
"rever" <vt> <vi> V PCP F S ‘to see again’, ‘to leak through’
"revista" <CI> <rr> <occ> <+n> N F S ‘magazine’, ‘inspection’

For nouns, the second largest group in this respect, class internal ambiguity is much lower, since its typical inflexions (the singular '-a' and '-o', as well as the plural '-s') are quite distinct, and few irregular exceptions exist. So most cases have to be lexical homonyms. A relatively common case are words in '-r' or 'l' which can cover 2 different lexemes, one masculine, one feminine (5a). Another possibility is lexicalised metaphorical use (5b).

(5a)

final
"final" <n> ADJ M/F S 'last'
"final" <occ> N F S 'finale'
"final" <cP> N M S 'end'

(5b)

cara
"cara" <anfeat> <sh> <topabs> <fazer+> <ter+> <+de> <H> N F S 'face'
"cara" <H> <Rare> N M S 'guy'
"caro" <+a> ADJ F S 'expensive'

A certain amount of ambiguity is even purely syntactic or semantic, like much of the ADV internal ambiguity where I have chosen to treat the relative (<rel>) and interrogative (<interr>) subclasses of words like *como, onde* and *quando* as distinct word classes, in order to achieve early disambiguation[78] (i.e., in this case, make syntactic class information available at the PoS tagging level). Another example is the topological - name ambiguity in *Salvador*, which can both be a place name (not allowing an article) *and* a personal name (allowing the definite article).

Only such word class internal *semantic* ambiguity has a chance to survive the tagger's disambiguation rule set, as the figures for the same VEJA text (6) show after complete analysis (i.e. including disambiguation)[79].

(6) **Table: PoS-ambiguity resolved**

	N	**ADJ**	**VFIN**	**INF**	**GER**	**PCP**	**ADV**	**PROP**	**all pairs**	**preci-sion**[80] (%)

[78] This is, of course, an exception, since secondary tags do not usually justify separate reading lines, and are not meant to be disambiguated at the morphological stage. However, the above distinction in complementizer adverbials is of great importance for the disambiguation of other - morphological - ambiguities, like the above mentioned FUT SUBJ vs. INF readings, as well as for syntactic mapping (FS versus ICL).

[79] Since the PoS error rate for automatic disambiguation is under 1% classes (cf. chapter 3.9.2) and fairly balanced between word classes, there is nothing wrong with using the tagger's output after disambiguation as a base line for measuring "disambiguation gain" in comparison with the ambiguity found before disambiguation.

[80] Here defined as the ratio of word forms and word form readings, *not* , as in Karlsson et. al. (1995), *correct* readings divided by *all* readings. The reason for my usage of the term is, that at nearly 100% disambiguation, the alternative definition of 'precision' doesn't make much sense, since it will be close to the recall figure, both of which I therefore combine as *correctness* (treated in 3.9). Recall *without* disambiguation (where it does make sense as an independent figure) is treated in 2.2.6).

N	107	-	-	-	-	-	-	-	107	99.7
ADJ		2	-	-	-	-	4	-	6	99.9
VFIN			14	-			2	-	16	99.9
INF				2	-	-	-	-	2	100.0
GER					-	-	-	-	-	100.0
PCP						13	-	-	13	99.7
ADV							9	-	15	99.8
PROP								10	10	99.9
all									163	99.8
before:	69603	17950	30619	4970	903	5335	13938	11704	121170	
after:	39394	9549	16023	4648	894	3818	8552	11522	94394	
decrease (in %):	43.4	46.8	47.7	6.5	1	28.4	38.6	1.6	22.1	
table 2 ambiguity index	42.2	91.5	96.0	93.8	3.9	94.3	50.4	32.2	45.1	

Cross word class precision is virtually 100% for all open word classes, with the only exception of the - not so open - adverb class (99.8%). But even when including word class internal ambiguity, precision is still as high as 99.8%.

Table (6) makes it clear, how huge the differences in "disambiguation gain" are for the different word classes, suggesting how and where it would be most economical for the grammarian to channel his rule writing effort. Very little is gained for proper nouns, infinitives and gerunds, while finite verbs, nouns and adjectives have a nearly 50% disambiguation gain. From this it is clear that it "pays more" to write CG rules aimed at the latter classes than for the first.

Even more striking is a look at the relations between ambiguity index and disambiguation gain: infinitives, for example, start as highly ambiguous word forms, but most cases are finally tagged as unambiguous infinitives anyway! For nouns, though not as ambiguous to begin with, the disambiguation tendency is even more lopsided, with an ambiguity index 20-times as high as the final disambiguation gain, meaning that there is a very strong bias in favour of the PROP reading in ambiguous cases. The most "profitable" situation is that encountered in nouns, where CG rules do most work: the gain percentage is about the same as the ambiguity index, meaning that nouns have no strong bias in their ambiguity distribution.

3.3 Borderline ambiguity: The limits of form and structure

In chapter 3.1.1 a number of cases of true ambiguity were presented, i.e. cases where sentence context is not large enough a window for full disambiguation, and - short of widening the context window to, say, paragraph size or using non-CG tools like pragmatic reasoning or scripts - portmanteau tags and dependency underspecification were suggested as tools fully compatible with and in fact elegantly supported by CG's tag-based "flat" notation.

	example	**ambiguity type**	**underspecification tool**
(1)	amamos	inflexional	portmanteau-tag (PR/PS)
(2)	homens e mulheres do Brasil	co-ordination	linear link (@CO)
(3)	o homem com a bicicleta da China	attachment hierarchy	left linear attachment (@N<)
(4)	o sacrifício da moça	bound - unbound	functionless attachment (@N<)

Ambiguities like the above are often cited as (linguistically) "interesting", and especially the structural ones, 2-4, are often the first input people come up with when asked to try out a new parsing system. Of course, the parser is usually trapped - it will either (if it is cautious) be criticised for not living up to human disambiguation standards, or (if it does make a choice) for not preserving true ambiguity.

In my view it is, however, pointless for a parser to specify ambiguity that it cannot resolve, - such ambiguity is best left to humans, implicit and waiting for pragmatic context.[81]

Though Constraint Grammar can produce structural analyses, it is not a structural tool as such, but rather an incremental context checking tool. Its power as a disambiguation tool can be increased incrementally by improving the lexicon or by adding more rules, without increasing the structural complexity - and ambiguity - of its description.

Constraint Grammar rules are very good at using tiny context clues, that are nearly unexploitable in PSG type rewriting rules, for example where the constituents to be linked by the context condition are disjunct and thus cannot be lumped into one, bigger, constituent, or where morphological or functional details have gone unmarked because they have no consistent impact on constituent structure.

[81] The real reason for the popularity of structural ambiguity among a large group of modern linguists is possibly not its cognitive or grammatical weight, but simply the fact that large quantities of this kind of ambiguity are an unavoidable side effect of a very fashionable disambiguation tool - phrase structure grammar (PSG) and its derivatives, where any increase in descriptive power unavoidably entails an increase in unresolvable ambiguity.

Consider the following "classical laboratory sentence" (5) with its ambiguity made explicit in CG- (5a) and PSG-terms (5b), respectively:

(5a) They saw ('see'/'saw') the girl with (@N< @<ADVL) the pair of binoculars.

(5b) They [saw(see) [the girl with the pair of binoculars]].
They [[saw(see) the girl] with the pair of binoculars].
They [saw(saw) [the girl with the pair of binoculars]].
They [[saw(saw) the girl] with the pair of binoculars].

The lexical ambiguity of *saw* together with the functional/attachment ambiguity of the PP *with the pair of binoculars* yields 4-fold ambiguity[82]. Note that the flat tag notation (5a) is capable of elegantly expressing this ambiguity in *one* string, while a traditional PSG (5b) would produce four trees or bracketed lines. Leaving aside the question of notational elegance, I would like to argue that (5) is not at all as ambiguous as it seems, not even with a mere sentence window, and can be tackled - provided the right tools for disambiguation. How?

Starting with the PP attachment ambiguity, the (morphological) feature of definiteness seems to make all the difference:

(6a) They **killed** the girl with @N< the gun. - What did they do?
(6b) They killed **a girl with @N< a gun**. - Who did they kill?
(6c) They killed the girl with @<ADVL **a gun**. - What did they kill him with?
(6d) They killed **a girl** with @<ADVL the gun. - Who did they kill with the gun?

Intuitively one would say that (6a) and (6b) have postnominal PP attachment, while (6c) and (6d) have ad-verbal (or ad-VP) attachment, the difference being, that in the first pair *girl* and *gun* have the same degree of definiteness, while they have different degrees of definiteness in the second pair. The secret of why definiteness can be used for disambiguation in this case, is topic-focus structure.

According to Togeby (1993), focus is the *last sentence constituent, that is not definite*. Topic material, by contrast, is normally known in advance (from the last sentence, or from extra-lingual context), and will therefore appear in definite form.[83] It follows from this that constituents will be assembled in the "receiving" mind according to matching definiteness or non-definiteness. Since topic and focus constitute *different* constituents, the constituents in question can be told apart by their definiteness or non-definiteness, respectively, which is why *the girl with the gun* and *a girl with a gun* are easily accepted as NPs, while *the girl with a gun* and *a girl with the gun* are not, - at least not as long as the argument of 'with' belongs to the semantic set of tools.

Returning to (5), it is to be deemed probable that it is the girl who has the pair of binoculars.

[82] Actually, the transobjective valency of 'see' also permits a fifth reading, that of object complement (@<OC) for the PP, as discussed later in this section.

[83] Togeby's model refers to Danish, but since language has a universal element of linearity, the model would seem appropriate for other languages with definiteness-marking, too

Given that a sentence like (5) is most likely to appear in fiction or direct speech, the second, lexical, ambiguity (between 'see' and 'saw') can be resolved by a heuristic CG-rule[84] that exploits the statistical fact that the distribution of present and past tenses is text-type-dependent, and prefers - in stories - past tense over present tense readings for main clause finite verbs without sentence-initial or -final quotes, or quoting speech-verbs:

```
(7) REMOVE (PRESENT) IF  (0 PAST)
                         (NOT *-1 VFIN)
                         (NOT @1 QUOTE OR CLB-ORD)
                         (NOT @-1 QUOTE)
                         (**1 @<SUBJ OR <<< BARRIER V-SPEAK) ;
```

The above rule highlights one of the major differences between corpus linguistics and “single sentence linguistics” - In a corpus, there is always context, and even if this context (due to the small size of the window of analysis) is not directly made use of, it still limits the range of meaning to the more common readings. Heuristic rules, at least, or CG-rules with C-(certain)-contexts[85], can therefore discard those readings that need a lot of artificially constructed corpus around them. It is semantic context in particular, that, for structural or lexical disambiguation to be feasible in the usual "laboratory examples", has often to be quite imaginative. However, in corpus linguistics, girls are never being sawed, and salt is only passed over the table, but never passed by in Utah. Therefore, another, semantic, road of disambiguation can be followed, too, - by defining semantic sets of words in their ordinary, "prototypical" uses (8) which are then drawn upon by semantic CG-rules (9):

<u>sets:</u>

(8a) LIST SEEING-TOOLS = "binoculars" "glasses" "looking-glass" "microscope" "telescope" ;
(8b) LIST CUTTING-TOOLS = <kniv> ; (knife-prototype)
(8c) LIST SAWABLE = <mat> ; (materials)

(8a) LIST SEEING-TOOLS = "binoculars" "glasses" "looking-glass" "microscope" "telescope" ;
(8b) LIST CUTTING-TOOLS = <kniv> ; (knife-prototype)
(8c) LIST SAWABLE = <mat> ; (materials)

<u>constraints:</u>

(9a) REMOVE ("saw") IF (*1C @<ACC LINK NOT 0 SAWABLE)
Discard 'saw', if the next safe following direct object is not a sawable
(9b) SELECT ("saw") IF (*1C PRP-WITH BARRIER CLB LINK 0 @<ADVL LINK NOT 0 @N< LINK *1 @P< LINK 0 CUTTING-TOOL)

[84] If the ambiguity wasn't lexico-inflexional, but purely inflexional and systematic, a portmanteau-tag would maybe be preferable within CG-philosophy, cp. the 1.person plural present tense and perfeito simples tense ambiguity in Portuguese.

[85] Cf. chapter 3.6 for a detailed description of the CG rule formalism used.

Choose 'saw', if within the same clause there appears the preposition 'with' with a sawing-tool argument and adverbial, but not postnominal, function.

(9c) SELECT ("see") IF (*1C PRP-WITH BARRIER CLB LINK 0 @<ADVL LINK NOT 0 @N< LINK 0 SEEING-TOOL)

Choose 'see', if there within the same clause appears the preposition 'with' with a seeing-tool argument and adverbial, but not postnominal, function.

Two of the semantic prototypes involved, 'knife' (8b) and 'material' (8c) have already been implemented in my Portuguese lexicon, the third can be fashioned in the grammar itself as a set of base forms (8a).

For (5), only rule (9a) will be applied, since the definiteness based topic-focus analysis assigns the PP a @N< tag, and not the @<ADVL tag necessary for the application of (9b) and (9c). (9c) would, however be useful in (10a) and (9b) in (10b).

(10a) They saw ('see') the girl with @<ADVL a pair of binoculars.
(10b) They saw ('saw') the log with @<ADVL a chain-saw.
(10c) They saw ('see') the girl with @N< the friends in high places.
(10d) They saw ('see') the girl with @<OC a friend.

Finally, on top of other features, valency can prove an important player in this case of ambiguity as well. Why is it that the instrumentality of the PP in (10b) sounds more convincing than the one in (10a)? Could it, in spite of the definiteness incompatibility, still be the girl that has the binoculars? I'd say this depends on the level of valency analysis: Substituting, for clarity, the non-instrumental 'friend' for 'binoculars', there is still a difference between the sentence with a definite PP-argument (10c) and the one without (10d), though in both cases it is the girl, who has a friend. But if we accept that the PP in (10d) cannot be a postnominal for lack of definiteness compatibility, what then can it be? The solution to the puzzles lies in the transobjective valency of the verb 'see', which is not shared by either 'saw' or 'kill', and can be seen in ACI constructions like *Can you see him climb the tree* or in semantic variations like *I saw her home.* Here, both *climb the tree* and *home* can be read as object complements (@<OC), a function that will also serve for discriminating (10d) from (10c).

3.4 Word internal (local) disambiguation

Sometimes an ambiguous word form is assigned readings of differing complexity, that is, some analyses are made up of more derivational elements than others. However, "Karlsson's law" of minimal derivational complexity[86] (Karlsson, 1992, 1995) claims that in such cases the cohort can be made less ambiguous by rejecting all but the least complex readings, which in almost all cases prove to be the contextually correct ones. Though the law was not specifically formulated for Portuguese, it seems to hold for that language, too[87].

When the morphological analyser program searches for analyses of a given word, it first looks for whole roots and inflexional endings (step 1) , then for suffixation with or without inflexion (step 2). Implementing Karlsson’s law, the program only progresses to step 2, if no readings are found at step 1. Suffixation itself is analysed iteratively with increasing "suffixation depth" for each round (step 2a: one suffix, step 2b: two suffixes etc.), maximum depth being 4 at the moment. Again, the process only goes on to the next round (depth), if no analyses are found. Thus the analysis cohort only contains the "shortest" readings, saving time and disambiguation effort.

Prefixation (step 3), though, is more problematic. Only undertaking step 3, if no analyses are found in step 1 and 2, would mean possibly neglecting a 2-element analysis with prefix and root only, just because the program already has found a - say - 4-element reading involving 3 suffixes. So, prefixation is done whenever suffixation has been done. For each prefix on the list step 1 and 2, too, are undertaken for the remaining part of the word. As before, depth is increased step by step if no analysis is found for that individual prefix, thus automatically discarding unnecessarily complex analyses. But, when searching for possible prefixes, the program has to look at all prefixes, because it cannot know in advance which particular prefix will yield the analysis with fewest elements, nor whether this analysis will be shorter than the shortest "suffixation only" analysis.

Therefore, after completed analysis, word internal disambiguation is undertaken summarily on the resulting cohort, discarding all readings that have more than the minimum number of derivation elements for that cohort.

When applied to the RNP literature corpus, local disambiguation - apart from obviously reducing overall ambiguity - has a peculiar "smoothing effect" on the ambiguity distribution curve by considerably lowering the percentage of 4-way ambiguous word forms (that previously had been higher than the one for 3-way

[86] Karlsson (1995) uses the term "local disambiguation" for this selection process, referring to the fact, that the rule concerned is applied to word forms in isolation, and does not make use of any context conditions whatsoever.

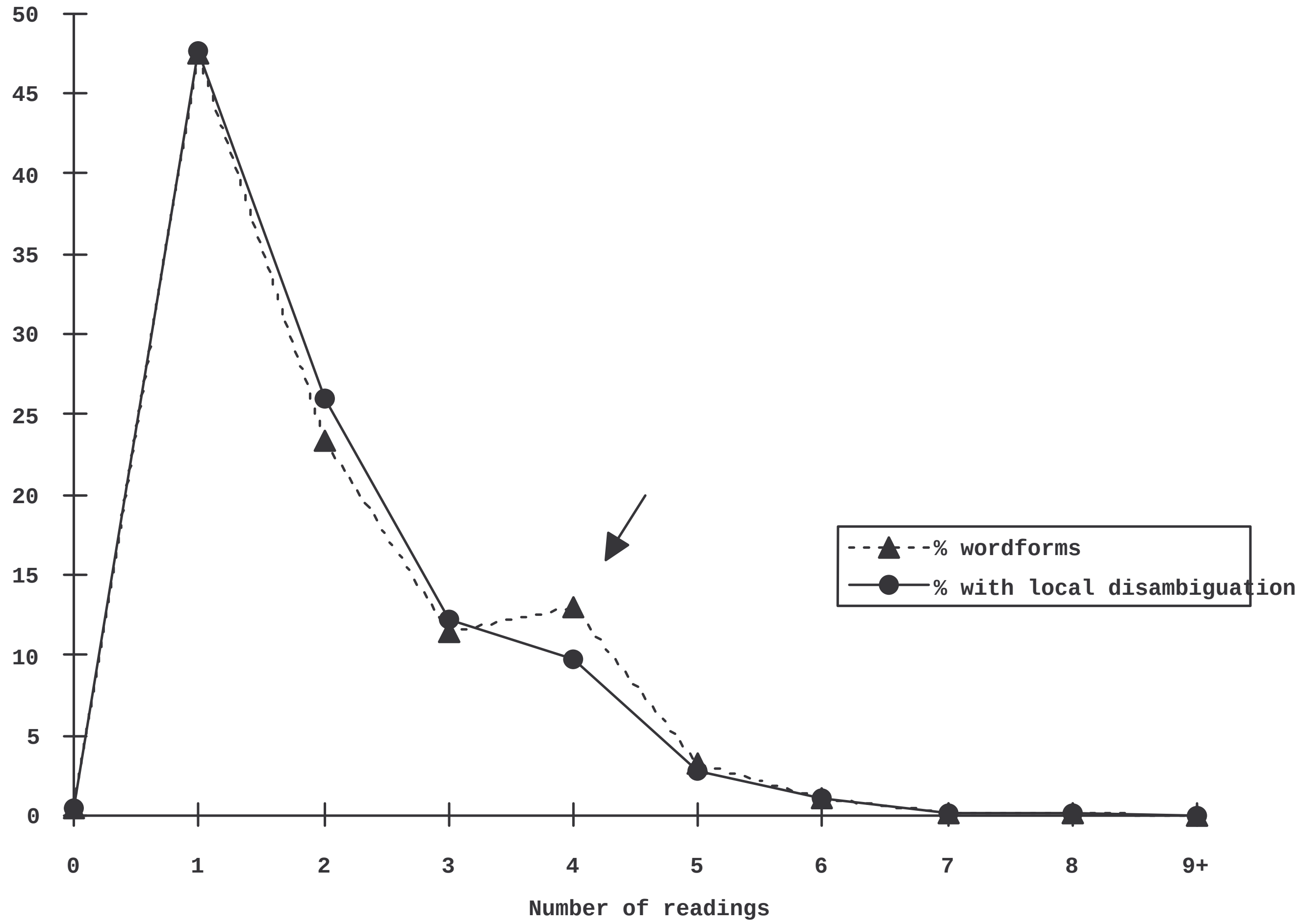
[87] The law was inspired by languages with productive compound formation, like Swedish and German, but can be extended to languages with few root compounds, as long as these languages have productive affixation, like English (Karlsson et. al., 1995) and, here, Portuguese. Though Karlsson's law is of a heuristic nature, it is all but impossible to find counter-examples.

ambiguous words), moving them to "lower" ambiguity groups. At the same time, there was nearly no effect on the very highly ambiguous word forms.

When judging the changes in the table below, one has to consider, however, that even the original percentages were arrived at by some degree of local disambiguation, in that my parsing program from the beginning has had successive level analysis for suffixes, so the changes involve only prefixed readings being rated against each other, and prefixed readings of low complexity substituting high complexity suffixation readings. The overall effect of local disambiguation may be even more visible when compared to an analysis technique with no "depth control" whatsoever.

(1) **Effects of local disambiguation by minimum derivation selection ("Karlsson's law"), data from the RNP literature corpus analysis**

number of readings	**number of word forms**		**% of all word forms**		**cumulative %**	
	without	with local disambig.	without	with local disambig.	without	with local disambig.
0	479	480	0.4	0.4	0.4	0.4
1	62527	62847	47.4	47.6	47.7	48.0
2	30860	34303	23.4	26.0	71.1	74.0
3	15075	16159	11.4	12.2	82.5	86.2
4	17126	12945	13.0	9.8	95.5	96.0
5	4209	4564	3.2	2.7	98.7	98.7
6	1437	1418	1.1	1.1	99.8	99.8
7	159	159	0.1	0.1	99.9	99.9
8	79	78	0.1	0.1	100.0	100.0
9	1	1	-	-	-	-
10	15	15	-	-	-	-
≥ 11	15	14	-	-	-	-
total	131981	131981	100.0	100.0	100.0	100.0



In the case of "derivational" ambiguity with equal depth, another tool for local disambiguation - not included in the above table - has been introduced: The tagger fuses readings of same derivational depth if their tag strings are identical, and the only difference is a word class difference in the *root* . Thus, in (2), the same '-ista' derivation is arrived at, departing from 4 different roots, with *parlamentar* lexicon-registered as both N and ADJ (since this difference is not visible in the base form tag, I have here retained the system internal lexeme identity numbers §...§).

(2)

```
parlamentaristas
  "parlamentar"  <attr> <DERS -ista [ADEPTO]> N M/F P   §37367§
  "parlamentar"  <attr> <DERS -ista [ADEPTO]> N M/F P   §37368§ ###
  "parlamentário"  <attr> <DERS -ista [ADEPTO]> N M/F P   §37370§ ###
  "parlamentarismo"  <attr> <DERS -ista [ADEPTO]> N M/F P   §37371§ ###
```

Finally, from a non-semantic perspective, the derivational path, telling *which* affixes have been used, is not important, as long as the resulting morpho-syntactic information is identical. Therefore readings with different paths (in fig. 3 '-ção' vs. '-ização'), but identical non-derivational tags (in fig. 3, N F S), may be fused.

(3)

```
        modernização
          "modernizar"  <DER:-ção (CAUSE)> N F S      §33620§
          "moderno"  <DER:-ização (CAUSE)> N F S      §33621§
```

Mixed prefix/suffix readings are preferred over double prefix or double suffix readings. In practice, this weeding out, unlike "true" local disambiguation (1), is first performed at the CG disambiguation level.

3.5 Tools for disambiguation

In corpus linguistics, most systems of automatic analysis can be classified by measuring them against the bipolarity of rule based versus probabilistic approaches. Thus, Karlsson (1995) distinguishes between “pure” rule based or probabilistic systems, hybrid systems and compound systems, i.e. rule based systems supplemented with probabilistic modules, or probabilistic systems with rule based “bias” or postprocessing. As a second parameter, lexicon dependency might be added, since both rules based and probabilistic systems differ internally as to how much use they make of extensive lexica, both in terms of lexical coverage and granularity of lexical information.

Typically, in terms of computational viability, probabilistic systems are good at lower level analysis, especially word class (part of speech, PoS) annotation and speech recognition, while rule based systems have been preferred for higher level annotation, like constituent trees and argument structure. As a result of this polarisation, the older - linguistically motivated - term "parsing", though derived from "pars orationis" (part of speech) has come to mean, more narrowly, higher level syntactic analysis, while the newer - computationally motivated - term "tagging" has mostly been limited to lower level PoS-annotation, - which is the obvious application for at least*word based* tags. Even implementationally, the bipolarity is quite distinct: The archetypal rule based systems, PSG grammars and their descendants, have embraced declarative programming languages like Prolog and Lisp, while probabilistic systems huddle together around the Hidden Markov Model using procedural programming languages like C or - for statistics proper - common UNIX-tools like sort, uniq, awk and perl.

With the advent of larger, multi-million word corpora, apart from annotation speed, error rates have become more crucial, since manual post-processing is becoming less and less feasible. On the one hand, this should favour rule-based systems, since they can - at least in theory - be made more "perfect", so the high initial price in man power for writing a grammar should pay off for large corpora - the larger the corpus the better the investment. On the other hand, large corpora supply better training facilities for the "cheap" probabilistic systems and should thus make them more accurate[88]. Yet again, since what is really needed, are *tagged* training corpora, co-operation between systems might be the best solution. This, however, presupposes more or less compatible category definitions and tag sets, which is, in spite of normalising initiatives like the EU's EAGLES convention (Monachini and Calzolari, 1996) far from being a reality today.

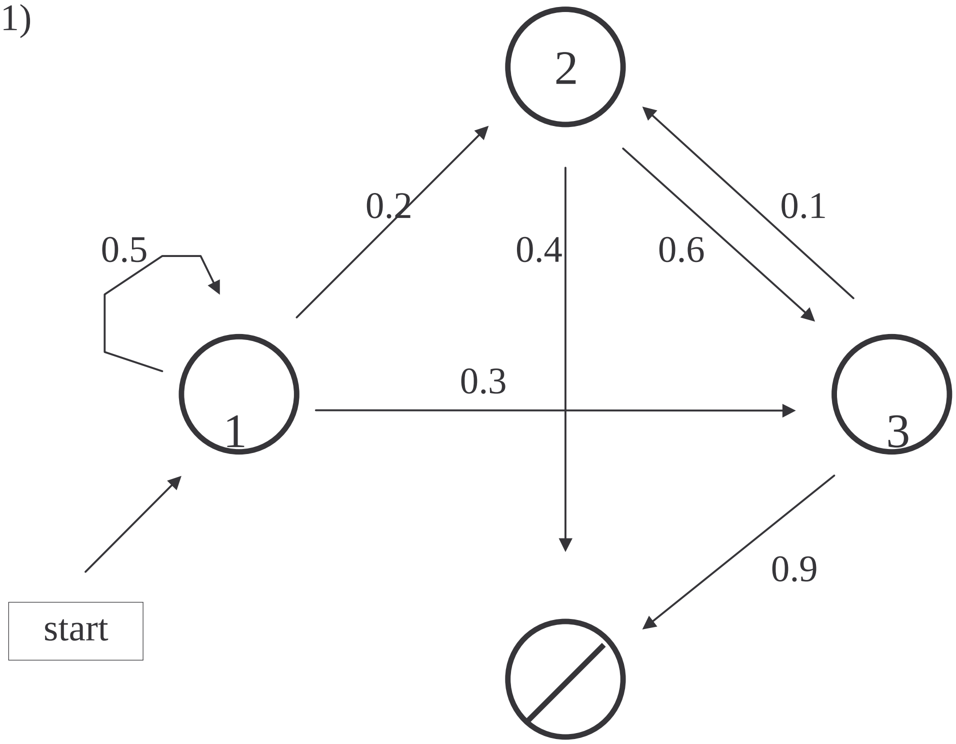
[88] For a tagset of 50 PoS-inflexion tags or tag chains, for example, it is as hard to train trigrams on a million word corpus as it is to train tetragrams on a 50 million word corpus, the reason being, that the number is only 8 times as high as the number of different n-grams. Training trigrams on a hundred million word corpus, however, yields on average 800 examples of each trigram combination - even when ignoring the relatively higher frequency of the more relevant trigrams -, which should be enough to do statistics on.

3.5.1 Probabilistics: The 'fire a linguist' approach

Most probabilistic NLP systems address part of speech tagging by automatic training and base themselves upon Hidden Markov Models (HMM), a mathematical model, where a surface-sequence of symbols is stochastically generated by an underlying ("hidden") Markov process with a state- and/or transition-dependent symbol generator.

A Markov Model consists of a finite number of states and describes processes (or sequences) as transitions (probability labelled arcs) between these states:

(1)



The MM in (1) has three states and a stop-state (Ø). When in state 1, for example, the MM has a 50% probability of staying there, a 20% probability of moving to state 2, and a 30% probability of moving to state three. The probability for a given sequence can be computed as the product of the individual transition probabilities. Thus, the sequence 1132 is assigned the probability 0.5 x 0.3 x 0.1 x 0.4 = 0.006. Since transition probabilities only depend on which state the process is in at a given point in time, such a MM is called a *first order* Markov Model. If the model's states represented the words of a language, sequences could be used to model utterances in that language, and transition probabilities could be computed as bigram frequencies in a text corpus. However, the lack of "contextual memory" in a first order MM makes it impossible to describe long distance correlations like subject-predicate agreement or valency. In theory, using higher order Markov Models can be used to somewhat soften this problem. In a n-th order MM, the networks history of the last n-1 states is taken into account, and transition probabilities are computed from n+1-gram frequencies (using so-called trigrams, tetragrams etc.). In practice, however, due to the exponential combinatorial growth of the number of possible n-grams, such an approach is not feasible for an MM where states are words (or rather, in this context, word forms). Even the 1 million bigrams of a 1 million word corpus have no great worth for predicting the 40.000.000.000 possible transitions for a language with 200.000 word forms.

This is why most part of speech taggers use *Hidden* Markov Models, where states stand for word classes, or morphologically subclassified word classes, like NS (noun singular) or even VBE3S (the verb "to be" in the 3.person singular), and each (PoS-) state generates words from a matrix of so-called lexical probabilities. An English article, for example, might be said to have a probability of 0.6 for being 'the', and 0.4 for being 'a'. The reason why the model is now called *hidden*, is that it is only the word-symbols that can be directly observed, whereas the underlying state-transitions remain hidden from view.

For word classes, trigram frequencies *can* be meaningfully computed from a tagged corpus of reasonable size, and the same corpus can be used to determine lexical frequencies. The trained tagger can then be used on unknown text, provided the existence of a lexicon of word forms, or at least inflexion and suffix morphemes. Interestingly, for small training corpora, the trigram-approach even performs slightly better than a variable context algorithm (Lezius et. al., 1996).

For making its decision, the HMM tagger computes the probability of a given string of words being generated by a certain sequence of word class transitions, and tries to maximise this value. The probability value (for a string w_1 w_2 w_3 ... w_n of n words) is the product of all n transition probabilities and all n lexical probabilities[89]:

for bigrams:
$p(T) * p(W|T) = p(t_1) * p(t_2|t_1) * p(t_3|t_2) * ... * p(t_n|t_{n-1}) * p(w_1|t_1) * p(w_2|t_2) * p(w_3|t_3) * ... * p(w_n|t_n)$

for trigrams:
$p(T) * p(W|T) = p(t_1) * p(t_2|t_1) * p(t_3|t_2t_1) * ... * p(t_n|t_{n-1}t_{n-2}) * p(w_1|t_1) * p(w_2|t_2) * p(w_3|t_3) * ... * p(w_n|t_n)$

[where p = probability, W = word chain, T = tag chain, w = word, t = tag]

Since $p(T|W) = p(W|T) * p(T) / p(W)$ and $p(W)$ is constant for all readings, $p(T|W)$ is maximised at the same time as $p(T) * p(W|T)$.

[89] Eeg-Olofsson (1996, IV, p.73) thinks that relative (i.e. lexical) and transitional probabilities are, in a way, complementary, with one of them being able to compensate for lack of information with regard to the other.

In a brute force approach, for an average word ambiguity of 2, two to the power of n combinations would have to be computed. Such exponential complexity growth is, of course, quite prohibitive. But since what is wanted is only the *most likely* reading (and not the probabilities of *all* possible readings), the program can be set to only use the highest probability chain encountered *so far* (i.e. from word 1 up to word i) when moving on to the next word in the string (i.e. making the transition i -> i+1). This so-called Viterbi-algorithm yields linear (and therefore manageable) complexity growth, where the number of operations is proportional to 2n for n words that are on average two-way ambiguous. Due to the problem of limited training data, zero-probability transitions have to be replaced by small default values or by lower-n-gram values (i.e. trigrams by bigrams). Other necessary ad-hoc solutions include heuristics for proper nouns and lexicon failures (e.g. the use of suffix/PoS probabilities).

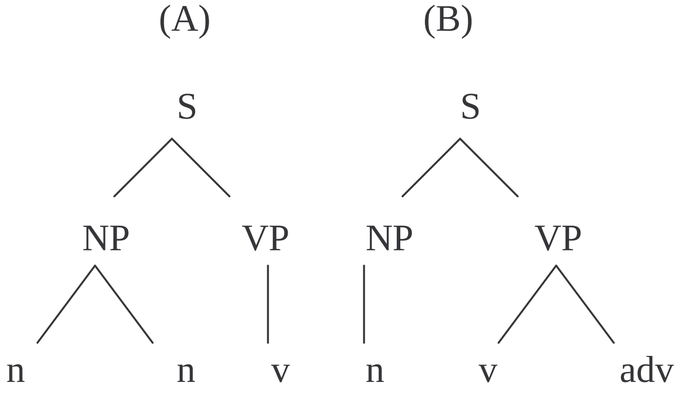
Interestingly, while the existence of PoS-lexica is a *conditio-sine-qua-non* for most languages, the lack of a tagged training corpus for the estimation of transition probabilities can be partly compensated for by estimating the parameters of the HMM by means of an iterative re-estimation process on a previously untagged corpus (called "forward-backward"-algorithm or Baum-Welch-algorithm). On the other hand, if a sizeable tagged corpus *is* available for the language concerned, even the lack of a lexicon is no real hurdle, since a lexicon file can be automatically compiled from the tagged corpus, and will have a fair coverage at least for texts from the same domain. Thus, the 1 million word Brown corpus contains some 70.000 word forms. The importance of good lexicon coverage has been tested by Eeg-Olofsson (1991, IV p.43) for a system combining lexicon entries with a heuristics based on 610 suffix strings: using a 50.000 word corpus of spoken English, the system had an error rate of 2.4% with full lexicon coverage, but 6% when using a lexicon compiled from one half of the corpus and then tested on the other. Even with a large suffix module a sizeable lexicon appears to be necessary, in order to cover those words that are *exceptions* to the suffix patterns.

The big advantage of probabilistic taggers is that they are fast, and can be trained in a short time, without the need of writing a real grammar of rules. Biasing a probabilistic tagger by adding hand written rules or exceptions, may actually have an adverse effect on its performance, since intervening on the behalf of a few irregular words, for example, would interfere with the much more important statistical modelling of the regular "majority" cases (Chanod and Tapanainen, 1994). Rumours have it that such phenomena, as well as development speed and cross-language portability of probabilistic tools, have made some commercial NLP enterprises believe that system improvement can actually be improved by firing a linguist (and hiring a mathematician instead). This view, of course, opportunistically ignores the fact that without linguists, there would be no lexica and no tagged corpora to train a probabilistic parser on in the first place.

Even trigrams, however, are far from expressing real syntactic structure, and the lexical collocation knowledge expressed in Hidden MMs is diluted considerably by the fact that it is seen through a word class filter. While the lexicalisation problem to a certain degree also haunts rule based grammars, the syntactic structure problem is "unique" to probabilistic HMM grammars and resides in the "Markov assumption" that $p(t_n|t_1 \dots t_{n-1}) = p(t_n|t_{n-1})$ (for bigrams), or $= p(t_n|t_{n-1}t_{n-2})$ (for trigrams). In generative grammar, syntactic structure is handled in an explicit way, and functions both as the traditional objective and as the main tool of disambiguation. In CG, finally, syntactic structure can be expressed, but results as a kind of by-product of sequential contextual disambiguation rules. Of course, it does matter what the objective of disambiguation is: in fact, as shown in chapter 3.7.3, two thirds of all *morphological* CG-rules make do without "global" rules, i.e. they could be expressed as statistical n-gram transitions (though even here, most rules use a larger-than-trigram window), while only 10-20% of *syntactic* CG-rules can manage without unbounded contexts.

One proposed solution to the syntax problem in probabilistic systems has been to expand context-free grammars (CFGs) into *probabilistic* context-free grammars (PCFGs), where CFG-productions are assigned conditional probabilities on the non-terminal being expanded, and the probability for a given syntactic (sub)tree can be computed as the product of the probabilities of all productions involved. The two readings of the sentence *'Einstein lectures last.'*, for instance, can be described by the following mini-PCFG, consisting of CFG-rules weighted with - arbitrary - production probabilities:

(2) *Einstein lectures last.*
1. S -> NP VP (p = 0.5)
2. VP -> v (p = 0.3)
3. VP -> v adv (p = 0.2)
4. NP -> n (p =0.4)
5. NP -> n n (p =0.1)



The complex NP reading (Einstein @>N lectures @SUBJ> last @FMV) involves productions 1, 2 and 5, yielding a complex probability of 0.5 x 0.3 x 0.1 = 0.015, while the single noun reading (Einstein @SUBJ> lectures @FMV last @<ADVL) can be generated by 1, 3 and 4, with a probability of 0.5 x 0.2 x 0.4 = 0.04, and will thus be chosen by the parser.

PCFGs address one of the most serious problems with ordinary generative grammars, that is, their tendency to produce either no parse or a parse forest of hundreds or thousands of trees without any obvious order or preference. Thus, PCFGs can, like CG, *make a choice*.

While undeniably involving more context than HMMs, probabilistic CFGs suffer from the same lexicalisation problem and to a much higher degree from scarceness of hand-tagged training material (while the higher complexity involved would demand *more* training data, there is actually *less* material available[90]). One of the core problems of PCFGs is deeply rooted in the assumption of "context-free-ness" itself: the probability of a given production is wrongly supposed to be the same *everywhere*. Still, linguistic context like the function and dependency of the non-terminal in question, will obviously have a strong influence on this probability. NPs, for instance, are more likely to be definite (i.e. expand into 'det-def N' or pronouns) in subject position than in direct object position. While function and dependency are easily available context conditions in Constraint Grammar, they would have to be expressed in a more implicit way in PCFGs. An NP's subject function, for example, might in English be expressed by stating that the NP in question is the first NP in a 'S -> NP VP' production happening to be describing the NP's mother node, and the conditional probability concerned would then read: p(NP -> det-def N | NP in S -> NP VP).

Current Constraint Grammars, on the other hand, have only crude tools at their disposal for exploiting statistical tendencies in collocational patterns, like lexically marking certain readings as <Rare>, or ordering rules in successively applied sets of less and less safe, or more and more heuristic character. Such rule hierarchies mimic, in a way, the rule probabilities of PCFGs, yet without the latter's mathematical precision.

State-of-the-art probabilistic PoS-taggers can now compete with traditional rule based systems and achieve correctness rates of 96-97%. Probabilistic taggers also provide a good base line against which to measure any other tagger: even zero-order HMM, i.e. where each word simply is assigned its post likely PoS, have a correctness rate of 91-92%, for English (Eeg-Olofsson, 1991).

Early systems computed both lexical probabilities and Markov Model PoS transition probabilities from tagged corpora, as - for English - in (Church, 1988) and in the LOB-tagging system, CLAWS (Garside, 1987), where a success rate of 96-97% is reported for a mixed tag sets of PoS, inflexion and - for a few words - base form. By using techniques like the Baum-Welch algorithm, lexica with different tag sets can be used as a starting point, with only ordinary text to train on. In (Cutting et. al., 1992), for example, 96% correctness is claimed for recovering PoS tags from the tagged Brown Corpus (Francis and Kucera, 1992), using only a lexicon and untagged training text from the same corpus. With yet another probabilistic approach, Ratnaparkhi's maximum-entropy tagger (Ratnaparkhi, 1996) claims 97% accuracy on WSJ text when trained on the Penn Treebank (Marcus et al., 1993). In (Brill, 1992) automatically learned trigram transformation rules are used in combination with a simple zero-order stochastic tagger, with error rates around 5% when using a tagged training corpus but

[90] For English, the 100.000 word syntactically annotated Suzanne corpus does provide such training data, but it must still be considered a corpus of rather modest size when compared to the market of purely PoS-tagged corpora.

no lexicon. By combining supervised and unsupervised learning, accuracies of up to 96.8% have subsequently be described (Brill, 1996). Results for languages other than English seem to confirm the 97% mark as a kind of upper ceiling for the performance of probabilistic PoS taggers. Thus, the Morphy system described in (Lezius et. al., 1996) achieved an accuracy of 95.9% for a tag set of 51 tags, using a lexicon of 21.500 words (about 100.000 word forms). Lezius cites 5 other German taggers or morphology systems with accuracy rates in the range between 92.8 - 96.7%[91] (ibd. p. 370).

For probabilistic (syntactic) parsing, performance is considerably lower, and such systems have not so far been able to replace manual annotation as a means of syntactic parsing. For standard PCFGs, which augment standard CFGs with probabilistic applicability constraints, accuracies of about 35% are supposed to be typical. Better results are achieved by conditioning production probabilities not only on the terminal in question, but also on the rule that generated it, as well as one or more subsequent words. On the short sentences of the MIT Voyager corpus, an accuracy of 87.5% is reported (Marcus, 1993). Some parsers make use of lexical information: For the SPATTER parser (Magermann, 1995) 84% accuracy is claimed for recovering labelled constituents in WSJ text. In (Collins, 1996) head-dependent relations between pairs of words are modelled in a probabilistic fashion, yielding 85% precision and recall on the same material. For longer sentences, systems do not fare as well: (Carroll and Briscoe, 1995) describes experiments with a probabilistic LR parser trained and tested on the Susanne-corpus (average sentence length: 20 tokens), which first had been relabelled with CLAWS-II tags using the Acquilex HMM-tagger (Elsworthy, 1994). Here, for bracketings matching the treebank, a recall of 73.56% and a precision of 39.82 is reported for the highest ranked 3 analyses of each sentence. 43.8% of sentences had the correct analysis ranked among the top 10. Parse fails amounted to 25.9% and time-outs to 0.2%. Nearly a third of all test sentences received more than one hundred different analyses, 5.8% were assigned more than 100.000 parses.

3.5.2 Generative Grammar: All or nothing - the competence problem

[91] For larger tag sets with hundreds of tags (presumably including inflexional information), considerably lower accuracy rates - around 80% - are cited for those members of the group of German taggers, that have this option. Of course, as Elworthy (1995) points out, what is important for performance, may not so much be the size of the tag set used, but the type of information encoded. From the point of view of disambiguation one might argue that larger tag sets leave more ambiguities to resolve, but they also provide more and better context to do so (for example, in the shape of inflexional agreement information). The relatively constant performance of different versions of CLAWS (Leech et. al., 1994), with tag set size varying by nearly a factor of three, seem to corroborate this assumption.

Generative Grammar, introduced and advocated by Noam Chomsky in the fifties[92] as Generative-Transformational Grammar, comes in many flavours. It is alive and well today in the shape of - for example - Government and Binding Theory (GB), Generalised Phrase Structure Grammar (GPSG) or Head Driven Phrase Structure Grammar (HPSG). One of the main - and most revolutionary - ideas of early Phrase Structure Grammar (PSG) was to express syntactic function as constituent structure. Thus, a subject would be implicitly defined as that noun phrase (NP) which is left after removing a sentence's other main constituent, the verb phrase (VP)[93]. A pure PSG would take word class information from a lexicon of full-forms, ignoring inflexion and semantic information. The grammar as such would then consist of rewriting rules that allow substitution of lower-level symbol sequences for higher-level symbol sequences (so-called "productions"). Symbols can be terminals (words or word classes) or non-terminals (complex units of words and/or symbols), and providing for a start symbol S (typically a sentence), we arrive at the following complete "grammar" for the PSG meta-language:

1. **T** terminal vocabulary set (e.g. words and parts of speech)
2. **N** non-terminal vocabulary set (e.g. noun phrase, verb phrase)
3. **P** set of productions a -> b (e.g. noun phrase -> determiner noun)
4. **S** start symbol, a member of N

A miniature grammar, capable of generating the sentence *'The cat eats a mouse'*, would consist of a lexicon of terminals (*'the'* det, *'a'* det, *'cat'* n, *'mouse'* n, *'eats'* v), non-terminals (NP = noun phrase, VP = verb phrase), and the following three productions:

S -> NP VP
VP -> v NP
NP -> det n

Agreement is hard to express by word class alone (plural nouns and 3.person singular verbs, for example, would have to be separate word classes), but can be incorporated in the form of Prolog style arguments, as in Definite Clause Grammar (DCG), for instance:

S -> NP(number) VP(number)
VP(number) -> V(number) NP(number2)
NP(number) -> det(number) n(number)

[92] Chomsky's *Syntactic Structures* was published in 1957.

[93] One can say, that this idea is further pursued in Categorical Grammar where all word classes and phrase classes are defined in terms of constituent categories, with only two basic categories, s (sentence) and t (referent, i.e. "noun").

Here, the variable ‘number’ can be instantiated with the values ‘singular’ or ‘plural’, and all instances of the same variable in a rule have to “match” (or to be “unified”, which is why such grammars are called unification grammars), that is, their values have to be the same in order for the production to be legitimate (note, that in the second rule, there are *two different* number-variables - for verb and direct object, respectively -, that do *not* have to match!). While pure PSG has a certain appeal for isolating languages like English - not least due to its pedagogical simplicity - unification grammars are unavoidable where generative grammar is applied to inflexional languages like French or German.

Higher level generative grammars, like HPSG, may incorporate other subcategorisation information, like valency and selection restrictions, into the lexicon, and thus build a more sophisticated rule set.

Traditionally, four levels of descriptional power are distinguished for generative grammars:

Chomsky's hierarchy of grammar classes (Chomsky, 1959)
(low number: more powerful, high number: more restricted

0 unrestricted PSG
1 context sensitive PSG
x -> y [where y has more symbols than x, e.g. A B -> C D E]
or: x A z -> x y z [other notation with "visible" context]
2 context free PSG
A -> x
3 regular PSG = finite state grammars
left linear: A -> B t, A -> t
right linear: A -> t B, A -> t
[where: T = terminal; N = non-terminal; A,B, C, D, E $\in$ N; t $\in$ T; x,y,z = sequences of T and/or N]

The computationally most interesting grammars are the least powerful, - finite state grammars, since they can be implemented as algorithmically very efficient transition networks (reminiscent of the above described Markov Models, without the transition probabilities). In such networks, the computer program starts from the start symbol and moves along possible transition paths (arcs) between non-terminal symbols. Every path is labelled with a non-terminal symbol (word or word class), and can only be taken, if the word class or word in question is encountered linearly at the next position to the right (in right linear grammars[94]). When it encounters a “dead end” (i.e. a non-terminal

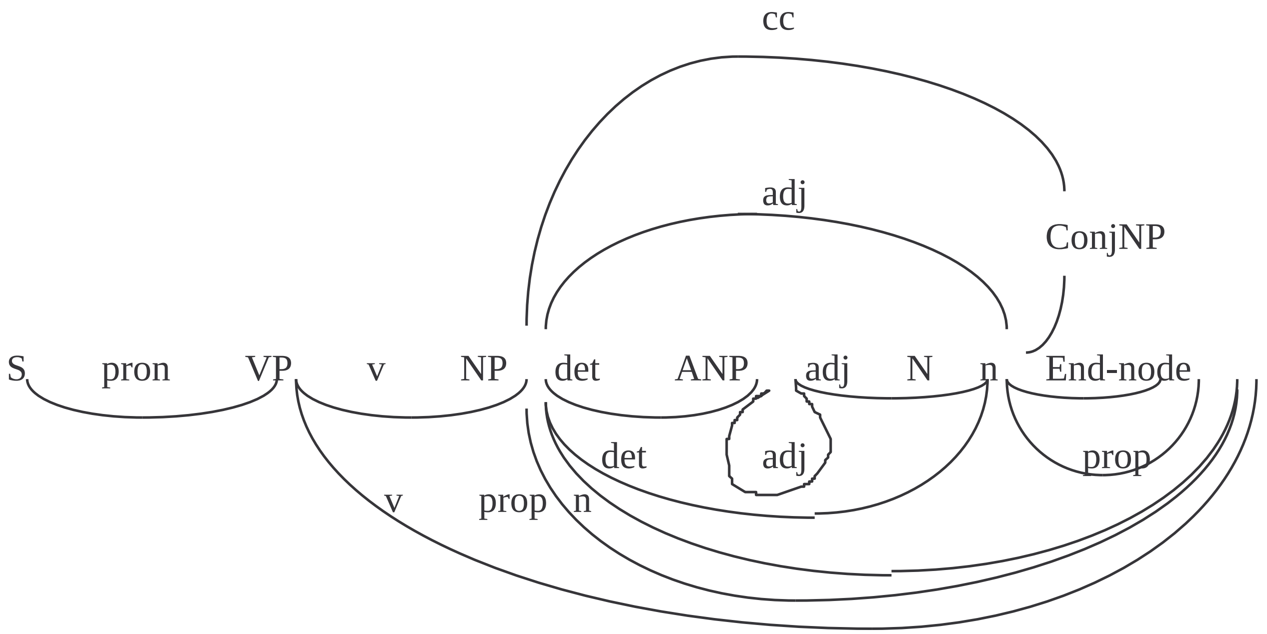
[94] In left linear grammars, the algorithm would have to work from right to left, in order to avoid infinite loops created by the possibility of reiterating non-terminal production of the type A -> A t, as in ADJP -> ADJP adj.

node without branches matching the next word or word class), the algorithm retraces its steps back to the last viable branching alternative. This way, the whole tree of possible constituent analyses is searched in a finite number of steps, and if a given sentence falls into the subsegment of a language described by the grammar in question, the algorithm will print out an analysis each time it encounters the ‘end’ symbol (i.e. takes a path matched by the last word of the sentence).

An example for a simple finite state transition network is shown below:

Example for finite state transition network:

S -> pron VP
VP -> v
VP -> v NP
NP -> n
NP -> det ANP
NP -> adj N
ANP -> adj ANP
ANP -> adj N
N -> n
N -> prop
N -> n ConjNP
ConjNP -> cc NP



E.g. *She offers green tea and red oranges and a song*.
He loves a fine story and Shakespeare.

Though Finite State Machines (FSM) are fast, finite and efficient, they have a number of serious shortcomings, due to the low power of the grammar types they represent:

* An FSM's memory is very short - once a transition is made, the network only looks at paths departing from that node, and its choice will not be conditioned by *how* the algorithm got there. The NP-section of the FSM in the above example can thus not be used for an NP-subject (by adding a direct path from S to NP, and a 'v'-path from NP to VP), because the FSM would confuse subject-NP and object-NP, trying, for example, a verb-path also after having used the NP-section for *object*. Therefore two separate NP-sections have to be incorporated into the FSM, for subject and object, linked to S and VP, respectively. For a similar reason, the co-ordinating conjunction path in the example is problematic, since it doesn't distinguish between adding am NP as co-ordinated object or as subject for a co-ordinated *sentence*. To make the distinction, different "conjunct networks" would have to be inserted into the network right after S and, and before the NP node, containing conjuncted copies of the relevant network sections. Thus, an FSM's complexity can grow enormously for long sentences with heavy subordination and co-ordination.

* Regular grammars cannot express inflexional agreement as such, - they'd have to run the whole network or large sections in many parallel versions, one for every instantiation of every category. This is why unification grammars have to be level 2 grammars (context free grammars), where no restrictions apply to the right side of a production. Number- and gender-arguments, for example, can be thought of as "affixes"[95], attached as additional affix-symbols to the "normal" symbols, both allowing for either terminal or non-terminal symbols. Number-agreement can then be added to an ordinary PSG rule by inserting an affix-variable for number in the rewriting chain of symbols:

regular grammar: S -> pron VP
context free grammar: S -> pron number VP number

Since the 'number'-variable has to be instantiated with the same value in both places, the production cannot be produced by simply working step-by-step from left to right.

By comparison, the difference between context free and context sensitive grammars is more subtle - at least when applied to natural languages. Context sensitive rules can usually be rewritten as one or more context free rules. A routinely quoted counter

[95] For Portuguese, I have worked with the AGFL formalism (Affix Grammars over a Finite Lattice), as described in (Koster, 1991).

example is a grammar of one terminal and the rule (x) -> ((x)) which produces an infinite language of “sentences” with paired brackets. One of the very few examples from the domain of natural language is Swiss German that has a construction where word order in two verbal sections of a sentence is cross-dependent. But since even such examples can be circumvented for constructions of finite depth, by writing *as* many context free rules to cover the phenomenon [for the bracketing example, (x) -> ((x)), ((x)) -> (((x))), (((x))) -> ((((x)))), ...], most generative parsers have been built around context free grammars.

In most languages, morphological structure is more linear than syntactic structure, and therefore easier to describe in an FSM framework. Thus, the TWOL-systems (Koskenniemi, 1983) used to supply analyser-input for most Constraint Grammars, describe words as linear morpheme transitions, allowing for phonetically motivated surface level changes at morpheme borders. Thus, the word ‘unrecognisable’ would be analysed as ‘un_recognis(e)_able’. Here, the FSM contains transition paths from preverbal prefix to verbal root, and from verbal root to postverbal suffix, expressed as so-called alternation of sub-lexica. A surface-level rule removes the ‘e’ of ‘recognise’ because of the clash with the ‘a’ of ‘-able’. All inflexion and most cases of derivation and compounding can be handled this way[96].

On a syntactic level, on the other hand, it is very hard to imagine a FSM capable of describing free natural language, though the technique has been explored in recent years by, for instance, Atro Voutilainen (1994:32ff).

The generative grammars of the context free type used for syntactic parsing, try to achieve several objectives at the same time: They analyse sentences by generating sentences, and they disambiguate both word class and function by assigning structure. While this does not by itself pose unresolvable technical problems, the conceptual priorities of generative grammar do seem to make it less efficient in a parsing context, i.e. for identifying “partes orationis”, or “parts-of-speech” on a morphological and functional level:

- 1. In generative grammar, there is a tradition of assigning low priority to broad lexicography, which can be explained by the fact that “toy lexica” are fine for *generating* sentences, while being unsatisfactory for research on *parsing* [free] sentences.

- 2. The constituent structure approach creates its own, theory-specific ambiguity priorities, some of which, like the scope of postnominal attachment or some cases of

[96] One might, of course, argue, that ‘un-’ as a preverbal prefix is limited to transitivized denominal verbs, usually ending in ‘-ize’/’-ise’ or ‘-ate’. For productive word composition one would then need a higher level (context free) rule to describe the interdependence of the causative suffix and the antonymous prefix.

co-ordination, are syntactically irresolvable. Such “surplus” ambiguity compromises notational clarity and creates huge “parse forests”.

- 3. Since structure is found by recursive generation of syntactic trees, a lot of “dead end” partial constructions are computed, rendering the technique very time consuming (to the point of “time out” for very long sentences).

Furthermore, Generative Grammar assumes a stable language system with clear-cut borders for what is correct. The objective is to generate “all and only” the sentences of a given language that are correct. The Chomskyan point of departure was an innate and trained “language faculty” rooted in the human brain and capable of making the distinction by means of “competence”. This approach contains the risk of fostering a “black-and-white”-attidude to language analysis, visible for instance when a generative grammar’s rule set is seen as *prescriptive* in nature rather than *descriptive* (since it rules out as “not part of the language system” or as “performance errors” what it cannot describe). In general, the generative approach also entails that the notion of “parsing failure” is acceptable[97], whereas probabilistic and CG-based systems assume that “the corpus is always right”, and can run on large chunks of running text without ever giving up on a sentence.

When comparing Constraint Grammar to Generative Grammar, one has to distinguish between conceptual differences and implementational differences. Conceptually, CG is - unlike PSG - parsing-oriented and next to useless for generating sentences. In CG, ambiguity is defined independently from structure, and ambiguity resolution is consequently more flexible. CG is reductionist rather than generativist, which makes it more tolerant (or robust) with regard to what Chomskyan grammar would call performance failures, incomplete utterances, dialectal variation and the like. In its objective, CG is descriptive rather than prescriptive, but technically, it follows a third road, which - in analogy with the other two - might be termed “prohibitive”.

Implementationally, a key difference is that, in Constraint Grammar, ambiguity can be reduced *gradually*, without retracing, and that rules tend to add or remove form and function labels for individual words, defining (in a reductionist way) what is *not* contextually feasible rather than expressing syntactic patterns (in a generative, productive way) for multi-word units.

The performance of most grammar based systems is difficult to compare to that of probabilistic or Constraint Grammar based parsers, since the theoretical potential is usually valued higher than practical applicability to unrestricted text, for example by trading lexical coverage for descriptive power. Still, such systems have been applied to

[97] Though the programming formalism as such would allow compromise solutions like partial parses or automatic ad hoc rule amendments.

wide coverage tagging and parsing tasks, as in the case of the GPSG based Alvey Natural Language Tools - ANLT - (Phillips & Thompson, 1987) or the ongoing TOSCA project (Oostdijk, 1991) using extended affix grammar. Since an existing CFG can be enhanced by probabilistic indexing of its production rules (cp. 3.5.2), hybrid systems may be one way to solve the recalcitrant problem of huge parse forests for long sentences, conceptually inherent to the constituent analysis approach. In (Wauschkuhn, 1996) a chart parser is used to implement 615 PSG rules for German, where every rule is assigned a "safety factor" measuring "usage plausability". The default for terminal productions is 1. The safety factor, though seemingly assigned by hand, works much the same way as rule probabilities in PCFGs, allowing to compute a ranking for every tree in the parse forest: here, the safety factor of the left side (non-terminal) of a production is the product of the safety factors of all right hand side symbols. Wauschkuhn's parsing system assigns complete analyses to 56.5% of the sentences in a 1.6 million word news text corpus, and partial analyses to 85.7%. Due to the lack of a benchmark corpus, no correctness rate is given[98]. In contrast to many other systems, a sentence is analysed in two steps: more than half the rules treat macrostructure (clause-trees), and the rest then parses each subclause's microstructure individually. Thus, even partial analyses still construct clause-trees, with less than a fifth of partial analyses exhibiting microstructure failures in more than one subclause. This additional robustness is reminiscent of Constraint Grammar, where *all* rules in principle are perceived as independent of each other, and most of the structure of a sentence will survive a locally wrong function tag or a wrong dependency marking.

3.5.3 Constraint Grammar: the holographic picture (addressing ambiguity directly)

Most words in natural language texts are - seen in isolation - ambiguous with regard to word class, inflexion, syntactic function, semantic content etc. It is, above all, sentence context (besides content coherence and the reader's "knowledge about the world") that determines how a word is to be understood. *Constraint Grammar* (CG), introduced by Fred Karlsson (1990) and shaped by the Helsinki School (cp. Karlsson et.al., 1995), is a grammatical approach that aims at performing such disambiguation by establishing rules for which of a word form's possible readings is to be chosen, and which readings are to be discarded in a given sentence context. In the parser itself these rules are

[98] Wauschkuhn did experiment with ambiguity (ibd., p. 366), reducing parse forest size by running input text through a PoS tagger first, but blames the available taggers' high error rate (3-5%) for a corresponding drop in parse quality. The interesting question is how the experiment would have worked with input from a Constraint Grammar tagger, since such taggers usually claim much lower error rates than probabilistic systems, cp. (Karlsson et. al., 1995) and (Bick, 1996).

compiled into a computer program that takes as input morphologically processed, but still fully ambiguous text, as provided by lexicon and inflexion rule based analysers like the one used in my own system, or the TWOL analysers (in the Helsinki systems, cp. Koskenniemi, 1983). The multiple ambiguity represented by alternative tag lines, will optimally be reduced to only one line[99] (the correct reading) by the CG-rule system.

(1) Constraint grammar input (morphological analyser output)

```
"<nunca>"
        "nunca" ADV
"<como>"
        "como" <rel> ADV
        "como" <interr> ADV
        "como" KS
        "como" <vt> V PR 1S VFIN
"<peixe>"
        "peixe" N M S
"<$.>"
```

[ADV=adverb, KS=subordinating conjunction, V=verb, N=noun, PR=present tense, S=singular, M=maskuline, 1=1.person, VFIN=finite verb, <rel>=relative, <interr>=interrogative, <vt>=monotransitive]

The four readings[100] of the word form *'como'* are - in CG terminology - called a *cohort*. A typical CG-rule[101] for disambiguating this ambiguity might be the following:

(2) SELECT (VFIN) IF (NOT *-1 VFIN) (NOT *1 VFIN)

[select for a given word form the VFIN reading (finite verb) if there is no (NOT) - neither to the left (*-1) nor the right (*1) - other word that can be VFIN.][102]

By first adding ("mapping") all[103] possible syntactic functions onto a word form, conditioned by its word class, inflexion etc., and then disambiguating this syntactic

[99] Of course, in the case of true ambiguity (which is surprisingly rare in the world of corpus linguistics), two (or more) correct tag lines are possible and should then be preserved.

[100] The difference <rel> ADV and <interr> ADV is not really motivated by morphological word class, but expresses a semantic-functional distinction (the English translation is 'like' in the first case, and 'as' in the second). It is of great importance to polysemy resolution to determine which of a word's potential valency patterns has been instantiated in a given clause context, and which semantic class fills a given valency slot. Here valency tags (and selection restrictions) gain importance not only as *secondary* tags (that exclusively are used for the disambiguation of morphological/syntactic tags), but also as *primary* tags in their own right, which can and must be ambiguated, like for the word form *'revista'* , where simple word class ambiguity (V-N) is turned into fourfold lexeme ambiguity:

rever <vt> V 'see again'	instantiated valency: transitive <vt>
rever <vi> V 'leak through'	instantiated valency: intransitive <vi>
revista <+n><rr> N 'news magazine'	instantiated valency: title <+n>, semantic class: reading matter <rr>
revista <CP> N 'inspection'	instantiated semantic class: +CONTROL, +PERFECTIVE

[101] The notation convention used here is the one used by Pasi Tapanainen's cg2-compiler, which among other things replaces the older operators '@w=0' and '@w=!' with the ordinary English words 'REMOVE' med ' SELECT'.

[102] The rule has been simplified, presuming that every sentence contains at least one finite verb, which isn't always the case, in head lines, exclamations etc. The rule can be made safer by conditioning it on the existence of a full stop (*1 PUNKTUM) or by exploiting the possible valency relation between the transitive verb *comer* and the 'safe' *peixe* (0 <vt>) (1C NP).

ambiguity, Constraint Grammar can also be used for syntactic parsing, as efficiently shown, for instance, in the Bank-of-English-project (200 million words, Järvinen, 1994).

(3) Input to the syntactic CG-rules (after mapping)

```
"<nunca>"
        "nunca" ADV @ADVL
"<como>"
        "como" <vt> V PR 1S VFIN @FMV
"<peixe>"
        "peixe" N M S @SUBJ @ACC @SC @OC
```

[@ADVL=adverbial, @FMV=finite main verb, @SUBJ=subject, @ACC=direct object, @SC=subject complement, @OC=object complement]

In (3), adding all possible syntactic tags (@) has resulted in fourfold syntactic ambiguity for *peixe*. The direct object reading (@ACC) can be selected in a positive way by means of a 'SELECT'- rule exploiting the transitivity of the verb, but it could just as well be identified indirectly, - by being the only surviving reading, after CG-rules have discarded all others:

(4) REMOVE (@SUBJ) IF (0 N) (NOT *-1 V3) (NOT *1 V3)
[discard the subject reading, if the target is a noun (N) and there is no verb in the 3.person]

REMOVE (@SC) IF (NOT *-1 <vK>) (NOT *1 <vK>)
[discard the subject complement reading (@SC) if there is no copula verb (<vK>) in the sentence]

REMOVE (@OC) IF (NOT *-1 @ACC) (NOT *1 @ACC)
[discard the object complement reading (@OC) if there is no direct object reading (@ACC) in the sentence][104]

It is this indirect disambiguation, that is most characteristic of Constraint Grammar, and it is the prime reason for the robustness of this method: even rare or incomplete constructions will receive at least *one* reading - the one that survives the most constraints. The incremental use of the rules, with safe contexts and safe rules before ambiguous contexts and heuristic rules, furthermore ensures that the parser will prefer a reading that is "almost correct" to one that is "quite wrong".

CG-grammars have first of all been described for English (e.g. Karlsson et.al., 1991), but there are - on the morphological level, at least - projects involving several

[103] In the mapping modul, constraint grammar rules are used, too, and the list of possible syntactic functions for a given word form can thus be made context dependent (and, of course, shorter).

[104] Note that all 3 rules make use of "unbound" contexts conditions:
*-1 = the context condition is to be true *anywhere* to the left (1 or more positions to the left)
*1 = the context condition is to be true *anywhere* to the right (1 or more positions to the right)
Of course,"bound" context conditions can be used, e.g. -2 = second word to the left, 3 = third word to the right. Bound context conditions can in principle be translated into n-gram rules (as used in probabilistic HMM parsers), while "unbound" (*-context) conditions are characteristic of Constraint Grammar and not easily translatable into probabilistic systems (cp. also chapter 3.7.3).

other languages from both the Germanic, Romance and Finno-Ugric language families (Swedish, German, French, Finnish etc.)[105]. A mature CG-grammar for the morphological level (word class or PoS disambiguation), typically consists of at least 1.000-2.000 rules. For the English ENGCG system, word class error rates of under 0.3% have been reported at a disambiguation level of 94-97% (Voutilainen, 1992).

In a recent direct comparison[106] between an updated ENGCG and a statistical tagger trained on a 357.000[107] word section of the Brown corpus, Samuelsson & Voutilainen (1999) found that error rates for the Constraint Grammar system were at least an order of magnitude lower than those of the probabilistic system at comparable disambiguation levels. Thus, ENGCG error rates were 0.1% with a 1.07 tag/word ratio and 0.43% with a 1.026 tag/word ration, while the statistical system achieved error rates of 2.8% and 3.72%, respectively.

Constraint Grammar type rules have also been used in hybrid systems, for instance where an automated learning algorithm is trained on a morphologically tagged corpus with the objective of constructing or selecting local context discard rules. Thus, Lindberg (1998), using Progol inductive logic programming[108] and a ±2 word context window, reports 98% recall in Swedish test texts, with a residual ambiguity of 1.13 readings pr. word, and a rule body of 7000 rules. Another hybrid system is decribed in Padró i Cirera (1997) , where a relaxation labelling tagger is applied to English and Spanish. In this system, CG style rules for POS-tagging were integrated with HMM tagging, creating a statistical model for for the distribution of tag targets and context conditions. Constraint rules were partly learned from a training corpus using statistical decision trees, and partly hand-written on the basis of output errors in probabilistic HMM taggers[109]. In comparison with HMM and relaxation labelling base line taggers, both types of constraint rules improved tagger performance individually, and resulted - when combined - in an overall precision rate of 97.35% for fully disambiguated Wall Street Journal text.

While hybrid systems thus seem to offer some advances in comparison with ordinary HMM modelling and related techniques, they are still far from achieving ENGCG level results, one likely explanation residing in the fact that the automatically learned rules of such systems (so far) lack the global scope (i.e. sentence scope) and

[105] For a short comparison of CG systems, cp. chapter 8.1.

[106] Both systems used the same tag set: CG-tags were filtered into the kind of fused single tags typical of statistical taggers. Both systems were tested on the same 50.000 word benchmark text, consisting of both journalistic, scientific and manual excerpts.

[107] At this training corpus size, the learning curve of the statistical tagger flattened out, suggesting that larger training corpora would not lead to any significant improvement in tagging performance.

[108] In addition, Lindberg used so-called “lexical” rules (not to be induced), removing rare readings of frequent word forms, much like the heuristic <Rare> rules in a regular CG - but with the important difference, that the CG <Rare> rules would be used *after* at least one round of regular disambiguation, whereas Lindberg’s lexical rules came into play *before* ordinary (induced) rules.

[109] With only 20 linguist-written rules, the balance was heavily in favour of the automatically generated constraints (8473).

syntactic reach of ordinary hand-crafted CG rules, and that linguist written rules have not (yet) been extensively employed.

3.6 The rule formalism

In principle, the paradigm of Constraint Grammar is independent not only of the particular notational conventions commonly associated with it (such as flat dependency syntax), but also of the rule formalism used to implement and compile Constraint Grammar rules. Up to now, however, only very few CG-compilers have been written, and the conventions established by Fred Karlsson's original LISP-implementation have largely been maintained in later implementations. Today, to my knowledge, only Pasi Tapanainen's two rule compilers, cg1 and cg2, are available to the research community, one licensed by Lingsoft (www.lingsoft.fi), the other by Connexor (www.conexor.fi).

For testing purposes I programmed (in 1996) a C-version of a cg1-compatible compiler myself, which handled the morphological disambiguation module in my parser, but only at about 50% the speed achieved by Tapanainen's cg1. Still, I gained valuable insight into the way CG-rules work and interact on a technical level. Thus, I was able to measure "reiteracy" on individual rule set levels: Though - in theory - rules are supposed to come into play gradually as their contexts grow safer by the work of other rules, in practice almost all test runs "dried up" already after 2 rounds (on the same heuristics level). In the face of 18% four-fold-or-higher morphological ambiguity (ch. 3.2.1), this may mean that CG-rules help each other somewhat more by focusing on different tags and contexts than by disambiguating each other's context. In other words, CG-rules can be thought to be complementary to a higher degree than they are interdependent.

This chapter is meant as a short but comprehensive introduction to Pasi Tapanainen's cg2 rule-compiler (Tapanainen, 1996), which is the one PALAVRAS is currently using (1999).

The cg2-compiler runs under UNIX, with the following command line:

dis —grammar rule-file < text.tagged > text.dis

(which reads a rule file into the compiler, and applies it to a tagged text, a disambiguated version of which is then written to an output file.)

Or (if mapping rules are included, typically at the syntactic level):

mdis—grammar rule-file < text.tagged > text.map&dis

Input from the morphological analyser must be verticalised text, i.e. one word form per line, followed by all possible readings for this word form, with one reading pr. line, typically arranged as a so-called *cohort* in the following way, conventionally with base forms in quotes, secondary tags in <>, and morphological tags (the ones destined for disambiguation) in capital letters.

```
word form
      "base form-1" <valency> .. <semantics> .. WORD CLASS-1 INFLEXION
```

“base form-1” <valency> .. <semantics> .. WORD CLASS-2 INFLEXION
“base form-2” <valency> .. <semantics> .. WORD CLASS-3 INFLEXION
“base form-2” <valency> .. <semantics> .. WORD CLASS-4 INFLEXION

A rules file ordinarily consists of the following sections:

DELIMITERS (1 section, defines sentence boundaries)
SETS (1 or more lists of set-definitions, compiled as one)
MAPPINGS (1 list of mapping rules for adding context dependent tags)
CONSTRAINTS (1 or more lists of CG-rules, compiled one section at a time)
END

In case there are several *constraints* sections with *constraint grammar* rules, these will be applied to the input text in the same order sections have in the file. This way, it is possible to distinguish, for instance, between morphological disambiguation, to be done before, and syntactic disambiguation, to be done after the mapping of syntactic tags.

Comments can be added anywhere in the rules file after a #-sign.

DELIMITERS

The compiler is told which text window the rules are to be applied to. In the case of PALAVRAS the following punctuation delimiters are included:

<$.> <$!> <$?> <$;> <$:> <$--> <$(> <${> <$}> ;

Note that quotes and single hyphens are not included. This may result in complex sentences with parenthetical clauses causing trouble for rules based on, e.g., the uniqueness principle. On the other hand, it is easier to satisfy, for instance, verbal valency in a larger window.

A few special non-punctuation delimiters are used: <$START> which is automatically added to mark the left hand border of the *first* sentence in a text, and <$¶> which is used for graphical line breaks in news paper corpora, in connection with otherwise undelimited headlines or pictures.

SETS

In the cg2 compiler, rules can not only apply to word forms or their tags, but also to *sets* of words or tags or combinations of these. A set definition is introduced by:

(a) ***LIST*** set-name =

followed by a list of set elements (tags or tag combinations), separated by blanks, or

(b) ***SET*** set-name =

followed by a list of pre-defined sets (or tags in parentheses), linked by set operators.

Elements in (a) can be:

(1) a tag, word form or base form, e.g. N [for noun], “<palavras>”, “ir”

(2) any combination of (1) appearing in the same reading in this order, flanked by parentheses, e.g. (N M P) [for *noun masculine plural*], (“ser” V).

Set-elements from (b) can be linked by the following operators:

union*: **OR** or* **|** *, e.g. set1 OR set2 OR (tag3) OR (N F S)*

concatenation: **+** , e.g.. set1 + set2, yields all possible combinations of the 2 sets' elements. SET set1 = (V) and SET set2 = (INF) (GER) (PCP) , for instance, yield, when concatenated, all non-finite verb forms: (V INF) (V GER) (V PCP).

difference: **-** , e.g. set1 - set2, meaning set 1 *without* those elements comprising set2. SET @ARG-NON-SUBJ = @ARG - (@SUBJ), in connection with a previously defined SET @ARG = (@SUBJ) (@ACC) (@DAT) (@PIV), for instance, yields all clause level arguments with the exception of the subject.

Operators + and - are handled first, before OR. The same operators may also be used outside the definition section, in the rules, in order to link sets or tags (which, in this case, must first be turned into "sets" by a pair of parentheses).

CONSTRAINTS

A CG-rule has the following general form:

WORD FORM *OPERATION* ***TARGET*** *IF* *(CONTEXT1)* *(CONTEXT2)* *...;*

OPERATION:

(a) REMOVE

Removes, if the context condition is true, the line containing the TARGET tag, - unless this reading is the *last* surviving tag. For @-targets - conventionally syntactic function tags - the TARGET tag is removed from its line, unless it is the last surviving @-tag.

(b) SELECT

In principle, the opposite of REMOVE, - it removes all *other* reading line but the one (or those) containing the TARGET tag. For @-tags, all others are removed *from this line*.

WORD FORM:

Optional part of a rule, limiting the rule for use with this word form only. Cannot be combined with other tags into a complex tag, but is otherwise like a context condition for position 0 (the word form itself).

TARGET:

Obligatory part of a rule, contains (in parentheses) that tag (e.g. N) or tag sequence (e.g. N F P) or (without parentheses) that set which the rule is designed to select or remove.

Base forms ("...") are tags like all others, only word forms may not be used (they have their own place, in the beginning of the rule, cp. above). Instead of rewriting the same rule for several targets, these can be combined by using the set convention[110]:

SELECT NOMINAL IF (-1C DET) ;

where NOMINAL has been defined as N, A, PCP, and the context condition demands an unambiguous ('C' for 'careful') determiner at the neighbouring position to the left (-1).

CONTEXT:

Contexts are delimited by parentheses, and by default AND-linked, that is, they must all apply at the same time, if the rule is to be used (true). A complete context consists of the following:

1. A **position** information, consisting of a number denoting the relative position to the **left (-)** or **right (+)**, where (or from where) the context condition is to be checked. NOT can be added in front, and will negate the context condition. An **asterisk (*)** before the position number means "unbounded context", i.e. the condition applies all the way left (-) or right (+) of the position given (absolute or LINKed), - even if the search for a fitting context should cross the TARGET position (position 0)[111]. For non-negated (positive) contexts, only the first instance of the context condition will be instantiated (used for matching the rest of the rule), unless one uses the **double asterisk (**)**, which makes the rule checker search all the way to a DELIMITER, even in non-negated contexts. An at-sign (@) before the position number means an absolute context, @1, for instance, refers to the first cohort, @-2 to the last but one cohort in the sentence.

2. A **context condition**, consisting of a set, a tag or a tag sequence (the last two in parentheses), which again can be linked by the operators **OR** (union), **+** (concatenation within the same reading) or **AND** (intersection of two tags from the cohort). A **C (careful)** directly after the position number means that the context condition must be the cohort's only tag. (-1C N), for instance, means a safe (= fully diambiguated) noun reading one position to the left. If the word to the left has a, say, (V)-reading at the same time, the context can not be instantiated (is not true).

3. A **linked (complex) context**, where the word **LINK** "hooks up" 2 contexts (within the same context parenthesis). The second context's relative position is calculated from the first context's instantiated position, which becomes the new "0-position". This way one can build long context chains (where all the LINKed contexts are oriented towards the same side, either right (+) or left (-). Also zero-links (adding more conditions to an instantiated context) are allowed.

[110] In terma of rule writing efficiency, not allowing for sets in targets is one of the main disadvantages of the older cg1.

[111] In the cg1 compiler, an unbounded search would *not* pass the target (0) position, accounting for one of the more substantial incompatibilities between cg1 and cg1.

4. A **blocking context**, where the word **BARRIER**[112], right after a context with a *-position (an unbounded context), supplies a context condition (tag, tag sequence or set), that must not appear *before* the context in question, as calculated from the (absolute, relative or linked) position that defines the starting point for the search. (*1 VFIN BARRIER CLB), for instance, looks for a finite verb (VFIN) anywhere to the right - but this context condition only counts as true, if there is no interfering clause boundary (CLB) between position 0 and the finite verb.

MAPPINGS

A MAPPING-rule has the following general form

OPERATION *(MAPTAG1 MAPTAG2 ...) (TARGET)* ***IF*** *(CONTEXT 1) ...*

A mapping rule adds mapping tags, usually syntactic tags marked by the mapping-marker @, to those readings (= cohort lines) that contain the target-tag, - provided that all context conditions apply. This part of a mapping rule (the context test) works exactly as for the *constraint rules.*

OPERATION can be:

- **MAP**: first-time mapping, for those cohort lines, that do not yet contain a tag with the mapping marker (@). This is the normal way to map syntactic function.
- **ADD**: mapping is performed regardless of any earlier @-tags on the readings line, in particular, it will also be applied to words featuring lexical mappings from the lexicon.
- **REPLACE**: all tags but the first (usually the base form) are deleted, and replaced by the mapping tags. REPLACE rules could, to a certain degree, compensate for mistakes preceding parser modules have introduced on the tag line, but are not supported in the current CG-2 compiler.

Mapping rules are applied in exactly the order they are listed in, - in contrast to constraint rules, which are best thought of as taking effect "simultaneously"[113] and can even be tried several times, until no further disambiguation is possible.

In the case of multi-element tag strings, individual tags or tag combinations trigger appropriate mapping rules in left-to-right tag order. For example, in the tag string “ser” V PR 3S IND VFIN, mapping rules targeting the word class V (verb) will come into

[112] In cg1, a barrier context would have to be expressed by a “backwards looking” **LINK NOT *(-)1** context, making continued “forward” linking difficult.

[113] If one wants to control the order in which constraint rules are applied, this can be achieved by grouping them into several CONSTRAINTS sections, for example separating safe rules from one or more heuristic levels. In my system, six such levels are used for morphology, and four for syntax.

play after base form rules targeting “ser”, and before mappings targeting PR 3S (present tense 3.person singular).

CG2 EXAMPLE FILE:

DELIMITERS = "<.> "<!>" "<?>" ; **# sentence window**

SETS # definitions

LIST NOMINAL = N PROP ADJ PCP ; # nominals, i.e. potential nominal heads

LIST PRE-N = DET ADJ PCP ; # prenominals

LIST P = P S/P ; # plural

LIST PRE-N-P = (DET P) (DET S/P) (ADJ P) (ADJ S/P) (PCP P) (PCP S/P) ; # plural prenominals

(also: SET PRE-N-P = PRE-N + P ;) # the same via set operation

LIST CLB = "<,>" KS (ADV <rel>) (ADV <interr>) ; # clause boundaries

LIST ALL = N PROP ADJ DET PERS SPEC ADV V PRP KS KC IN ; # all word classes

LIST V-SPEAK = ("say" V) ("talk" V) "suggest" ; # speech verbs

LIST @MV = @FMV @IMV ; # main verbs

CONSTRAINTS # morphological level disambiguation

REMOVE (N S) IF (-1C PRE-N-P) ; # removes a singular noun reading if there is a safe plural prenominal directly to the left.

REMOVE NOMINAL IF (NOT 0 P) (-1C (DET) + P) ; # removes a nominal if it isn't plural but preceded by a safe plural determiner.

REMOVE (VFIN) IF (*1 VFIN BARRIER CLB OR (KC) LINK *1 VFIN BARRIER CLB OR (KC)) ; # removes a finite verb reading if there are to more finite verbs to the right none of them barred by a clause boundary (CLB) and co-ordinating conjunction (KC).

"<que>" SELECT (KS) (*-1 V-SPEAK BARRIER ALL - (ADV)) ; # selects the subordinating conjunction reading for the word form 'que', if there is a speech-verb to the left with nothing but adverbs in between.

MAPPINGS # syntactic possibilities

MAP (@SUBJ> @ACC>) TARGET (PROP) IF (*1C VFIN BARRIER ALL - (ADV)) (NOT -1 PROP OR PRP) (NOT *-1 VFIN) ; # a proper noun can be either forward subject or forward direct object, if there follows a finite verb to the right with nothing but adverbs in between, provided there is no proper noun or preposition directly to the left, and a finite verb anywhere to the left.

CONSTRAINTS # syntactic level disambiguation

REMOVE (@SUBJ>) IF (*1 @MV BARRIER CLB LINK *1C @<SUBJ BARRIER @MV) ; # removes a forward subject (SV case) if there is a safe backward subject (VS case) to the right, with only one main verb in between

3.7 Contextual information in constraint building

3.7.1 Implicit syntax: Exploiting linear structure

A Constraint Grammar has at its disposal three types of information, of which the morphological level usually[114] only exploits two (a/b):

(a) lexical information, part disambiguated or being disambiguated (base form, word class and inflexion tags), part not (secondary valency and semantic tags)

(b) the linear order of words and non-words (punctuation symbols, numbers) in a sentence.

(c) At the syntactic level, in addition, non-lexical information (syntactic function and dependency tags) is made "lexical" (mapped onto word forms) and disambiguated creating a third type of information to be used by the CG rules.

What a CG rule does, is - in principal - stating whether a certain sequence of word based tags is grammatical or not. The actual compiled grammar is handed (partially ambiguous) information of type (a) and (b) from the morphological analyser (or its own mapping module), and then extracts information of type (b) from a given sentence, trying to instantiate one or more matching tag sequences from the rule body.

Since they basically express word/tag sequences, all CG rules could be called syntagmatic, - even the morphological ones. For example, a CG grammar does not state agreement rules per se, and does not operate with the concept "noun group" (np) as such. Still, both syntactic concepts are implicitely employed even on the morphological level. Consider the following tag sequences (DET = determiner, N = noun, A = adjective, V = verb, M = masculine, F = feminine, S = singular, P = plural, 3 = third person, *agrammatical):

...	DET-MS	NMS	AMS	V3S	...
...	DET-FS	NFS	AFS	V3S	...
...	DET-MP	NMP	AMP	V3P	...
...	DET-FP	NFP	AFP	V3P	...
...	DET-M	*NF	AM	V3	...
...	DET-F	*NM	AF	V3	...
...	DET-S	*NP	AS	V3S	...
...	DET-P	*NS	AP	V3P	...

[114] That is, if the morphological and syntactic levels are kept apart in a strict way, not least for linguistic reasons. Technically, a CG-grammarian can choose - rather than apply too heuristic rules at the morphological level proper - to run an additional round of morphological rules *after* the syntactic mapping and disambiguation phases, in order to address the remaining "hard" morphological ambiguity with more (i.e. syntactic) context information.

The above sequences are examples of grammatical 3-part-np's (DET-N-ADJ) with number and gender agreement, followed by a finite verb in agreement with the noun group. The sequences can be sanctioned as grammatical by CG rules like the following:

```
        SELECT NMS IF (-1C DET-MS) (1C AMS) (2 V3S)
        SELECT DET-FS IF (1C NFS) (2C AFS) (2 V3S)
        REMOVE NS IF (-1C DET-P) (1C AP) (2 V3P)
```

On the syntactic level, linear structure is exploited more directly. Not least, adjacency of syntactically "friendly" word classes is used to establish dependency relations. In the above example, the np will be implicitely delineated by flat dependency links (cp. chapter 4.1 and 4.6), with mapping or selecting rules expressing the grammaticality of the following sequence (@>N = prenominal modifier, @N< = postnominal modifier):

```
...     DET_@>N         N       ADJ_@N<     V3      ...
```

```
MAP (@>N) TARGET (DET) IF (1C N) (2C ADJ) (3C V3)
MAP (@N<) TARGET (ADJ) IF (-2C DET) (-1C N) (1C V3)
```

In a language like Portuguese, without case marking for nouns, the implicit syntax of linear structure is also very important for the assignment of subject and object categories. Relying on lexical information about word class and valency potential, rules can be coined about the probability of sequences like SVO, VSO, SV or VS. De Oliveira (1989), for instances, cites the following frequencies for valency dependent constituent order (for utterances without zero constituents, and without a relative pronoun as subject or object, in a spoken language corpus):

SVO for "direct transitives" (<vt>):	96%
SVO for "indirect transitives" (<vp>):	97%
SVOO for "bitransitives" (<vtp>):	89%
SV for intransitives (<vi> and <ve>):	44%
VS for intransitives:	56%

The percentages for transitives verbs are high enough to justify direct "translation" into CG rules at a heuristic level, and with additional context conditions, at the non-heuristic level:

```
REMOVE (@<SUBJ) IF (*-1 @MV BARRIER CLB LINK 0 <vt> LINK NOT 0 <vi> OR <ve>)
REMOVE (@ACC>) IF (*1 @MV BARRIER CLB) (NOT 0 <rel> OR ACC)
```
[115]

For a list of the parser’s valency tags, and their statistical prominence in the CG rule set, cp. chapter 3.7.2.1.

[115] The percentages given by de Oliveira do not seem to include clitic objects in OV constructions, and in any case, excepting pronouns morphologicaly marked as ACC from a heuristic @ACC-remove rule, is more than sensible.

3.7.2 Making the most of the lexicon

3.7.2.1 Level interaction: The secondary tags of valency

In my system, valency proper is defined as the power of a dependency head to govern optional or obligatory dependents in a functional way. For this kind of valency Portuguese obeys the linear precedence principle (i.e. heads precede dependents), though it is obligatory only in the case of nominal valency (the valency of nouns and adjectives) and adverbial valency. For verbs, there are numerous exceptions with fronting of valency bound material, like subjects of non-ergative verbs, relative pronouns and focusing. The most important valency classes are listed below, with a list of examples, and of the constituents involved[116]:

for verbs:

<vt>	monotransitive	SUBJ V ACC	*comer ac., amar alg.*
<vd>	monotransitive	SUBJ V DAT	*obedecer, agradar, convir*
<vp>	monotransitive	SUBJ V PIV	*contar com, gostar de*
<va>	monotransitive	SUBJ V ADV	*durar TEMP, custar QUANT, morar LOC, ir DIR*
<vK>	copula	SUBJ V SC	*estar, ser, parecer, chamar-se*
<vi>	intransitive inergative	SUBJ V	*trabalhar, nadar, dançar, correr*
<ve>	intransitive ergative (= inaccusative)	V SUBJ	*desaparecer, chegar, desmaiar, cair, crescer, desmaiar, nascer*
<vdt>	ditransitive	SUBJ V ACC DAT	*dar ac. a alg., mostrar, vender*
<vtp>	ditransitive	SUBJ V ACC PIV	*confundir ac. com, trocar por, transformar em, afastar de*
<vta>	ditransitive	SUBJ V ACC ADV	*pôr ac. LOC, collocar ac. LOC, mandar alg./ac. DIR*
<vtK>	transitive prædicative	SUBJ V ACC OC	*achar alg./ac. ac., considerar*
<vU>	impersonal	V	*chover*

[*abbreviations used for verbal valency:* SUBJ = subject, V = verbal constituent, ACC = direct (accusative) object, DAT = indirect (dative) object, PIV = prepositional object, SC = subject predicative complement, OC = object predicative complement, ADV = adverbial object, TEMP = time quantity adverbial, QUANT = quantity adverbial, LOC = place adverbial, DIR = direction adverbial]

for nouns:

[116] Verb-dependent valency bound constituents, i.e. clause level arguments, need not necessarily come in the order given in the third column of the table. Portuguese allows (almost) free positioning of subjects, predicative complements and objects (with the exception of clitic object pronouns that always come in DAT ACC order).

<+a>, <+com>, <+de> ...	N PP	*contraste com, respeito para*
<+de+INF>, <+para+INF>	N PRP INF	*capacidade de, licença para*
<+que>	N FS-que	*convicção que, esperança que*
<+num>	N NUM	*século, capítulo*

[*used for non-verbal valency:* N = noun, ADJ = adjective, ADV = adverb, PRP = preposition, PP = prepositional phrase, NP = noun phrase, INF = infinitive, FS = finite subclause, NUM = numeral]

for adjectives:

<+a>, <+com>, <+de> ...	ADJ PP	*cônscio de, rico em*
<+de+INF>, <+para+INF>	ADJ PRP INF	*capaz de, hábil para*
<+que>	ADJ FS-que	*atento que, esperança que*

for adverbs:

<+de>	ADV PRP	*antes de, depois de*
<+de+INF>	ADV PRP INF	*antes de, depois de*
<+NP>	ADV NP	*inclusive*

for prepositions:

<+que>	PRP FS-que	*até que*

In a broader way, valency is understood as lexical co-occurence rules, so valency-like tags are used to inform, for instance, that measuring nouns like 'segundo' (second) or 'metro' (meter) are regularly preceded by numerals (<num+>). Typically, such information treats "reverse linear precedence", providing information about the *left hand* context. The tag <+num>, by comparison, used with words like 'capítulo' (chapter) or 'número' (number), signals *real* (functional) valency and *right hand* context. Another example for "co-occurrence valency" are <PRP+> tags, where more or less fixed PP-expressions are targeted by providing information about the governing preposition *at its argument nominal*, e.g. 'graça' <de+>, where the assembled PP forms a fairly independent lexical unit, 'de graça' (‘free of charge’).

Accordingly, valency information can be exploited for disambiguation in two ways: a) by "local" rules, typically using close lexical or word class context for morphological disambiguation, and b) by "global" rules, for determining functional dependency. Table (1) attempts to quantify the importance of valency tags for disambiguation on different levels:

(1) number of CG rules containing at least one valency context condition

	Morphological rules		**Syntactic rules**	
	"safe"	heuristic	"safe"	heuristic
verbal valency	17.3%	10.0%	29.8%	28.4%
nominal LP valency <+...>	6.5%	10.9%	4.8%	3.4%
"left valency" <...+>	1.5%	3.0%	1.4%	-

It can be seen, that, on the whole, verbal valency is quantitatively more important to disambiguation than nominal valency, which is not surprising given the fact that all verbs receive valency information, while the figure for nouns and adjectives is only 10% for nouns and 7% for adjectives[117]. In analogy with what is said about the distribution of "global" vs. "local" rules in chapter 3.7.3, verbal valency information is used in a third of all syntactic rules, but only in one sixth of all morphological rules. Since heuristic morphological rules are most likely to lack global contexts altogether, they will obviously also be the ones least likely to make use of verbal valency, since the dependencies concerned cannot be guaranteed to be contiguous. In contrast, syntactic rules need verbal valency information even if they are heuristic (the percentages for safe resp. heuristic syntactic rules are nearly the same).

Nominal valency and left valency, on the other hand, since they are about group structure and lexical neighbourhood, are primarily used for close context morphological disambiguation, a rationale that becomes even clearer for *heuristic* morphological disambiguation.

Quite another aspect of the valency discussion are semantically motivated selection restrictions. At present, the parser lexically assigns unambiguous ±HUM/ANIM head tags to 35.7% of all adjectives, and ±HUM/ANIM subject tags to 48.2% of all verbs in running newspaper text:

<vH> verb with obligatorily human subject ('discutir' - ‘to discuss’)
<vN> verb with obligatorily inanimate subject ('explodir' - ‘to explode’)
<vA> verb with obligatorily animal subject ('coaxar' - ‘to croak‘)
<vB> verb with obligatorily plant subject ('espigar' - ‘to sprout’)
<adj.h> adjective with obligatorily human head ('assassudo' - ‘wise’)
<adj.n> adjective with obligatorily inanimate head ('asséptico' - ‘sterile’)

[117] These are token frequency related numbers for disambiguated running newspaper text. In the PALAVRAS lexicon, the percentage of nouns and adjectives featuring valency information is lower, since many very infrequent nominals lack valency patterns.

<adj.a>	adjective with obligatorily animal head ('carnívoro' - kødædende)
<adj.b>	adjective with obligatorily plant head ('epífito' - epifytisk)

The remaining verbs and adjectives are assigned tags for all 4 possibilities (<sH>, <sN>, <sA>, <sB> for verbs, and <jh>, <jn>, <ja>, <jb> for adjectives), *after* the morphological and syntactic levels, which are then disambiguated on the valency level and used for polysemy resolution in the semantics module.

This way, only the "safe", unambiguous selection restrictions are accessible on the first two levels of disambiguation, and so far (1998), on the morpho-syntactic levels only some 25 rules make direct use of this kind of information, like in

MAP (@>N) TARGET (ADJ) IF (0 <ante-attr>) (-1 <art> LINK 0 MS) (1 INF) (NOT 0 <h>); # *e.g. um leve erguer de ombros.* (Map prenominal function onto an adjective preceded by the male singular definite article and followed by an infinitive, if its doesn't obligatorily select for a human head.)

REMOVE (@#ICL-SUBJ>) (*1 V3S BARRIER @#FS LINK 0 V-HUM); (Remove the subject reading for an infinitive clause, if the next third person singular verb takes a human subject and is not isolated by a finite subclause complementiser.)

On the other hand, the <h> tag for adjectives can be used in order to determine whether an ambiguous NP head noun is +HUM or not, a feature that is more widely used in the grammar, and thus linked to pre-existing rules. The hybrid nominal set HUM-N/A, that lists +HUM semantic class tags proper for nouns alongside with the "left selection" tag <h> for adjectives, is another example of the present - indirect - use of the feature.

3.7.2.2 Level interaction: Secondary semantic tags

One of the big syntactic ambiguities for nouns is the one between subject (@SUBJ) and (direct, "accusative") object (@ACC). Other functions, like appositions (@APP) or argument of preposition (@P<) have a clearer context. Since Portuguese does not have a fixed word order, both subjects and objects can appear before *or* after their main verb, giving rise to the @ACC> - @SUBJ> and @<ACC - @<SUBJ ambiguity. Worse, in the case of embedded subclauses, an NP between two main verbs can also be ambiguous as to clause membership - it may, for instance, be either direct object of the first (subclause-) main verb or subject of the second (main clause) main verb. Sometimes, clause-boundary punctuation helps, but it can be absent (for example, in the case of relative clauses unless they are parenthetic), or be mistaken as an iterator mark (in a chain of co-ordinated subjects or objects). In other cases, the uniqueness principle helps, i.e. there may already be a "safe" - positioned - subject to the left of the first verb or a "safe" object to the right of the second verb. In many cases, however, the contextual clues are much more subtle, and semantic information may be needed to make an educated guess.

Intuitively, one might assume

(a) that a subject reading is more likely *before* the predicator than after it, and
(b) that noun phrases denoting humans, are more likely to function as agent than others, and might therefore have a larger affinity to subject function

Whereas (a) is a syntactic rule and fits in naturally with the CG-rules on the syntactic level, (b) presupposes semantic lexical information, that must be expressed as *secondary* tags, i.e. tags, that are not (on this level!) intended for disambiguation themselves.

In order to test the two assumptions, I have statistically analysed the computer's parses for one and a quarter million words, as shown in table (1). Since shorter, manually controlled texts show the parser's syntactic error rate to be lower than 3% (cf. chapter 3.9), the dubious cases will disappear in a sea of safe correct readings (like those where the uniqueness principle can be applied, or where verbs have *obligatory* direct objects), - and therefore distributional patterns may be trusted even when derived from automatic analysis alone. Even if all errors were subject-object errors (which they are not!), a ±3% margin of statistical significance would not change much in the ratios calculated below.

(1) **The influence of the semantic feature <+HUM> on the probability of subject tags vs. direct object tags** (573.285 words from VEJA, plain numbers, and 690.269 words from the Borba-Ramsey corpus, numbers in italics). Percentages measure the frequency of a given function within a certain semantic group.

	PROP (proper nouns)				**N** (nouns)							
	top (places)		**hum** (names)		**H** (persons)		**HH** (groups)		**inst** (institutions)		**all N**	
	n	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
@SUBJ> presubject	239 *187*	7,9 *9,0*	5409 *3410*	15,3 *18,4*	4200 *3688*	23,3 *21,9*	995 *981*	18,8 *18,5*	597 *436*	10,9 *9,6*	16573 *17811*	12,5 *11,9*
@<SUBJ postsubject	6 *29*	0,2 *1,4*	710 *344*	2,0 *1,9*	566 *421*	3,1 *2,5*	61 *83*	1,2 *1,6*	37 *51*	0,7 *1,3*	2291 *3461*	1,7 *2,3*
@ACC> preobject	2 *2*	0,1 *0,1*	24 *31*	0,1 *0,2*	24 *31*	0,1 *0,2*	0 *3*	0,0 *0,1*	3 *5*	0,1 *0,1*	295 *546*	0,2 *0,4*
@<ACC postobject	90 *89*	2,9 *4,3*	1409 *873*	4,0 *4,7*	2114 *1900*	11,7 *11,3*	526 *503*	9,9 *9,5*	452 *366*	8,2 *8,1*	23279 *26725*	17,4 *17,9*
all words in class	3011 *2084*		35378 *18573*		18037 *16856*		5297 *5291*		5491 *4519*		132673 *149125*	
ratio @SUBJ> @<ACC-	2,7 *2,1*		3,8 *3,9*		2,0 *1,9*		1,9 *2,0*		1,3 *1,2*		0,7 *0,7*	

As to intuition (a), an SVO word order - though not fixed for Portuguese - is definitely preferred, pre-posed noun subjects (@SUBJ>) being at least 7 times more likely than post-posed subjects (@<SUBJ), while pre-posed direct objects are nearly non-existent in the noun class (objects pronouns, of course, are another matter).

More interestingly, "subject-ivity" is higher and "object-ivity" is lower for *human* nouns than for others. The relevant ratio in favour of the @SUBJ> tag (as compared to @<ACC) is highest for names (3.8-3.9) and persons and human groups (both 1.9-2.0). Even the institutions class (1.2-1.3) has a subject/object-ratio twice as high as the noun class as a whole (0.7), which has the opposite tendency - i.e. occurring more often in direct object than in subject position.

With these figures, a purely guessing parser would have a 4-in-5 chance to resolve the @SUBJ>/@<ACC ambiguity for names, and a 3-in-4 chance to resolve it for person or human group nouns.

In absolute terms (i.e. when looking at subject and object probability in isolation), some special cases can be observed in the table:

- proper nouns that are *not* person names - but all place names - have a high subject/object ratio, too, but both subject and direct object[118] readings are less frequent than in average nouns, probably because most incidents are in locative PP-

[118] Portuguese can completely avoid using personal names as syntactic direct objects, by using the preposition *'a'* before them: *Maria ama a Pedro*.

constructions. Something similar is true of the human noun subclass of institutions, that share the semantic feature of +LOC with place names.

- Both person and place names are much more frequent in the VEJA-newsmagazine corpus than in the mixed Borba-Ramsey corpus.
- There is a slightly higher frequency of *post*-positioned subjects for person nouns in the VEJA texts, probably due to journalese quote constructions (e.g. " ...." diz o estudante (@<SUBJ) Alberto da Mata, 25, de São Paulo).

3.7.3 Local vs. global rules: Constraint typology

In this section I shall as far as possible detach myself the CG grammar critic from myself the CG rule writer, inspecting and quantifying the types of rule architecture used in the system, and trying to map and interpret possible system immanent structural regularities or tendencies. The point of this exercise is:

(a) to provide other CG-grammar writers with some *standard for comparison* and CG novices with some guidelines as to how a CG may be expected to develop, what grammar size and complexity to expect, which pitfalls to avoid etc., and

(b) to facilitate *cross-system comparison*, like when the author of a probabilistic HMM tagger/parser wants to decide on the possibility to match or "emulate" a CG rule set (a problem the relevance of which I have personally been confronted with when discussing with NLP-researchers outside the CG camp).

What a CG grammar architecture looks like, may, of course, depends not only on general linguistic and analytic factors, but also on the individual grammarian's approach to grammatical problem solving in general, and the technical limitations imposed by the few presently available CG rule compilers in particular, - and with very few Constraint Grammars around (and even fewer published), real proof of any structural universality claim must therefore await future research. Still, even regularities found within one system (and with one type of compiler), may help other researchers understand why CG rules look the way they do, and how best to learn from their not so bad performance.

One of the ways to assess a given Constraint Grammar in a typological way is to quantify rule types with regard to their contextual scope and complexity, as suggested by Anttila in his discussion of the Helsinki group's English CG (Karlsson et. al., 1995, p.352). Contextual scope is what ordinarily distinguishes probabilistic grammars (narrow scope) from generative grammars (wide scope). Within Constraint Grammar, bounded context conditions, especially of low order (i.e. close to the target), are natural narrow scope tools, whereas unbounded context conditions are characteristic of a wide scope approach. Thus, a CG rule set can be typologically located between probabilistic and generative grammars, mimicking the first for part-of-speech discrimination, and the second for syntactic parsing.

In table (1), a rule count is given for rules with unbounded context conditions (henceforth "global" rules) or without ("local" rules[119]), for all three operations supported by the cg2-compiler. The columns containing numbers for non-heuristic rules

[119] The concept of "local" rules is not to be confused with that of "local disambiguation" - the first term is used to describe rules without unbounded context conditions (i.e. rules where all contexts conditions are bounded), while the second concerns word-internal disambiguation (the minimal derivational complexity rule, or "Karlsson's law")

are shaded, as well as the sum-column. 'morf1-3' and 'syn1-3' refer to the heuristic levels in the morpology and syntax module, respectively.

(1) Rule scope

	morf 0	**morf 1**	**morf 2**	**morf 3**	**syn 0**	**syn 1**	**syn 2**	**syn 3**	**all**
REMOVE tag (only local contexts)	403	112	13	27	153	37	4	2	651
REMOVE word (only local contexts)	66	12	1	18	-	-	18	10	125
REMOVE tag (≥ 1 global contexts)	183	44	5	5	941	219	17	1	1415
REMOVE word (≥ 1 global contexts)	63	16	2	1	4	1	4	-	91
local/global tag	**2.2**	2.5	2.6	5.4	**0.2**	0.2	0.2	2.0	**0.5**
local/global word	**1.0**	0.8	0.5	18.0	**-**	-	4.5	-	**1.4**
SELECT tag (only local contexts)	271	70	8	7	60	2	1	1	420
SELECT word (only local contexts)	162	33	4	7	-	-	-	-	206
SELECT tag (≥ 1 global contexts)	129	23	9	2	209	57	3	-	432
SELECT word (≥ 1 global contexts)	135	73	11	5	-	-	-	-	224
local/global tag	**2.1**	3.0	0.9	3.5	**0.3**	0.0	0.3	-	**1.0**
local/global word	**1.2**	0.5	0.4	1.4	**-**	-	-	-	**0.9**
IFF tag (only local contexts)	3	-	-	-	-	-	-	-	3
IFF word (only local contexts)	7	-	-	-	-	-	-	-	7
IFF tag (≥ 1 global contexts)	-	-	-	-	-	-	-	-	-
IFF word (≥ 1 global contexts)	4	-	-	-	-	-	-	-	4
local/global tag	**-**	-	-	-	**-**	-	-	-	**-**
local/global word	**1.8**	-	-	-	**-**	-	-	-	**1.8**

For the grammar as a whole, REMOVE rules account for two thirds of all disambiguation rules, with a higher incidence for global and tag targeting rules.

For global context syntactic rules (i.e. rules containing at least one unbounded context condition), selecting is even more risky than ordinarily. For example, it is safe to assume that a direct object reading can be removed in a sentence without a transitive verb, while the "inverse", *choosing* the object reading in the presence of a transitive verb, is risky, rules would have to thoroughly check for other direct objects and direct object candidates, clause boundaries and the like.

In the case of tag targeting rules, cautiousness is necessary because a tag-target has to cover a range of possibly quite different lexical items, whereas a word-form target is really the equivalent of a complete tag sequence, including the lexical base form tag. Since very few syntactic rules have word-form targets, the effect is only visible in the morphological rule portion, with a remove/select ratio of 1.5 for tag targeting rules as compared to one of 1.0 for morphological rules on a whole.

The elevated remove/select ratio for syntactic rules (over 4) is not only due to a higher degree of "structural globality" (as addressed by the valency based uniqueness principle), but also to the grammar specific fact that clause function tags have been attached to non-finite verbs and complementiser words (relatives, interrogatives, conjunctions), in addition to these words' clause internal function tag. Since double tag targets are not allowed in *syntactic* SELECT rules in the available cg-compilers, such words can only be disambiguated by REMOVE rules - a SELECT rule targeted at either the internal or the external function tag would "kill" the other of the two.

Apart from low error rates, Constraint Grammar parsers are famous for their processing speed. The actual speed, even when using the same compiler on the same machine, is of course dependent on both text type and grammar size. For text type, the relevant parameters are sentence length and word form ambiguity (average number of readings per word form); for grammar size, parameters are the number of rule contexts (subsuming the number of rules as well as their complexity) and the proportion of unbounded contexts. Since the parser has to apply for every word and every one of its readings all rules that target that reading, a first approximation for sentence processing time would be one of linear complexity:

(3a) time ~ n * a * R

where

n = number of words in the sentence
a = average ambiguity (number of readings per word form
R = rule number constant, depending on, but less than proportional to the number of rules Rn in the grammar

However, since the parser - in applying a rule to the target reading found - must check ("instantiate") all the rule's context conditions as true, the relevant constant is not the number of rules Rn, but the number of context conditions in the grammar, Cn. While obviously slowing down the parser, adding more *absolute* contexts does not change the linear complexity characteristic as such, a good algorithm that avoids checking contexts twice for different rules, may even make R grow slower than Rn. *Unbounded* context conditions, however, force the parser to look, if necessary, at *all* words and their readings in its half of the sentence. Processing time will therefore grow binomially ($(n*a)^2$) with sentence length for that proportion G% of contexts that is unbounded.

(3b) time ~ (n * a * C) * (n * a * G)

where

C = context number constant, depending on, but less than proportional to the number of contexts Cn in the grammar

G = globality constant, depending on, but less than proportional to the proportion of unbounded contexts, G%, in the grammar

Finally, processing time is also proportional to the proportion RM of REMOVE rules, since REMOVE rules have to look at all readings, while SELECT rules, when hitting the right reading (on average by trying half of them), discard all others automatically. Therefore[120], the variable *a* has to be replaced by a*(RM+1)/2 in the first parenthesis of equation (3b). Likewise, *a* in the second parenthesis is influenced by the proportion SC of safe context conditions (NOT and C) in unbounded contexts.

(3c) time ~ n * a * (RM+1)/2 * C * (n * a * (SC+1)/2 * G)

where:

RM = proportion of REMOVE rules

SC = proportion of safe unbounded context conditions

Binomial complexity growth is tolerable, and compares favourably with the exponential complexity growth[121] seen when a parser has to look at all analysis *paths* for a sentence parse (a^n*C).

Having discussed *a* in the chapter on ambiguity, and *G* as well as*RM* earlier in this chapter, I will now try to shed some light on rule complexity (*C*) and context certainty (*SC*).

[120] With SE for the SELECT rule proportion, the formula would be a*RM + a*SE/2, with SE =1-RM we get a*RM + a*(1-RM)/2, which can be transformed into a*(RM+1)/2.

[121] In a probabilistic HMM *PoS tagger* this problem can be solved by not "remembering" all paths, but only the highest probability path when progressing from left to right through the sentence. Complexity will then grow in a linear way (~ n * a * N), with N being the constant reflecting the size of the n-gram window. A probabilistic *syntactic parser*, evaluating whole sentence paths, will, of course, have to deal with the above mentioned exponentiality problem.

In table (4), three types of contexts are subsumed:

- direct contexts, addressed by absolute or unbounded position instantiation, with the target word form as position 0.
- relative contexts, addressed by the LINK feature and related to another, preceding, context functioning as new position 0.
- BARRIER contexts which are always negative, with their scope defined by the instantiation of the unbounded context they refer to.

Most rules, with the exception of some default mapping rules, have at least one direct context, and 75% of the 2739 "global" rules (with at least one unbounded context condition) have LINK or BARRIER contexts, or both. Thus, 2085 rules feature at least one LINK context, and 2017 rules have at least one BARRIER context.

(4) Rule complexity

number of contexts	**morf**				**syn**				**map**	**all**
	morf0	**morf1**	**morf2**	**morf3**	**syn0**	**syn1**	**syn2**	**syn3**		
0	-	-	-	-	-	-	-	-	39	39
1	172	**77**	**16**	**44**	54	8	12	**14**	57	54
2	317	70	14	14	159	26	**15**	-	127	742
3	**383**	62	8	8	**219**	41	3	-	170	**894**
4	249	**73**	2	2	214	46	3	-	**187**	776
5	143	51	10	2	204	**47**	4	-	150	611
6	81	29	2	2	173	42	-	-	101	430
7	37	7	-	-	115	29	2	-	71	262
8	23	5	-	-	58	27	1	-	33	147
9	12	4	-	-	39	12	4	-	23	94
10	4	4	-	-	30	10	2	-	10	60
11	1	-	-	-	27	10	1	-	3	42
12	1	-	-	-	15	10	-	-	2	28
13	1	-	-	-	23	7	-	-	5	36
14	1	-	-	-	17	-	-	-	1	19
15	-	1	-	-	11	1	-	-	-	13
16	-	-	-	-	3	-	-	-	-	3
17	1	-	-	-	-	-	-	-	-	1
18	-	-	-	-	1	-	-	-	-	1
19	-	-	-	-	1	-	-	-	-	1
20	-	-	-	-	1	-	-	-	-	1
21	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	2	-	-	-	-	2
23	-	-	-	-	1	-	-	-	-	1
all rules	1426	383	53	72	1367	316	47	14	979	4657
average per rule	**3.37**	3.40	2.60	1.75	**5.28**	5.75	3.66	1.00	**4.22**	**4.14**

On average, 4 context conditions have to be true before a rule can be successfully applied to its target. Syntactic rules are more complex (5.28 contexts for non-heuristic rules), and morphological rules less complex (3.37 contexts for non-heuristic rules) than the average. Mapping rules display an intermediate degree of complexity (4.22 contexts), on the one hand they apply to morphological targets, on the other they add syntactic structure. Also, adding more context conditions is not the only way to make a mapping rule more safe, a common alternative is to keep some ambiguity, map a longer string of function tags, and leave disambiguation to the syntactic module proper.

Generally, there is less complexity on the higher heuristic levels, reflecting the fact that these rules are less safe, incorporating fewer context conditions that would limit the rule to less general - and thus safer - cases. Still, on the first heuristic level rules appear to be slightly *more* complex than non-heuristic rules, the reason for this being the fact that the heuristic level distinction is also used for non-heuristic purposes: - for all but the mapping rules it is the only way to determine *in which order* rules will be applied. Also, some of the hard, multi-context, cases are postponed to heuristic level 1, because there is a hope that other rules will resolve or restrict the ambiguity in question in some indirect way. This double functionality of heuristic level 1 can even be seen in the statistics in table (4), as a double peak curve in the *morf1* column. The first peak (1 context) reflects pure heuristic uses, like the removal of readings with a <Rare> tag, the other (4 contexts) is related to rule ordering and the postponement of hard cases.

While both a high remove/select ratio and a high percentage of safe contexts (NOT and C) make a grammar more cautious (and robust), they also make the parser a little slower. Among other things, table (5) contains the data necessary to understand the second part of this trade-off, which is related to context type distribution. The relevant parameter, *C-percent*, measures "certainty" and is computed as the ratio between the combined number of NOT and C conditions and the number of*all* contexts at a given position. For the zero position (the target itself) the current cg-compilers do not permit C-conditions, so here, the "safe" portion will consist of the NOT conditions alone.

(5a) Context position, polarity (±NOT) and certainty (±C)
[absolute contexts]

number of contexts	**morf**				**syn**				**map**	**all**
	morf0	**morf1**	**morf2**	**morf3**	**syn0**	**syn1**	**syn2**	**syn3**		
0	554	181	20	53	812	258	36	13	473	2400
NOT 0	268	92	8	2	230	48	4	-	43	695
all 0	*822*				*1042*				*516*	*3095*
C-percent	*32.6*				*22.1*				*8.3*	*22.5*
+1	250	78	2	2	97	15	8	-	220	672
+1C	310	48	1	-	28	1	-	-	4	392
NOT 1	191	66	4	3	62	11	2	-	133	572
all +1	*751*				*187*				*357*	*1636*
C-percent	*66.7*				*48.1*				*38.4*	*59.0*
-1	409	103	21	12	177	29	1	-	468	1218
-1C	381	52	-	6	43	-	-	-	3	485
NOT -1	275	103	13	4	70	20	-	-	62	547
all -1	*1065*				*290*				*533*	*2250*
C-percent	*61.6*				*39.0*				*12.2*	*45.9*
+2	42	18	-	-	32	-	-	-	41	133
+2C	73	5	-	-	24	-	-	-	-	102
NOT 2	53	6	-	-	13	1	-	-	8	81

all +2	*158*				*69*				*49*	*316*
C-percent	*79.7*				*53.6*				*16.7*	*57.9*
-2	122	22	2	5	43	6	-	-	184	384
-2C	98	16	-	-	19	5	-	-	15	153
NOT -2	84	17	1	4	26	3	-	-	26	161
all -2	*304*				*88*				*225*	*698*
C-percent	*59.8*				*51.1*				*18.2*	*45.0*

(5a), continued

number of contexts	morf				syn				map	all
	morf0	morf1	morf2	morf3	syn0	syn1	syn2	syn3		
+3	18	-	-	-	1	-	-	-	10	29
+3C	4	2	-	-	1	-	-	-	-	7
NOT 3	3	-	-	-	1	-	-	-	2	6
all +3	*25*				*3*				*12*	*42*
C-percent	*28.0*				*66.7*				*20.0*	*31.0*
-3	50	5	-	1	7	-	-	-	61	124
-3C	24	1	-	1	2	-	-	-	10	38
NOT -3	12	1	-	1	5	-	-	-	2	21
all -3	*86*				*14*				*73*	*183*
C-percent	*42.4*				*50.0*				*17.8*	*32.2*
≥ +4	8	-	-	-	-	-	-	-	1	9
≥ +4C	8	-	-	-	-	-	-	-	-	8
NOT ≥ 4	4	-	-	-	1	-	-	-	-	5
all ≥ +4	*20*				*1*				*1*	*22*
C-percent	*60.0*				-				-	*59.1*
≤ -4	24	1	-	-	-	-	-	-	10	35
≤ -4C	4	1	-	-	-	-	-	-	-	5
NOT ≤-4	1	-	-	-	7	-	-	-	-	8
all ≤ -4	*29*				*7*				*10*	*48*
C-percent	*17.2*				-				-	*27.1*
all +	318	96	2	2	130	15	8	-	272	843
all +C	395	55	1	-	53	1	-	-	4	509
all NOT +	251	72	4	3	77	12	2	-	143	664
all +	*964*				*260*				*419*	*2016*
C-percent	*67.0*				*50.0*				*35.1*	*58.2*
all -	605	131	23	18	227	35	1	-	723	1761
all -C	507	70	-	7	64	5	-	-	28	681
all NOT -	372	121	14	9	108	23	-	-	90	737
all -	*1484*				*399*				*841*	*3197*
C-percent	*59.2*				*43.1*				*14.0*	*44.9*

Looking at absolute positions first, the following observations can be made:

(a) *The C-percent parameter is higher for morphological than for syntactic rules, and lowest for mapping rules.*

This fact reflects the order in which disambiguation is performed in progressive level parsing, morphology first, then syntax. Thus, the amount of unambiguous context (where no certainty restrictions are necessary) increases from level to level. In particular, mapping rules may - at least within the current cg-compilers - only refer to pre-existing tags, i.e. morphological tags which have all (or nearly all) been disambiguated already, not to syntactic tags introduced by other mapping rules (which ordinarily would be highly ambiguous). Therefore, C-contexts are redundant and very rare in mapping rules. For the same reason, in syntactic disambiguation rules, too, all contexts referring to *morphological* information do not usually need C-tags. And even for syntactic tags the C-option has a handicap: the double tags used for words bearing clause function, which cannot be AND-grouped in the current compiler's set-definitions (both cg1 and cg2 allow only OR-grouping for syntactic tags).

(b) *The C-percent parameter is higher for right hand positions than for left hand positions, and lowest for the zero position. Left hand contexts are more common than right hand contexts of the same distance, and the disparity increases with distance, from ca. 50% for the +1/-1 pair (1636 and 2250 rules, respectively) to 400% for the most distant contexts.*

It is quite hard to find a clear and general explanation for this interesting finding. It seems to imply that for disambiguation, left hand context is more important (or easier to use) than right hand context, and that left-looking rules can be applied before right-looking ones, since the latter would have to wait for the creation of safe right-hand contexts by left-looking rules.

The real reason may even be a psycholinguistic one: Language has evolved as speech, and is therefore processed in a linear way. It will therefore be a communicational advantage, if the listener be able to anticipate the next word or word group, or at least its type and function. Empirical priming tests and the existence of the linguistic garden path problem seem to indicate that, in fact, humans tend to choose that reading for a word that is suggested by its left hand context. So a right hand context has to be "extra safe" in order to be allowed to make a difference.

Since Portuguese valency structures reflect this left-to-right approach, on the syntactic level, where word classes are unambiguous, and functions are ambiguous, arguments can be identified by finding a word of the relevant head word class to the left, like in the case of an ambiguous @ACC after VFIN, or @P< after PRP. In the morphological module, with its more local (narrower) rule scope, (group level) modifiers are more important than (usually clause level) arguments, and it seems logical that articles, determiners, numerals and intensifiers (which all typically precede their head in Portuguese) are more essential to the type of head[122] they attach to, and "need" their head more, than adjectives and prepositional phrases which as modifiers usually come to the right of their head, and *could* be non-group, clause level constituents (predicatives or adverbials). This explains the natural dominance of left hand contexts

[122] Left context determiners, article determiners and numerals help recognize (disambiguate) nouns, immediate left context intensifiers help recognize adjectives and adverbs.

in the sequential “understanding” of an utterance, - and the extra safety tax imposed on right hand contexts.

c) Close context conditions are more common than distant context conditions

This final observation matches one's intuition about structural cohesion, close contexts have a higher probability of being structurally linked to the target than distant ones. Also, when looking at a distant context, it is mandatory to check the closer context in between, too, - for potential clause boundaries or other blocking elements. So the -1 position will be checked both for its own sake, and (also) every time the -2, -3 or -4 contexts are addressed.

All of the above findings are true of unbounded contexts, too:

(5b) context position, polarity (±NOT) and certainty (±C) [unbounded contexts]

number of contexts	morf				syn				map	all
	morf0	**morf1**	**morf2**	**morf3**	**syn0**	**syn1**	**syn2**	**syn3**		
***1**	117	48	9	6	409	70	14	-	238	911
***1C**	48	14	-	1	107	24	1	-	2	197
NOT *1	61	22	2	4	33	2	-	1	10	135
*all *1*	*226*				*549*				*260*	*1243*
C-percent	*48.2*				*25.5*				*4.6*	*26.7*
***-1**	147	44	12	2	565	190	16	-	349	1325
***-1C**	64	23	-	2	275	65	-	-	8	437
NOT *-1	61	27	2	1	64	9	-	-	16	180
*all *-1*	*272*				*904*				*373*	*1922*
C-percent	*46.0*				*37.5*				*6.4*	*31.1*
***2**	30	7	-	-	10	2	-	-	20	69
***2C**	8	-	-	-	3	-	-	-	-	11
NOT *2	4	2	-	-	-	-	-	-	-	6
*all *2*	*42*				*13*				*20*	*86*
C-percent	*28.6*				*23.1*				*-*	*19.8*
***-2**	62	13	4	-	58	2	-	-	77	216
***-2C**	9	4	3	-	3	-	-	-	-	11
NOT *-2	4	1	-	-	6	3	-	-	26	14
all -2	*75*				*67*				*103*	*241*
C-percent	*17.3*				*13.4*				*34.2*	*10.4*
≥ *3	5	2	-	-	-	-	-	-	-	7
≥ *3C	-	-	-	-	-	-	-	-	-	-
NOT ≥ *3	1	-	-	-	-	-	-	-	-	1
*all ≥ *3*	*6*				*-*				*-*	*8*
C-percent	*16.7*				*-*				*-*	*12.5*
≤ *-3	8	2	-	-	5	1	-	-	16	32
≤ *-3C	1	1	-	-	-	-	-	-	-	2
NOT ≤*-3	-	-	-	-	2	-	-	-	-	2
*all ≤ *-3*	*9*				*7*				*16*	*36*
C-percent	*11.1*				*28.6*				*-*	*11.1*
all *+	152	57	9	6	419	72	14	-	258	987
all *+C	56	18	-	1	120	24	1	-	2	208
all NOT *+	66	25	-	4	33	2	-	1	10	142
*all *+*	*274*				*572*				*270*	*1337*
C-percent	*44.5*				*26.7*				*4.1*	*26.2*
all *-	217	59	16	2	628	193	16	-	442	1573
all *-C	74	28	3	2	278	65	-	-	8	450
all NOT *-	65	28	2	1	72	12	-	-	42	196

all -	*356*				*978*				*492*	*2219*
C-percent	*39.0*				*35.9*				*10.2*	*29.1*

However, though the relative distribution remains similar to that of absolute contexts, the proportion of safe conditions (the C-percent parameter) is consistently lower for unbounded contexts. Unbounded contexts are usually defined in a more restrictive way than absolute contexts that only check for one particular tag set. In contrast, as mentioned above, most unbounded contexts have BARRIER or LINK (often even LINK 0) conditions attached that further restrict instantiation of the context. BARRIER conditions are unbounded backwards-looking NOT conditions[123] and are thus part of the safe context group, and many LINK contexts are themselves specified as C. With such a wealth of linked information, the chance of error when instantiating an unbounded context is thus smaller than for ordinary, absolute contexts, and the parser's philosophy is: Rather find an ambiguous word context that matches all the additional, relative context specifications than not use it just because it happens to have another *local* reading not itself sustained by further relative context.

The leftward leaning tendency for unbounded contexts is about the same as for absolute contexts, about 60%.

(6) Proportion of leftward context conditions (% left/all contexts)

	morf0	**syn0**	**map**	**all**
absolute contexts	60.6	60.5	66.7	61.3
unbounded contexts	56.6	63.0	64.6	62.4

Interestingly, this is not what Anttila finds for the English CG. In (Karlsson et. al., 1995, p. 352) he cites 81% for unbounded and 42.6%[124] for absolute contexts, supposedly for the syntactic segment of the English grammar. As an explanation for the high figure for unbounded contexts, Anttila refers to the fact that such rules are about phrase structure generalisations and that, in English, heads usually precede their complements. In the same vein one can argue that Portuguese here displays a lower figure, because its word order is not as strictly regulated as that of English.

Still, for absolute contexts, the Portuguese figure is *higher* than the corresponding English one, and not significantly different from that for unbounded contexts in the Portuguese CG. An explanation may be that the English rules concerned were meant as small window rules akin to heuristic rules (as suggested ibd., p. 353), whereas the

[123] the BARRIER condition was not present in the cg1-compiler. There, it would have to be expressed as a hooked (now: LINKed) unbounded context condition in the opposite direction. But even then, barrier function was intended - unbounded context searches would not be allowed to cross the zero position.

[124] My computation, - the article cites absolute figures, not percentages, for the absolute contexts.

Portuguese rules use *more (distant)* context, even in the unbounded case[125]. Thus, for the English CG, the -1 context accounts for 81.9% of all left contexts, and the +1 context for 87% of all right contexts. For the Portuguese CG, the figures are 72.7% and 71.9%, respectively, for syntax, and 70.4% and 81.2% for the whole grammar. The fact that the -1/left percentages are lower than the +1/right percentages for both languages, suggests that the left hand context is not only overrepresented, but also extends further away from the target position, - both possibly due to the linearity feature of language discussed above.

[125] Of course, the Portuguese rules may use a larger window and *still* be just as "heuristic" as their Englich counterparts, if it could be shown that Portuguese *needs* a larger window due to lower structural cohesion.

3.8 Mapping: From word class to syntax

Rules at the mapping level of a Constraint Grammar exploit (more or less) disambiguated morphological/PoS information for assigning a context dependent syntactic function potential to each target word in the text. Rules can address words individually, but are usually bundled for certain types of word class targets. Contextually safe mapping rules are able to map a more precise syntactic tag list (ideally, one tag only) than more broad rules. Therefore, unlike disambiguation rules, which *remove* information rather than add it, mapping rules are inherently sequential and mutually exclusive: Safe, specific, context rich rules have to be applied before more general, poor context rules, and once targeted, a word has to be "closed" for further mapping. Otherwise, every word will receive the full combined syntactic tag potential of all mapping rules targeting it, which would "erase" the visibility of any individual, more specific rule. In the rule compiler formalism used here, ordinary sequential mapping rules are marked by the MAP operator, and they are applied in the order given in the rules file of the grammar. Rules with the alternative ADD operator are cumulative and, in principle, non-sequential. Basically, ADD rules provide a way of splitting a complex MAP rule into smaller, more manageable parts.

There is no clear border line between mapping rules and (syntactic) disambiguation rules. In theory, all mapping rules could be crafted with a perfect and complete list of context conditions such that no mapping would need to be ambiguous - with no need for ordinary disambiguation rules. However, in none of the presently available rule compilers can mapping rules "see" the output of other (earlier) mapping rules, making it difficult if not impossible to address *syntactic* context (@tags) other than that provided by lexicon entries. Also, a perfect (i.e. unambiguous) MAP rule is like a SELECT rule in the way it works - a risky kind of rule, stating a grammatical "fact" all in one go. REMOVE rules, operating on broadly mapped - and therefore *ambiguous* - @tag strings, are much more cautious and robust, working together step by step, relying on each other's context condition safety nets.

The basic skeleton of syntactic mapping is the target word class condition. Even without further context conditions, word class mapping can provide a working mapping module for a syntactic disambiguation CG to work on. The Portuguese mapping rule set is structured in word class "chapters", with each chapter concluded by a "pure" word class mapping rule, preceded by more specific rules for that word class, and headed by a section with word or base form mapping rules. Though there are some prototypical relations, most form-function pairs (PoS-@tag pairs) are not very closely knitted. Thus, nouns are typical of subject (@SUBJ) and direct object (@ACC) function, but still, subjects do come as infinitive clauses (#ICL), too, and objects can be finite subclauses (#FS, "acho que não faz nada"). Adjectives and participles often occur with adnominal and predicative function, but they can head noun phrases, too, and thus usurp typical

NP-functions like @ACC or @SUBJ ("os ricos vivem bem"). Adverbial function (@ADVL), finally, is by no means restricted to adverbs - prepositional phrases ("fica em casa", "espere até amanã") and even nouns ("chegou segunda-feira", "dormia dez horas") can fulfill this function. Also, form-function relations are not necessarily symmetrical. Prepositional object function (@PIV), for instance, is only mapped onto prepositions (PRP), but the inverse is not true, since prepositions also occur as (heads of) argument and adjunct adverbials (@ADV, @ADVL), post-adjects (@N<, @A<, @KOMP<) and bound or free predicatives (@SC, @PRED, @N<PRED).

In the table below, I have listed, for each syntactic function label, the maximal set of word classes eligible as mapping targets. Prototypical mapping targets are in bold face.

Syntactic tags mapped: Word class targets for mapping:

Clause level arguments

@SUBJ =subject, @ACC =direct ("accusative") object, @DAT =indirect (dative) object, @PIV =prepositional object, @ADV =adverbial object, @SC =subject predicative complement, @OC =object predicative complement

@SUBJ> @<SUBJ	[**N PROP** A PCP **PERS-nom SPEC** DET #ICL #FS]
@ACC> @<ACC	[**N PROP** A PCP **PERS-acc SPEC** DET #ICL #FS]
@DAT> @<DAT	[**PERS-dat**]
@PIV> @<PIV	[**PRP**]
@ADV> @<ADV	[N-temp/quant **ADV PRP-loc/dir** #FS-onde #AS-onde]
@SC> @<SC	[N PROP **ADJ PCP** SPEC DET PRP #ICL #FS-que/interr]
@OC> @<OC	[N PROP **ADJ PCP** SPEC DET PRP #ICL #FS-que/interr]

cp. @#ICL-SUBJ> @#ICL-<SUBJ @#FS-SUBJ> @#FS-<SUBJ
cp. @#ICL-ACC> @#ICL-<ACC @#FS-ACC> @#FS-<ACC
cp. @#FS-ADV> @#FS-<ADV @#AS-<ADV
cp. @#ICL-<OC @#ICL-<SC @#FS-<SC

Clause level adjuncts

@ADVL =adjunct adverbial, @PRED =free (adjunct) predicative

@ADVL> @<ADVL	[N-temp **ADV PRP** #ICL #FS #AS]
@PRED> @<PRED	[N-indef/attr **ADJ PCP PRP**]

cp. @#ICL-ADVL> @#ICL<ADVL @#FS-ADVL> @#FS-<ADVL @#AS-ADVL> @#AS-<ADVL

Unbound utterance level constituents

@NPHR =isolated nominal expression, @ADVL =isolated adverbial expression, @VOK =vocative

@NPHR	[**N PROP ADJ** PCP SPEC DET]

@ADVL	[N-temp **ADV PRP** #ICL #FS #AS]
@VOK	[**PROP**, <poss 1S> + N]

cp. @#ICL-ADVL @#FS-ADVL @#AS-ADVL

Argument or modifier adjects in NP

@>N =prenominal adject (modifier), @N< =postnominal adject (modifier or argument), @APP =apposition, @N<PRED =postnominal nexus predicative, @PRED =free (adject) predicative

@>N	[**DET** A PCP ADV-focus]
@N<	[N-attr PROP **ADJ PCP** DET-post **PRP** #ICL #FS]
@APP	[**N-def PROP**]
@<PRED	[N-indef/attr **ADJ PCP PRP**]
@N<PRED	[ADJ PCP PRP GER, after "com/sem"]

cp. @#ICL-N< @#FS-N<

Argument or modifier adjects in AP (including attributive participle clauses)

@>A =adverbial (intensifier) preadject, @A< =adverbial postadject (intensifier or argument), @ADVL>A - @A<ADVL - @A<ADV - @A<PIV - @A<SC = "adjuncts" and "arguments" in attributive post-nominal participle "clause", @A<PASS =passive agent after attributive participle

@>A	[**ADV-intensifier** ADV-focus]
@A<	[PRP ADV-demais #AS]
@ADVL>A	[ADV-temp/loc PRP-temp/loc]
@A<PASS	[**PRP-por**]
@A<ADVL, @A<ADV	[**ADV-temp/loc PRP-temp/loc**]
@A<PIV	[**PRP**]
@A<SC	[N **ADJ**]

cp. @#AS-A<

Argument or modifier adjects in PP

@P< =argument of preposition, @>P =intensifier or focus modifier of PP

@P<	[**N PROP** A PCP **PERS-piv SPEC** DET ADV-loc/temp #ICL #FS]
@>P	[ADV-focus ADV-intensifier]

Argument of complementiser in averbal subclause

@AS<	[**N PROP A PCP PERS SPEC** DET PRP #ICL]

cp. @#ICL-AS<

Verb chain elements

@FAUX =finite auxiliary, @FMV =finite main verb, @IAUX =non-finite auxiliary, @IMV =non-finite main verb, @PRT-AUX< =auxiliary particle in verb chain

@FAUX	[**VFIN**]
@FMV	[**VFIN**]
@IAUX	[**INF GER PCP**]
@IMV	[**INF GER PCP**]
@PRT-AUX<	[**PRP** KS-que]

Conjunctions

@CO =co-ordinator, @SUB =subordinator, @KOMP< =argument of comparative, @COM =comparative subordinator, @PRD =predicative subordinator

@CO	[**KC**]
@SUB	[**KS**]
@KOMP<	[PRP-de #FS **#AS**]
@COM	[KS-que/do=que ADV-como/quanto/qual DET-quanto/qual]
@PRD	[**ADV-como**]

cp. @#FS-KOMP< @#AS-KOMP<

Finite subclauses

@#FS-SUBJ> @#FS-<SUBJ	[KS ADV-interr SPEC-rel/interr DET-rel/interr]
@#FS-ACC> @#FS-<ACC	[KS ADV-interr SPEC-rel/interr DET-rel/interr]
@#FS-ADV> @#FS-<ADV	[ADV-rel DET-rel]
@#FS-<SC	[KS ADV-interr SPEC-rel/interr DET-rel/interr]
@#FS-P<	[KS-que ADV-rel/interr SPEC-rel/interr DET-rel/interr]
@#FS-ADVL> @#FS-<ADVL	[KS ADV-rel]
@#FS-N<	[ADV-rel SPEC-rel]
@#FS-KOMP<	[KS-que/do=que ADV-como/quanto/qual DET-quanto/qual]
@#FS-S< "sentence apposition"	[SPEC-que/o=que]

Non-finite subclauses

@#ICL-SUBJ> @#ICL-<SUBJ	[INF]
@#ICL-ACC> @#ICL-<ACC	[INF]
@#ICL-<SC	[INF]
@#ICL-<OC	[INF]
@#ICL-ADVL> @#ICL-<ADVL	[INF GER PCP]
@#ICL-N<	[INF]
@#ICL-P<	[INF]
@#ICL-AUX< "argument of auxiliary"	[INF GER PCP]
@#ICL-AS<	[GER PCP]

Averbal subclauses

@#AS-A<	[ADV-rel]
@#AS-<ADV	[ADV-rel]
@#AS-KOMP<	[KS-que/do=que ADV-como/quanto/qual DET-quanto/qual]
@#AS-ADVL> @#AS-<ADVL	[KS-app ADV-rel]

3.9 Performance: Measuring correctness

3.9.1 Training texts

Working on "known" bench mark texts of 10-20.000 words, by constantly testing rule performance on manually introduced <Correct!> - markers, the Portuguese morphological tagger (analyser and disambiguator together) can be geared to resolve nearly all ambiguity while retaining a 99.9% correctness rate. For unknown texts, results are obviously lower. Yet, performance on training texts is not irrelevant, since it shows that the CG approach does not suffer from system immanent interference problems to the same degree as, say, a probabilistic tagger based on a pure trigram HMM, where (to my knowledge) even retraining and measuring on the *same* corpus seldom yields more than 97% correctness, even for parts of speech.

Aiming at maximal precision, I have also worked on a larger, untagged text (170.000 word from the Borba-Ramsey corpus) on both the morphological and syntactic levels. Though it wasn't possible single-handedly to produce manually tagged benchmark-corpora of that size, or to fully inspect the outcome of an automatic tagging run, it still made sense automatically to extract and quantify surviving ambiguities after tagging runs, since *precision* (defined as the percentage of surviving readings, that are correct) can be approximated by minimising ambiguity, at least as long as intermittent bench mark runs ensure that new rules discard few correct readings, and the ambiguity percentage thus still remains high in comparison with the other factor in the precision calculus, error frequency. With a PoS error rate of 1%, for instance, and 10% two-fold ambiguity, precision would compute as 99/110 = 90%, and cutting ambiguity in half (while retaining the same error rate) would entail a nearly equivalent improvement in precision (99/105 ≈ 94.3%). Surviving ambiguity, then, easily measured without manual control on any text corpus, can be used as an approximate guide to how precision is progressing during the grammar writing process. In contrast, *recall* (defined as the percentage of correct readings, that survive disambiguation) has - in the absence of a large tagged and proof-read Portuguese corpus for measuring - to be calculated manually on smaller sample texts.

When forcing the parser into full disambiguation, where all words - with the exception of the rare cases of true ambiguity - end up with one reading only, recall and precision will obviously assume identical values, and one can regard the recall/precision figure as a direct measure for the parser's performance, which is why I will henceforth use the more general term *correctness* to mean *recall/precision at 100% disambiguation.*

3.9.2 Test texts

During the project period I have done some such correctness evaluation on unknown texts, too. These test runs, while being fairly small, consistently suggest a correctness rate of over 99% for morphology and part of speech, when analysing unknown unrestricted text. For syntax the figures are 98% for classical literary prose (Eça de Queiroz, "O tesouro") and 97% for the more inventive journalese of newspaper texts (VEJA, 9.12.1992), as shown in table (1) below. At evaluation time, modifiers were tagged for dependency (adnominal adjects @>N, @N< and adverbial adjects @>A, @A<), but no functional subdifferentiation (like @A<PASS or @N<PRED) had been introduced. Of the 54 word/group function errors in the first test run, 13 concerned modifiers and 11 involved adjuncts (@ADVL>, @<ADVL, @PRED>, @<PRED), while nearly half (25) were mistaggings clause-level arguments (of verbs). In 3 cases verbal function itself was misanalysed, and 2 errors concerned the argument of a preposition (@P<).

(1) Correctness and error distribution for unknown prose fiction and news texts

Text:	*O tesouro*		*VEJA 1*		*VEJA 2*	
	ca. 2500 words		ca. 4800 words		ca. 3140 words	
Error types:	errors	correctness	errors	correctness	errors	correctness
Part-of-speech errors	16		15		24	
Base-form & inflexion errors	1		2		2	
All morphological errors	17	**99.3 %**	17	**99.7 %**	26	**99.2 %**
syntactic: word/group function	54		118		101	
syntactic: subclause function	10		11		13	
All syntactic errors	64	**97.4 %**	129	**97.3 %**	114	**96.4 %**
"local" syntactic errors due to PoS/morphological errors	- 27		- 23		- 28	
Purely syntactic errors	37	**98.5 %**	106	**97.8 %**	86	**97.3 %**

A contrasting run on another two whole articles with two different subjects (video games and arts), did not yield much topic dependent variation in the error rates:

(2) Correctness and error distribution for different news topics

Text:	**"VEJA"**	**"VEJA"**	**all**
	(videogames)	**(art)**	

	2412 words		1837 words		4249 words	
Error types:	errors	% correct	errors	% correct	errors	% correct
Morphology (all)	29	**98.8 %**	7	**99.6 %**	36	**99.2 %**
unknown English words in headlines	- 10 - 3		- 1 - 0		- 11 - 3	
Morphology (pure)	16	**99.3 %**	6	**99.7 %**	22	**99.5 %**
Syntax (all)	66	**97.3 %**	46	**97.5 %**	112	**97.4 %**
syntax caused by morphology	- 37		- 7		- 44	
Syntax (pure)	29	**98.8 %**	39	**97.9 %**	68	**98.4 %**

3.9.3 Text type interference and tag set complexity

However, a closer look at the texts involved reveals that the news texts are quite different from the prose fiction example, both lexically and syntactically. First of all, there is a rather high percentage of complex names (e.g. 'Massachussets Institute of Technology'), abbreviations ('MIT') and English loan words and vogue terms like 'joy stick', 'bad boy' and the like. Thus a single word, *console*, which - used as an unknown English noun ['video console'] and not as a Portuguese verb ['to comfort'] - is responsible for a third (!) of all errors in the video game text. Second, VEJA news texts are - syntactically - very rich in free predicatives (typically information about persons, institutions or abbreviations, like age, place, definition etc.) all acting as false "argument candidates" , as well as other types of parenthetical information, bracketing, head lines and interfering "syntactically superfluous" finite verb forms in the form of quotations, which all tend to blur the clause boundaries that otherwise would be important structural information for the parser.

Still, none of the above problems are in principle intractable for the CG-approach, and by providing for special features like these in the rule set (and lexicon) error rates can be reduced for any text type.

One might assume that errors are evenly spread throughout the text, which would - for an average sentence length of 15 words - mean about one morphological error in every tenth sentence, and a syntactic error in every third. However, this is not true: for all text types, errors appear in clusters, obviously most morphological errors also appear in the list of syntactic errors, and many syntactic errors interfere with readings in their neighbourhood, due to rules that depend on clause boundary words, uniqueness principle and so forth. Thus, a V-N word class error not only affects syntactic mapping and disambiguation for the word in question, but can cause 2 or 3 syntactic errors around it, by providing faulty disambiguation context. This clustering tendency of syntactic errors is good news both for the overall robustness of the result (there are many unaffected sentences, which are completely error free), and for the work of the grammarian: mending the grammar at one point may remove a whole chain of secondary interference errors. Likewise, when seen in isolation, - that is, when supplied with error-free morphological input -, the syntactic parser on its own can yield even better results. Thus, for VEJA newspaper texts, the correctness rate will rise by 0.5-1.0 %, to about 98%.

Also, when comparing the above correctness figures to the results of other approaches, one has to bear in mind the complexity of the tag set and the information content of the categories used. Thus, the attachment and functional information that my parser provides for prepositional phrases (such as post-nominal adject @N< , post-adjectival/adverbial adject @A<, adjunct adverbial @<ADVL, @ADVL>, adverbial @ADVL, adverbial object @<ADV, @ADV>, prepositional object @<PIV, @PIV>, subject complement @<SC, free predicative @<PRED, complementiser argument @AS<) can potentially give rise to numerous errors, that would just not be visible if all these tags were collapsed into a bare syntagmatical 'PP' (prepositional phrase) or a rudimentary "functional" '@ADVL' (adverbial). Thus, in the last two VEJA texts, error pairs *inside* the PP-group account for 15 cases, or 22%, of the purely syntactic errors.

3.10 Speech data tagging: Probing the limits of robustness

3.10.1 Text/speech differences in a CG perspective

In terms of test texts, the probably most difficult task for the parser has been a pilot project on the tagging of transcribed speech data (Bick, 1998-2) from the NURC corpus of Brazilian educated urban speech (Castilho, 1989). While the morphological/PoS tagger module, with a success rate of around 99% even without additional rules, proved quite robust in test runs on spoken language data, syntactic analysis fared somewhat worse, with an initial correctness rate of 91-92% for the - rule-wise - unmodified system (cp. 3.10.5).

In order to explain this discrepancy between morphological robustness and syntactic failure, a number of hypotheses were formulated and subsequently put to the test by changing the system's preprocessor module and CG rule set accordingly:

- In my parser, rules with **morphological targets** mostly use a **shorter context range** (group structure) than those with syntactic targets (cf. chapter 3.7.3). Thus, the proportion of rules without and with unbounded contexts is 10 times as high for rules targeting morphological tags than for syntactic targets, and 70-80% of all syntactic rules stretch their context all the way to the sentence delimiters – making these rules vulnerable to the speech specific absence or vagueness of such delimiters.
- **Incomplete utterances** tend to leave group structure intact more often than clause structure, - at least if one doesn't count repetitional modifications/corrections of prenominal modifiers *(essas esses progressos, esta este caminho, da dos nomes)*, where word class adjacency rules can often override agreement rules.
- Speech data **lacks punctuation** and has **unclear sentence window** borders, which is especially bad for syntactic CG analysis which tends to use many unbounded context restrictions (cp 1).
- Speech data is filled with "**syntactic noise**", repetitions and false starts of one- or two-word chunks, as well as pause and phatic interjections *(ahn, uh, eeh etc.)*.

3.10.2 Preprocessing tasks

In order to make these problems more accessible to Constraint Grammar rules, and to improve syntactic performance, a preprocessor was designed with the specific goal of establishing utterance or sentence boundary candidates and removing syntactic noise. Its task areas are the following:

1. Orthography and layout normalisation (character set, line numbers)

2. Repetitions and false starts (automatically commented out by $-signs)

mas é vo/ voluntária né?
becomes: -->*mas é <$vo/> voluntária né?*
então então vem tudo aquilo de cambulhada e im/ e im/ im::POSto sobre nós
becomes:--> *então <$então> vem tudo aquilo de cambulhada e <$im/> <$e> <$im/> <stress> imposto sobre nós*

3. Phonetics

* Vowel length markers are removed, e.g. *u::ma pessoa* --> *uma pessoa*
* In-word stress marking is commented out, e.g. *esnoBAR* --> *<stress> esnobar*

4. Introducing "dishesion marker candidates" (eee)

* Due to a complete lack of full stops, colons and commas (only question marks [?] and turn taking [¶] are used), other means of marking syntactic windows become necessary, and strings like '...', 'eh', 'éh', '()' are marked as "dishesion elements", as well as quotes if they enclose more than 1 word. Dishesion marker candidates are subsequently mapped as

a) <break> (major syntactic break, clause or sentence boundary)
* <break> markers can be used by the CG rules to establish maximal group size or valency scope; e.g., <break> should not occur between a premodifier and its head, or between main verb and direct object.

b) <pause> (non-word hesitation/pause marker)
* <pause> markers are not allowed to break up group og clause continuity.

The preprocessor also performs a certain degree of dishesion marker disambiguation (leaving part of the job to the CG rules proper which are better at handling complex contexts). To this end, the following (very local) rules are employed:

a) "xxx" --> eee xxx eee --> <pause> xxx <pause>
If a single word is surrounded by dishesion markers, these are treated as <pause>

b) eee (e) que/quando/embora ... --> <pause>
If a dishesion marker is followed by a conjunction or relative, possibly with an interfering coordinator, it is treated as <pause>.

c) que/quando/embora ... eee --> <pause>
If a dishesion marker is preceded by a conjunction or relative, it is treated as <pause>

d) eee + PRP --> <pause>
If a dishesion marker is followed by certain prepositions (de, em, com, sem, por), it is to be treated as <pause>

e) PRP/det + eee + NON-art/dem --> <pause>

If a dishesion marker is preceded by a preposition or a determiner (or a fused presposition+determiner), it is to be treated as <pause>, unless it is directly followed by an article or demonstrative (in which case the <pause>/<break> ambiguity is retained)

3.10.3 Grammar tasks

The next adaptation effort concerned the CG rule grammar as such, where dishesion marker candidates had to be integrated in those tag sets that denote possible syntactic breaking points. The PAUSE set, for example, includes not only the dishesion marker, but only certain interjections:

LIST PAUSE = "uhn" "ahn" "eh" "eee" <pause> <break> IN ;

The <break> tag is useful in NON-sets since these are often used in BARRIER conditions in CG-rules, baring group attachment, for instance:

LIST NON-NP = PERS SPEC ADV VFIN INF PRP KS KC <rel> <interr> "<$\,>" <break> >>> <<< ;

On the sentence level, <break> is a potential clause boundary marker, the same way certain complementizers, comma and hyphen are:

LIST CLB = KS <interr> <rel> "<$\,>" "<$->" KOMMA <break> ;

Also, rules had to be crafted for further disambiguation of cohesion markers, deciding whether to treat them as breaks denoting “sentence” window borders, or just as pauses embedded in the syntactic flow of speech.

For instance, dishesion markers are not <break> (but <pause>) if they intervene:

(a) between a “name bearer” and its name: *o rei $$ Alfonso*
(b) between a noun and the preposition ‘de’: *pai $$ de muitos filhos*
(c) between an intensifier and an attribute: *uma maneira um pouco $$ calcada*
(d) between a noun and a potential postmodifier or object complement of the same gender and number: *estou vendo **a TV** evidentemente $$ muito **presa** a ...*
(e) between a transitive main verb and its direct object.

These cases translate into the following CG-rules, where rule (a’) relates to example (a) etc.:

(a’) REMOVE (<break>) (-1 (<+n>)) (1 <*>)
(b’) REMOVE (<break>) (-1 N) (1 PRP-DE)
(c’) REMOVE (<break>) (-1 <quant>) (1 ATTR/<attr>)
(d’) REMOVE (<break>) (*-1 NFP BARRIER ALLuPAUSE/ADV) (*1 ATTR-FP BARRIER ALLuPAUSE/ADV)
(e’) REMOVE (<break>) (-1C @MV LINK 0 <vt>) (*1C @<ACC BARRIER @NON->N)

Of course, the use of dishesion markers and their introduction in NON-sets and CLB-sets, has to be balanced between the advantages of providing better defined analysis windows, and the draw-backs of disallowing many long range rule contexts that have BARRIER conditions involving CLB-items and NON-sets.

While disambiguated dishesion markers help to establish the kind of “syntactic chunking” essential to any CG grammar, a number of specific speech data problems remained to be treated directly by rule additions or rule changes:

a) Premodifier clashes (da dos)

In a simple correctional article clash (‘comeu a o bolo’) both articles will receive the @>N (premodifier) tag, but in more complex cases there may be problems, for instance, where a preposition is repeated as well. Here, the first determiner will be analysed as @P< (argument of preposition).

eu	não	estou	agora	por	dentro	**de**	**a**	**de**	**os**	nomes	sabe	?
SUBJ>	ADVL>	FMV	ADVL>	<SC	P<	A<	**P<**	N<	>N	P<	FMV	

b) “Faulty” noun phrases: stranded premodifiers in incomplete np’s and agreement errors

In the parser’s output, stranded premodifiers (here: ‘um, uma’) tend to assume np-head function in a syntactic parse, which may seem odd, but is hard to avoid, and may well be the logical solution - after all, in a word-based tagger/parser there are no zero constituents, and every function has to be attached *somewhere.*

Another np-problem for the syntactic section of the parser is the risk of a long distance between head and modifier resulting in agreement lapses as in the gender clash below (‘codificação nada normativo’). Also this variation is probably more commen in speech than in text.

(i)

e	$e	não	havendo	uma	codificação	não	$pause
CO		ADVL>	IMV ICL-ADVL>	>N	<ACC	ADVL>	

$break	$eee	**um**	**uma**	$pause	nada	**normativo**
		<ACC	<ACC		>A	N<

In the speech data in question, agreement failure (here SG - PL) *does* occur in adjacent position, too. The examples are taken from a transscription where the speaker (a lecturerer) admitted to being nervous on being taped

(ii)

a	demanda	de	moeda	por	transação	$pause	é	$pause

>N	SUBJ>	N<	P<	N<	P<	FMV

principal	**motivo**	por	**os=quais**	as	pessoas	$pause	retêm	moeda
>N	<SC	ADVL>	P< FS-N<	>N	SUBJ>		FMV	<ACC

(iii)

nós	podemos	resumir	isso	**em**	**um**	**exemplinhos**	numérico
				<ADVL	>N	P<	N<

c) Difficulties in identifying subjects:

Consider the following example, where three subject tags have to be found and tolerated in the same speech chunk without clear clause boundaries:*televisão, ela*, *telespectador:*

porque	a	**televisão**	sendo	estatal	**ela**	é	muito	$stress
SUB FS-<ADVL	>N	SUBJ>	IMV ICL-<ADVL	<SC	SUBJ>	FMV	>A	

uniformizada	$pause	$break	não	há	espectáculos	diversificados	o
<SC			ADVL>	FMV	<ACC	N<	>N

telespectador	$pause	$break	**o**	fica	sempre	$pause	preso
SUBJ> <ACC			ACC>	FMV	<ADVL		<SC

a	filmes	ou	a	$a	conferências
A<PIV	P<	CO	<PIV		P<

Here, **‘ela’** is semantically anaphoric to ‘televisão’, which syntactically belongs to its own non-finite subclause. **‘telespectador’** lacks a sentence/analysis window marker (before its article), which is why function has not been fully disambiguated in this case.‘**o**’ before the main verb ‘fica’ might be part of yet another subject candidate with only its article left, but since the grammar strongly disallows adjacency of articles and finite verbs, ‘o’ is treated as a personal pronoun in the accusative. ‘o’ does not bear any meaning in this sentence, and would be ignored by a human listener, but once uttered and transscribed, the word has to be handled in the grammar one way or another.

d) Synatctic speaker interaction and overlap in multi-speaker data

In a notation that uses only one time line, utterances of speaker S2 may syntactically “cut” an utterance of speaker S1. Also, speakers S1 and S2 may interact syntactically, finishing each others groups or clauses. In the example, ‘adequado’ (S1) is subject complement (SC) for ‘está’ (S2), ‘perfeitamente’ (S2) is premodifier (>N) for ‘adequado’ (S1):

S2 para aquele ... está ***perfeitamente*** *...*
S1 ***adequado***
S2 ***adeQUAdo:: do*** *... é muito mais interessante ... é uma*
[
L1()
L2 grande oportunidade para os nossos artistas nåo é ?
L1 isso é muito bom:: eh:: e ain/ e:: e a novela puxa o disco porque parece que na vendagem dos discos eles såo muito ... requisitados esses discos de novelas né ?

This last problem can only be addressed superficially by altering the CG rules set. A thorough solution would probably have to involve a harmonisation of transcription conventions and the CG formalism.

3.10.4 Positive side effects: Robustness

As also discussed in chapter 4, CG's flat dependency analysis is quite robust, and as a "side effect" often nicely handles unclear clause/sentence boundaries or nested sentences, both of which are frequent in speech data. Consider the 5 main verbs in the following comma- and coordinator-free sentence:

e	**é**	uma	grande	atriz	$break	então	**choca**	demais	$pause	$break
CO	**FMV**	>N	>N	<SC		ADVL>	**FMV**	<ACC		

aquela	paulista	$stress	quatrocentona	**que**	ele	**faz**	bem	$stress
>N	<SUBJ		N<	ACC> FS-N<	SUBJ>	**FMV**	>A	

grifado	$break	aliás	de	uma	maneira	um=pouco	$pause	calcada
<OC		ADVL>	ADVL>	>N	P<	>A		N<

demais	**porque**	esse	tipo	**acho**	**que**	já	se	**diluiu**
A<	SUB	>N	SUBJ>	**FMV**	SUB FS-<ACC	ADVL>	ACC>	**FMV**

nem	**existe**	mais	$pause	mas	...
<ADVL	**FMV**	<ADVL		CO	

Even double main verb constructions[126] without any sensible traditional syntactic analysis, and breaches of the uniqueness principle are tolerated fairly well by the CG-grammar:

$break	isto	**é**	**levava**	a	um	tipo	de	vida	nômade
	SUBJ>	**FMV**	**FMV**	<PIV	>N	P<	N<	P<	N<

[126] One possible integral analysis of the example given makes 'é' not a main verb, but a focus marker particle.

Where all goes well, the system tolerates overlapping clauses with double uncoordinated subjects and a shared direct object, as well as - to a certain degree - complex and interrupted np's and np-modifiers (boxes):

		problems:
papai	N M S @SUBJ>	
mesmo	DET M S @N<	
tem	V PR 3S IND @FMV	obligatorily transitive verb *without* direct object
em	PRP @<ADVL	
os	DET M P @>N	
<$nos>		
livros	N M P @P<	
de <sam->	PRP @N<	
ele <-sam>	PERS M 3S NOM/PIV @P<	
ele	PERS M 3S NOM/PIV @SUBJ>	2 subjects without co- or subordination
tem	V PR 3S IND @FMV	2 main verbs without co- or subordination
muitas	DET F P @>N	
expressões	N F P @<ACC	direct object serving verbs in 2 clauses
$pause		
completamente	ADV @>A	
caídas	V PCP F P @N<	heavy postnominal with adjunct and argument
em=desuso	VPP @A<PIV	
e	KC @CO	
portuguesas	**N** F P @<ACC**??**	less heavy postnominal *after* heavy postnominal
e	KC @CO	
<$por/>		
e	KC @CO	
$pause		
de	PRP @SC> @N<	very distant pp-postnominal with false start
português	N M S @P<	
clássico	ADJ M S @N<	
não	ADV @ADVL>	
é	V PR 3S IND @FMV	finite clause without clause boundary item
$?		

3.10.5 Evaluation

A quantitative comparison of the two versions of the parser (before and after adaptation to speech data) yielded the following results, with correctness defined as *recall at near 100% disambiguation*, counting both false tags, missing tags and false ambiguity as errors.

Parser performance on speech data (before/after grammar adaptation)
(NURC *[norma lingüistica urbana culta]*, São Paulo)

data sample	**sample size**	**morphological correctness**	**syntactic correctness**
2 speaker dialogue (topic: cinema, television, actors) females, 60 yrs (journalist - writer)	2810 words	99.2 %	95.7 %
secondary school teaching monologue (history), female 36 yrs	2080 words	99.5 %	96.3 %
university teaching monologue (economics), male 31 yrs	1600 words	99.0 %	95.4 %
unadapted parser speech base line: 2 speaker dialogue (same as above) analysed with unmodified grammar	1100 words	98.9 %	92.6 %
written text parser base line: typical performance on VEJA texts (cp. 3.9.2)	-	98.8 - 99.7 %	96.4-97.4 %

Providing for some incertainty due to the relatively small size of the individual test sample, the above performance table seems to indicate that the *unadapted* CG parser, though originally designed for written Portuguese, was able to more or less maintain its performance on speech data *morphology* (word class etc.), while error rates tripled for speech data *syntax*..

Judging from the effectiveness of according rule changes and preprocessing in the *adapted* parser, one can conclude that at least one of the reasons for this considerable difference between morphological and syntactic robustness resides in the fact that the disambiguation of morphological ambiguity involves mostly short range group context that is left intact even in the grammatically often incomplete utterances of spoken language, while rule based syntactic analysis depends on long range context patterns, working less than perfect without a clear sentence window, without full complementation of obligatory valency, and with breaches of the uniqueness principle. The hypothesis was tested by tagging - through a preprocessor module - what I call *dishesion markers* ("...", "eh" etc.) in the corpus as both <pause> and <break> for later disambiguation, thus introducing "sentence boundary" candidates, which may be disambiguated by either crude word form context or elaborate long range CG rules. Once disambiguated, the <break> markers provide more "traditional" syntactic window delimiters for the system's Constraint Grammar, considerably improving syntactic tag recall. Examples where modification of the syntactic rules as such proved necessary are violations of the uniqueness principle due to iterations or modified ("corrected") iterations, or cases, where one speaker complements the valency pattern of a syntactic unit uttered by another speaker. Especially problematic are clashes, where a speaker strands dependents without their heads (for instance, subjects without a verb, or a premodifier without its nominal head) and departs on a new syntactic path.

All in all, the preliminary quantitative results suggest that break markers and rule modifications can narrow the gap between the parser ‘s performance on written and spoken Portuguese, respectively, to a few percentage points (i.e. 95-96% correctness) for syntax and nearly eliminate it for part of speech tagging.

4

The syntactic level: A dependency description of Portuguese

4.1 Flat functional dependency grammar

4.1.1 Dependency markers and function tags: Syntactic form and function

In its essence, CG embraces a robust disambiguating philosophy, which does not build a specific sentence structure, but carves away what cannot be part of any structure. That way neither the carving method (rule system) nor the carving tools (rule compilers) are determined by the Constraint Grammar idea as such. And even less the finished sculpture. Every carpenter is free to apply his own beauty ideals. Or isn't he? Which *kind* of Constraint Grammar should he choose?

Historically, CG has its roots in morphological analysis, most systems run with a two-level morphological analyser (TWOL) as preprocessor, and focus on morphological features and parts of speech. Therefore, information is traditionally word-bound and coded as tags (to be attached to words). "Flat" grammar is a natural consequence of this, and my parser, too, makes use of a "flat" representation of syntactic structure.

The description contains information about both *syntactic function* (e.g., arguments like @SUBJ, @ACC) and constituent structure *(syntactic form)*. The latter is expressed by so-called dependency markers (<, >) which point towards the head of the syntactic unit concerned, assembling the constituent into a coherent whole, with implicit constituent borders. Where the head is not the main verb, it will be marked at the arrow point (e.g., N for nominal head, A for adject-head[127]). If there is a function tag (e.g., @<SUBJ, @ADVL>, @N<PRED), the dependency marker arrow's base will be attached to that tag. Otherwise, where function is implied directly by modifier status, the dependency marker base is left tag-less (e.g. @>N for [modifier-] prenominals).

```
(5)    Temos      [ter] <vt> V PR 1P IND VFIN        @FMV
       em         [em] <sam-> PRP                    @<ADVL
       este       [este] <-sam> <dem> DET M S          @>N
       país       [país] <top> N M S                  @P<
       uns        [um] <art> DET P S                  @>N
```

[127] In this terminology, adject heads are the nuclei of Aps (adjective phrases) or ADVPs (adverb phrases). Attributively used participles are included in the adject class, too.

```
castelos    [castelho] <hus> N M P        @<ACC
muito       [muito] <quant> ADV             @>A
velhos      [velho] ADJ M P                @N<
```

This way each word needs only "remember" its own immediate "upward" dependency relation (i.e. what the word itself is dependent of), and all of a sentence's syntactic structure can be described *locally* (in the form of word bound tags), - as in a mobile, where every thread (only) "knows" exactly 2 of the mobile's many moving parts: at one end the bar it is attached to (the head to which a dependency marker points), and at the other the object (or bar) which it holds (the dependent, from which the dependency marker points away). It is enough to note for every piece in the mobile to which other piece it attaches, and one will be able to cut the whole thing into pieces, store it in a shoe box, and reassemble it next Christmas, - without losing structural information[128].

While the mobile metaphor nicely captures the high degree of constituent order mobility in Portuguese sentences, a two-dimensional shadow projection of the mobile would yield "frozen" (dependency-) tree diagrams for individual sentences, and the description should ultimately contain all the structural information needed to draw PSG-like syntactic trees, too (cp. 4.6.3).

In (5), *muito* is located far down in the mobile, but it "knows" its 'adverbial-adject- (@>A) thread-link' to *velho*. This in turn is attached leftward as a 'postnominal' (@N<) to *castelo*. *Castelo* , itself, knows that it is direct object (@<ACC) of a 'main verb' to the left (<), *temos*, which functions as root in the dependendy mobile.

Without special dependency links, such a flat description works fine only as long as individual words bear all of a syntactic unit's functional burden. The description may well get into trouble when more complex dependency relations are involved. Thus, a CG-description without subclause-[function]-tags is bound to suffer from shortcomings like the following:

- 1. Clause boundary markers (or their rule context equivalents) are not hierarchically motivated, so there may be problems with unclear clause continuation after, e.g., centre embedded relative clauses.
- 2. Certain valency features may be left "unsatisfied", e.g. missing subjects in English (*'Visiting the Louvre was not his only reason for coming to Paris')*, or missing accusative objects ('that/que/at'-clauses after "cognitive" verbs).
- 3. Surplus arguments due to unclear clause level resolution, like in *'O perigo de os inimigos atacarem à noite era imanente.'*, where both *perigo* and *inimigos* are

[128] The idea to both mark and process structural information locally (at the word level), is at the very heart of CG's syntactic philosophy, and I will discuss below some of the advantages (and draw backs) of such a "flat" description, hopefully showing how even more complex dependencies (subclauses etc.) can be handled this way.

subjects, but the second subject can only be fully structuralised by attaching its main verb *(atacarem)* as clausal infinitive argument to the preceding preposition *(de).*

- 4. Reduced information content as compared to tree structures (cp. above).

I believe that, by distinguishing between CG as a disambiguation technique, on the one hand, and the descriptional system to be carved, on the other hand, some kind of flat representation can be designed that is functionally equivalent to tree structures, and can express argument and valency structures in a hierarchical way.

My approach has been (a) to add attachment direction markers to *all* argument tags, and (b) to apply double tags to the central linking word in subclauses , - that is, to the “complementiser” (subordinating conjunction, relative or interrogative) in finite and averbal subclauses, and to the infinitive, gerund or participle in non-finite subclauses[129]. These words, then, bear both an “internal” tag (@...) which describes their function inside the subclause, and an “external” tag (@#...), that describes the function of the subclause as a whole when integrated into the next higher level in the clause hierarchy of the sentence. Technically, the disambiguation process works on two lists of @- and @#-tags, respectively, so that internal and external function tags can be treated individually.

(6)	Sabe	[saber] <vq> V PR 3S IND	@FMV	
	que	[que] KS	**@#FS-<ACC**	**@SUB**
	os	[o] <art> DET M P		@>N
	problemas	[problema] N M P		@SUBJ>
	são	[ser] <vK> V PR 3P IND		@FMV
	graves	[grave] ADJ M/F P		@<SC

[@FMV = finite main verb, @#FS-<ACC = finite subclause, functioning as direct (accusative) object attached to a main verb to the left, @SUB = subordinator, @>N = prenominal modifier, @SUBJ> = subject for a main verb to the right, @<SC = subject complement for a (copula) verb to the left, V = verb, KS = subordinating conjunction, DET = determiner, N = noun, ADJ = adjective, PR = present tense, IND = indicative, 3S = third person singular, 3P = third person plural, M = male, F = female, S = singular, P = plural, <art> = article, <vq> = cognitive verb, <vK> = copula verb]

Let's look at a more complex example: *O baque foi atenuado pelo fato de sua mulher ter um emprego que garante as despesas básicas da família.* The analysis in (7) explains how dependency relations can assemble a sentence's building bricks into hierarchical structure. The boxes mark (from the outside in) the main clause, a passive complement, a non-finite subclause (functioning as preposition-argument) and a finite

[129] Another method for functional tagging of subclauses is described by Voutilainen (1994). Here it is the main verb, that bears the subclause's tag (...@), while dependency relations are made more explicit by introducing markers for subclause borders, and by distinguishing between arguments of finite and non-finite main verbs, respectively. Tapanainen (1997) has developed a dependency grammar proper, which is built upon a CG-based morphological disambiguation. Here, heads and dependents are linked by identifier numbers on the tag line.

subclause (functioning as postnominal attributive). NPs are shaded, and the syntactic macrostructure is shown to the left.

(7)

SUBJ	o	[o] <art> DET M S **@>N** 'the'
	baque	[baque] <cP> N M S **@SUBJ>** 'fall'
VP	foi	[ser] <x+PCP> V PS 3S IND VFIN **@FAUX** ‘was’
	atenuado	[atenuar]<vt><sN>V PCPMS **@IMV@#ICL-AUX<**‘buffered’
PP-PASS	por	[por] <sam-> <+INF> <PCP+> PRP **@<PASS** 'by'
P<	o	[o] <-sam> <art> DET M S **@>N** ‘the’
	fato	[fato] <ac> <+de+INF> N M S **@P<** ‘fact’
PP-N<	de	[de] PRP **@N<** ‘of’
SUBJ	sua	[seu] <poss 3S/P> DET F S **@>N** ‘his’
	mulher	[mulher] <H> N F S **@SUBJ>** ‘wife’
VP & **ICL-P<**	ter	[ter] <vt> <sH> V INF 0/1/3S **@IMV @#ICL-P<** ‘having’
ACC	um	[um] <quant2> <arti> DET M S **@>N** ‘a’
	emprego	[emprego] <stil> <ac> N M S **@<ACC** ‘job’
SUBJ & **FS-N<**	que	[que] <rel> SPEC M/F S/P **@SUBJ> @#FS-N<** ‘that’
	garante	[garantir]<vt><v-cog>V PR 3S IND VFIN **@FMV**‘guarantees’
ACC	as	[a] <art> DET F P **@>N** ‘the’
	despesas	[despesa] <ac> N F P **@<ACC** ‘expenses’
	básicas	[básico] <jn> ADJ F P **@N<** ‘basic’
PP-N<	de	[de] <sam-> PRP **@N<** ‘of’
P<	a	[a] <-sam> <art> DET F S **@>N** ‘the’
	família	[família] <HH> N F S **@P<** ‘family’

finite relative subclause, postnominal modifier
non-finite (infinitive) subclause, argument of preposition
preposition phrase, agent of passive adjunct
finite main clause

[@>N =prenominal modifier, @SUBJ> =subject (of verbal constituent to the right), @FAUX =finite auxiliary (head of the verb chain), @IMV =non-finite main verb, @AUX< =argument of auxiliary, @<PASS =passive agent, @P< =argument of preposition, @<ACC =direct (accusative) object (of main verb to the left), @N< =postnominal modifier, @FMV =finite main verb]

The word chain below shows how a dependency grammar "upward attachment sequence" can be constructed by moving from the lowest level (here the article *a*) to the highest level, the verb chain in the finite main clause ('>' means "attaches to", a colon means "makes"):

a > família:NP > de:PP > despesas:NP > garante:FS > emprego:NP > ter:ICL > de:PP > fato:NP > por:PP > atenuado:ICL > foi:S

[NP =noun phrase, PP =prepositional phrase, FS =finite subclause, ICL =non-finite subclause, S =main clause]

4.1.2 Dependency relation types: Clause and group structure

In my dependency notation the following attachment rules are implied for attachment markers (< = head to the left, > = head to the right):

- Clause arguments, e.g. @SUBJ>, @<ACC, @<SC, @#FS-<ACC, attach to the nearest @MV to the left (<) or right (>).[130]
- Adnominal adjects, @>N, @N<, including clausal ones (like @#ICL-N<, @#FS-N<), attach to the nearest NP-head (i.e. a N PROP ADJ DET or INF, that is not an adnominal itself). SPEC and PERS allow only post-adnominal adjects. @>N may point to an NP-head, that functions as an adverbial adject [adverbial modifier], i.e. is not itself an adnominal adject):*um @>N professor @NPHR* ***um @>N tanto @>A*** *iconoclasta @N<.*
- Adverbial adjects, @>A, @A<, including clausal ones (like @#ICL-A<, @#FS-A<), attach to the nearest ADJ PCP ADV or N-<attr>.
- "Forward" free predicatives, @PRED> refer to the following @SUBJ>, even when incorporated in the VP. "Backward" free predicatives, @<PRED, refer to the nearest NP-head to the left, or to the nearest @SUBJ to the left. In the first case @<PRED functions a group level modifier, in the second it is a clause level adjunct.
- Appositions, @APP, attach to the preceding NP-head.
- Verb chain arguments of auxiliaries are marked @#ICL-AUX<, and refer to the nearest auxiliary to the left. Mediating prepositions in verb chains are tagged @PRT-AUX<, also referring to the closest auxiliary to the left. The rightmost part of a verb chain is the main verb (@IMV @#ICL-AUX<). Intermediat auxiliaries are themselves tagged as verb chain argument (@IAUX @#ICL-AUX<). The leftmost auxiliary is the verb chain head (@FAUX or @IAUX).

In the following, I want to distinguish (a) between clause and group level constituents, and (b) between valency bound arguments and free constituents at these levels. Group level constituents will be called *adjects*, independent of their valency status, while at the clause level valency bound constituents will be called *arguments*, and free constituents

[130] In terms of agreement, it would make sense to have subjects attach to the first (finite) verb in the verb chain, - main verb or not. Thus, a finite auxiliary would have two arguments, (a) the subject and (b) its auxiliary argument, the latter consisting of a non-finite subclause comprising all the other constituents of the sentence, centered around the non-finite main verb. Semantically, however, even the subject is still subject to the selection restrictions of the main verb, and in a Portuguese grammar allowing cfor omplex heads, the whole verb chain could well be regarded as one constituent (the predicator) functioning as head of both the subject and all other arguments. However, the flat Constraint Grammar dependency notation as such does not force this distinction, leaving it to grammatical add-on filters (e.g. 7.2 and 4.6.3).

adjuncts. By combination of (a) and (b) 4 main types of constituent structures result, exemplified in table (1), and then discussed individually.

(1) **Table: Constituent structure types**

	a) clause arguments (bound clause constituents)	**b) group arguments (bound adjects)**	**c) clause adjuncts (free clause constituents)**	**d) group modifiers (free adjects)**
valency	valency bound	valency bound	not valency bound	not valency bound
uniqueness principle[131]	valid (unless co-ordinated)	valid (unless co-ordinated)	not valid	not valid
focus by extraposition (clefting)	possible	impossible	possible	impossible

a) Clause, argument structure:

Argument	Head	Argument
João PROP **@SUBJ>**	*come* V VFIN **<vt>** @FMV	*carne* N **@<ACC**
João	eats	meat.
	quer V VFIN **<x>** @FAUX	*jogar* V INF **@#ICL-AUX<**
[He/she]	wants to	play

b) Group, argument structure:

Argument	Head	Argument
	rico ADJ **<+em>** @N</<SC	*em* PRP **@A<** *ouro* N **@P<**
	rich	in gold

Dependency relations are, in the case of arguments, marked at upper "end" of the strings of the mobile: The dependency head word bears a valency marker; a tag like <vt> ('monotransitive verb'), for instance, "expects" a direct object (@ACC) somewhere in the clause[132]. *Rico em ouro* is an example of how the description handles cases with several hierarchical levels: the preposition *em* functions as head in a PP (which is

[131] The principle says that within one clause or group, there must not - without co-ordination - be more than one argument with the same syntactic function. For example, the main verb in a clause may not govern more than one direct object. The uniqueness principle holds explicitely for *arguments*, and can not be applied to other - free - constituents (here called adjuncts).

[132] In a purely syntactic context, however, valency markers are regarded as secondary, in contrast to the primary @-tags, and a word can boast a long list of (potential) valency markers, and still be syntactically unambiguous, with only one @-tag. It is only the @-tags that *have to* be disambiguated at the syntactic level. Still, disambiguation of valency markers can be very useful at a higher level of analysis, where the objective is polysemy resolution (cp. 6.2.3).

complemented by the @P< mark on *ouro*), while at the same time itself playing the role of argument (here @A<) for the AP's head *rico* (marked for this valency trait with <+em>).

In contrast to English or Danish, a pronominal subject can in Portuguese be incorporated into the finite verb (e.g. *quer jogar*), and must therefore be described not as clause constituent, but as (optional and valency bound) constituent of a "wider VP". One could say that the Portuguese clause does not consist of two equal constituents, subject and predicate, but of a head (the "smaller VP" or verb chain) and argument or adjunct dependents, among them the subject.

c) Clause, adjunct structure:

Adjunct	Head	Adjunct
Ontem ADV **@ADVL>**	*ele* PERS @SUBJ> *veio* V VFIN <ve> @FMV	*muito* ADV @>A *tarde* ADV **@<ADVL**
Yesterday	he came	very late.
Zangada PCP **@PRED>**,	*saíu* V VFIN @FMV	*sozinha* ADJ **@<PRED**
Annoyed	[she] left	alone.

Free adjuncts are not valency bound, and dependency is therefore only marked at the dependent: adjunct-adverbials (@ADVL[133]) point towards the main verb, and free (adjunct-) predicatives point towards a nominal group (often, but not always, the subject, which again can be incorporated into the finite verb).

d) Group, modifier structure:

Prenominal modifier adject	Head	Postnominal modifier adject
O DET <art> **@>N** *grande* ADJ **@>N**	*poeta* N M S	*fluminense* ADJ **@N<**
The big	poet	from Rio.
	caro ADJ M S	*demais* ADV **@A<**
	expensive	too [expensive]
mais ADV <quant> **@>A**	*interessado* PCP M S	
more	interested	

Modifiers are those dependents that have the closest link to the group head, and one might argue that a modifier's syntactic function is exactly, and only, this - modifying.

[133] Valency bound circumstantial adverbials (both adverbs and PPs) are tagged as @ADV (adverbial object), like also nominal arguments of time, place and quantity. PPs, that cannot be replaced by simplex adverbs, are tagged @PIV (prepositional object).

Any further, more detailed, syntactic function (attributive, quantifier etc.) is already obvious from the word's word class tag[134] and its lexeme specific semantic traits. This is why I have decided to abide by "mere" dependency marking, i.e. not to add a functional tag at the base side of my dependency marker arrows.

[134] One could say that even certain group types can be defined in terms of modifier word class rather than head word class. Thus, while prepositions both head and define PP's, I find it more modifier based definitions attractive for NP's (which can be definded as article or DEP allowing groups), and AP's (the traditional adjective or adverb groups, which can be defined as intensifier allowing groups). This way 'os ricos' and 'o fazermos nada' become NP's, despite being headed by an adjective and an infinitive, respectively.

4.1.3 The clause: Arguments and adjuncts

@SUBJ> @<SUBJ	subject
@ACC> @<ACC	accusative (direct) object
@DAT> @<DAT	dative (indirect) object (only pronominal)
@PIV> @<PIV	prepositional (indirect) object
@ADV> @<ADV	adverbial object (place, time, duration, quantity)
@SC> @<SC	subject predicative complement
@OC> @<OC	object predicative complement
@ADVL> @<ADVL	adjunct (free) adverbial
@PRED>	"free" (subject) predicative adjunct
@<PRED	"free" predicative (subject) adjunct or predicative post-adject

All above clause arguments [@SUBJ, @ACC, @DAT, @PIV, @ADV, @SC, @OC] and the adverbial complements [@ADVL] attach to the nearest main verb to the left [<] or right [>]. @PRED has dependency attachment to the main verb and its subject (clause level), or to the closest nominal head to the left (group level adject @<PRED)

@ADVL	stray adverbial (in non-sentence expression)
@NPHR	stray noun phrase (in non-sentence expression without a top node verb)
@VOK	"vocative" (e.g. "free" addressing proper noun in direct speech)
@S<	sentence apposition ('não venceu *o que* muito o contrariou')
@CO	co-ordinating conjunction
@SUB	subordinating conjunction

	head	*arguments*	*adjuncts*
VP	@MV main verb (the main verb can be a) finite [@FMV], or b) non-finite [@IMV], i.e. INF, PCP, GER, when b1) part of a verb chain (formally a complex VP-head), @#AUX< after @AUX, or b2) head in a non-finite subclause [@#ICL])	@SUBJ subject @ACC direct object *adorar ACC* @DAT indirect object ***lhe** dou um presente* @PIV prepositive (prepositional object) *gostar **de + NP**,* *contar **com + NP*** @ADV adverbial object *morar ADV* @SC subject complement *ser/estar/parecer SC* @OC object complement *achar alg. OC*	@ADVL adverbial (time, place, quantity, quality, manner) *na França,* *cada dia,* *atenciosamente* @PRED free predicative *nadava **nua***

As I have tried to define word classes by purely morphological criteria (avoiding syntax wherever possible), I will now try and define my syntactic categories by formal and syntagmatic criteria, avoiding semantics wherever possible. Parts of the following discussion are inspired by Perini (1989), who provides a detailed discussion of formally defined syntactic categories for Portuguese.

First, clause level function can be distinguished from group level function by **clefting**[135]**:**

Foi **um grande lobo** que comeu a menina. (@SUBJ)
*Foi **grande** que um lobo comeu a menina. (@>N)

Next, we have to distinguish between (valency bound) arguments and (free) adjuncts, which can be achieved by the **predicate isolation** test. When the predicate is replaced with ”dummy verbs” like ’fazer’ (for non-ergative verbs) or ’acontecer’ (for ergative verbs), predicate-internal arguments are covered, while adjuncts are not. Thus, only adjuncts can be isolated and appear alongside the predicate dummy:

Pedro confiava **na mulher**.	*O que fazia na mulher?	(argument-PIV)
Pedro dormia **no carro**.	O que fiz no carro?	(adjunct-ADVL)
A rainha morreu **em 1690**.	O que aconteceu em 1690?	(adjunct-ADVL)

Subjects are not part of the predicate and pass the predicate isolation test with the predicate-dummy ’fazer’ (1). They do fail it, however, with ’acontecer’, the reason being that the ’acontecer’-dummy includes both predicate and subject. With ergative verbs, the (patient) subject is part of the predicate, and consequently the test only sounds ”natural” with ’acontecer’ – and fails (2a). It can, however, be forced with ’fazer’ even with ergative verbs (2b):

1 - Os romanos construiram casas altas.
O que os ramanos fizeram? - Construiram ..
2 - A rainha morreu em 1690.
2a - *O que a rainha aconteceu?
2b - (?) O que a rainha fiz? - Morreu ....

Perini (1989) suggests **fronting** as a test to distinguish between on the one hand object complements (which he calls “predicatives”) and on the other hand subjects, subject

[135] Constituents from subclauses can not normally be clefted on main clause level, and clefting can therefore be used to test the ”tightness” of a verb chain. Compare the following cleftings where ’no sertão’ can be read either as where the city is to be built (main clause attachment to the verb chain ’VFIN + construir’) or as the place where the “wanting”/”suggesting” of town-construction takes place (attachment to VFIN with ’construir’ isolated on subclause level):

Foi no sertão que ia construir uma cidade.
Foi no sertão que quis construir uma cidade.
Foi no sertão que propôs construir uma cidade.
Foi na sertão que os viu construir uma cidade.

complements, direct objects (the last two lumped together as “objects”) and free subject predicatives (which he calls “attributes”). Of all clause level arguments discussed here, OC is the only one that can not be topicalised by fronting:

Os antigos (SUBJ) consideram este monte um lugar sagrado.
Este monte (ACC) os antigos consideram um lugar sagrado.
*Um lugar sagrado (OC) os antigos consideram este monte.
Sagrado (SC) não é mais, mas alto ainda.

In four cases, **agreement** can be used as a criterion for establishing a function category:

Constituent function	Agreement base
SUBJ	agrees with the predicator (number and person)
SC and PRED	agree with the subject (number and gender)
OC	agrees with the direct object (number and gender)

A constituent passes the agreement test if either (a) its inflexion agrees with the agreement base concerned, or (b) it can be co-ordinated with a word or group that features such agreement. Thus, exceptions like ’As guerras na África (SUBJ-plural) são uma desgraça (SC-singular)’ can be made to work: ’As guerras na África são [**eternas** (feminine plural) **e**] uma desgraça’.

In most traditional systems of syntax, case is interpreted as a function marker, and in Romance languages, **pronoun substitution** is used to elicit case in the face of case-less nouns and np-groups. I have made this case-relation explicit by using case-abbreviations in some of my CG function tags. For @SUBJ and @ACC, pronoun substitution seems to be a straightforward test. Subjects can be replaced by nominative pronouns (ele, ela, eles, elas, eu ...), direct objects by pronouns in the accusative (o/a/os/as). Some @ACC, however, do not allow pronoun substitution:

Comeu 7 bolos.	Os comeu.
Detesta lavar-se.	?O detesta.
Comeu peixe.	*O comeu.
Comeu bananas.	*As comeu.

The problem in the examples with ’peixe’ and ’bananas’ is that pronouns are referential, while the word forms used are generic. ’lavar-se’ is problematic, since it is a non-nominal object. In order to make substitution tests work in these and other cases, they should be read as:

”If a constituent X can be (a) replaced with Y or (b) co-ordinated with a constituent, that – on its own – *could* be substituted with Y.”

Thus, we get ”Comeu peixe/bananas [e 7 bolos]” and ”Detesta [o seu chefe e] lavar-se”, where the np’s in square brackets do allow pronominal substitution with o/a/os/as, and in the second case also with *a quem* (’detesta ao chefe’), which is a special interrogative pronoun test for *human* (+HUM) direct objects (@ACC) and dative objects (@DAT).

If @DAT is defined not *as* a dative pronoun, but as a constituent which can be substituted by one, a pp introduced by the preposition ’a’ – or in some cases, ’para’ – could qualify as a dative object. Likewise, some pp’s with ’a’ as a head and a +HUM ”body”, should qualify as @ACC, if replaceable by ’o/a/os/as’:

Deu o presente ao pai. (@PIV or @DAT)
Não ama mais a mim. (@PIV or @ACC)

Prepositional objects (@PIV) could then be defined as frontable pp-arguments (predicate isolation test) that can *not* be replaced by a (any) pronoun, rather than simply as frontable pp-arguments that cannot be replaced by *adverbial* pronouns.

Subject complements (@SC) can be replaced by ’o’, but no inflected forms are allowed. Object complements cannot be replaced by ’o’, and in addition to non-frontability and object agreement this is a third formal criterion for distinguishing object from subject complements:

Alexandra é muito linda (@SC), mas a sua irmã não o (*a) (@SC) é.
Ela considera a oferta (@ACC) uma ofensa (@OC), mas o seu marido não *o/a (@OC) considera a oferta (@ACC).

Both subject and object complements allow interrogative completion with (o) que, like subject and objects, while free predicatives don’t – they behave like (manner) adverbials in this respect, selecting ’como’:

Ele parece doente. Ele parece o que? (@SC)
O considera perigoso. O considera o que? (@OC)
Nadava nua. Nadava como/*o que? (@PRED)
Nadava depressa. Nadava como/*o que? (@ADVL)

The pronoun substitution test can be applied to adverbials, too. It works fine for all argument adverbials (@ADV) and many ”heavy” adjunct adverbials (@ADVL). ’Assim’ covers manner adverbials, ’lá’ space adverbials, ’então’ time adverbials, and ’tanto’ covers noun- or np-adverbials after transitive measuring verbs like ’custar’ and ’durar’. Ignoring verbal constituents, and postponing, for the moment, the definition of other adjunct adverbial subcategories, we can construct the following table of formal definitions of clause level function:

test	SUBJ	ACC	DAT	PIV	SC	OC	ADV	ADVL	PRED
clefting	+	+	()	+	+	+	+	+/-*	+
predicate isolation: fazer	+	-	-	-	-	-	-	+	+
acontecer	-	-	-	-	-	-	-	+/-*	-
fronting	+	+	()	+	+	-	+	+/-*	+
agreement	VFIN	-	-	-	SUBJ	ACC	-	-	SUBJ
pronoun substitution	ele/ela etc. (nominative)	o/a etc. (accusative)	() lhe etc.	-	o/¬a (tal)	- (tal)	assim lá então tanto[136]	assim etc. -*[137]	-
interrogative completion	o que quem	o que a quem	-	-	o que qual	o que ?qual	como quando onde quanto	como etc. -*	como -

There are, however, adverbials (or what I would like to call adverbials) that do not pass the pronoun substitution test, and even those that do, don't all behave in the same way syntactically. Consider:

Provavelmente ele segunda-feira não foi ao banco de bicicleta.
ADVL-1 ADVL-2 ADVL-3 ADV ADVL-4
[Probably he Monday didn't go to the bank by bicycle]

In this sentence, only 'ao banco' cannot be isolated from the predicate (*o que fez ao banco), therefore it is an argument (ADV). 'segunda-feira' and 'de bicicleta' are "ordinary" time-place-manner adverbs that can be clefted, fronted and replaced by (adverbial) pronouns. Still, there is a difference: First, adverbial 2 can replace 4, but not vice versa – the syntagmatic position between subject and predicator is forbidden for

[136] np-arguments after 'durar' [7 semanas], 'custar' [7 dólares], or 'nadar' [7 quilômetros] can be replaced by 'tanto', but *not* by 'a/o/as/os', which makes them adverbials (ADV in the case of 'durer' and 'custar', ADVL in the case of 'nadar') rather than direct objects.

[137] The category of adjunct adverbial must be seen as opposed to both ADV and PIV. ADVL differs from both ADV and PIV in that it isn't valency bound to the verb (i.e. passes the predicate isolation test), but it covers *both* pp-constituents that can be replaced by adverbs (assim, lá, tanto, as in ADV) *and* pp-constituents that cannot (pp's that would be called PIV if they failed the isolation test).

type-4 adverbials. Second, adverbial 2 passes the predicate isolation test with both ’fazer’ and ’acontecer’, while adverbial 4 fails it with ’acontecer’:

O que fez/aconteceu segunda-feira? (type 2)
O que fez/*aconteceu de bicicleta? (type 4)

The difference is that the ’acontecer’-dummy is a statement-dummy, not a mere predicate-dummy. Adverbials isolated by the ’acontecer’-test therefore must modify – or ”contextualise” - the whole statement, and not any constituent part of it - like the subject in the case of @PRED, or the predicator in the case of type 4 adverbials.

ADVL-1 and ADVL-3 are examples of what I will call meta-operator adverbials (’provavelmente’, ’dubitavelmente’) and set-operator adverbials (’não’, ’até’, ’só’), respectively. Operator adverbials do allow neither clefting nor pronoun substitution. The difference is that meta-operators allow fronting[138], while set-operators don’t. Also, set-operators forbid other kinds of adverbials to appear between themselves and the clause’s predicator, while meta-operators (like non-operator adverbials) *can* be separated from the predicator by other adverbials (type 2 and 3). A third kind of operator-adverbials (i.e. ”non-cleftables”) are time-operators (’ainda’, ’de=novo’, ’mal’, ?’frequentemente’) which *can* be fronted like meta-operators, but – like set-operators – don’t tolerate non-operator adverbials between themselves and the predicator.

For a further discussion of adverbial function as well as of lexical types of adverbs, cp. chapter 4.5.4.

[138] As a group, only operator adverbs do not allow fronting, though some semantically ”result-related” adverbs like ’totalmente’, ’completamente’ etc. don’t either.

4.2 Group types and group level function

@>N prenominal adject, usually determiner
(attaches to the nearest NP-head to the right, that is not an adnominal itself)
@N< postnominal adject, usually adjective, PP or relative clause
(attaches to the nearest NP-head to the left, that is not an adnominal itself)
@>A adverbial pre-adject, usually intensifier
(attaches to the nearest ADJ/PCP/ADV or attributively used N to the right)
@A< adverbial post-adject (rare, e.g. 'caro demais')
@APP “identifying” apposition, usually definite nominal (always after NP + comma)
@<PRED free predicative post-adject (usually adjective or participle), - or predicative adjunct
(refers, as an adject, to the nearest NP-head to the left;
as an adunct, it refers to a main verb and its subject to the left)
@N<PRED postnominal nexus predicative in small clause introduced by 'com/sem'
(rare, e.g. 'com a mão *na bolsa*', 'sem o pai *ajudando*, não conseguiu';
in constituent grammar also used for predicative adject @<PRED)
@P< argument of preposition

Groups (or phrases) are here defined as syntatic constituents that are not clauses, and consist of more than one word. In order not to be clauses, none of the group’s immediate constituents must be a verbal constituent (a predicator) or a complementiser (subordinator). The typical word class inventory of a group’s head and dependents defines the group’s form category. Here, two hypotactic groups will be recognized, **np** and **ap**, plus the katatactic group of **pp**. The concept of **vp** will on the Constraint Grammar level be substituted by the linear concept of **verb chain** (ch. 4.3.), and **paratactic groups** (co-ordinated units) will only be CG-marked after the syntax module proper, by mapping a secondary tag onto the co-ordinating conjunction (e.g. <co-acc> for linking 2 direct object conjuncts, <co-subj> for linking 2 subject conjuncts). While pp’s are easy to define in terms of a heading preposition, np’s and ap’s are more unpredictable with regard to their head. Thus, instead of the prototypical noun head, an np can also feature infinitives, adjectives and even determiners as heads:

(1a) o **comermos** mais carne do que nunca ‘our eating more meat than ever’
(1b) os **pobres** da África ‘the poor in Africa’
(1c) **os** que vi ontem ‘those I saw yesterday’

I would like to argue that what really evokes the concept of np in these cases, is *not* the head, but the prototypical dependents (@>N or @N<): Articles, in (1a-b), or a relative clause, in (1c).

Applying this same defining principle to the other traditional hypotactic groups, adjective phrases (2a, 2d), adverb phrases (2b) and determiner phrases (2c), reveals that they all share one, and only one, common type of modifier (@>A, @A<): Intensifiers like ‘muito’ (‘very’), ‘extremamente’ (‘extremely’) or ‘nada’ (‘not at all’):

(2a) muito/nada **religioso**	‘very religious’, ‘not religious at all’
(2b) muito/nada **devagar**	‘very slowly’, ‘not at all slowly’
(2c) muito **poucos**	‘very few’
(2d) **rico** demais	‘too rich’
(2d) muito **com pressa**	‘in great haste’

All these cases share their internal group structure, and I will therefore lump them together as ap’s (“adpositional” phrases). The concept of ap can even handle otherwise cumbersome complex group types like (2d), where a pp is premodified by an intensifier[139].

4.2.1 The nominal group or noun phrase (NP)

	head	*argument adjects*	*modifier adjects*

[139] In traditional CG dependency notation, the intensifier will first be flatly marked as @>P (premodifying a preposition to the right), and later (in the tree structure module) be filtered into the functionally more correct @>A (intensifier pre-adject).

NP	N noun PROP proper noun (PROP can not normally govern arguments or comparatives, and only very few post-adjuncts) (independent pronouns, PERS and SPEC, substitute for whole NPs, and can't usually have dependents, with the exception of set operators and, sometimes, 'todo':*todo ele*)	PP som @N< (postnominal) *respeito* ***por*** *cumplicidade* ***com*** *afinidade* ***para***	@>N or @N< (pre- or post-nominal), consisting of: AP adjective phrase ***grande*** *sertão* *sugestão* ***íntima*** PP (only post-nominal) ..*a janela* ***da sala*** **..***Pedro* ***da Silva*** DET, usually pre-nominal (also more than one) ***estas três*** *árvores,* ***o*** *João,* ***mais*** *leite,* *proposta* ***sua*** ADV as "set operator" ***só*** *o pai* , ***até*** *o pai ..* *dinheiro* ***demais*** AS-KOMP comparative small clause *um homem* ***como um forte, qual um touro***

Prenominal adjects(@>N) are always modifiers (i.e. not valency governed), and can be filled iteratively with more than one element from the word classes of DET, NUM, ADJ and, rarely, PCP. Though only one syntactic function (@>N) is used by the parser, permutation tests show that different subtypes do exist and can be defined in syntagmatic terms. These subtypes are the basis for the lexical subdivisions of the DET class listed in ch. 2.2.5.2, and they enter the system at the secondary tag level. CG-rules depend very much on word order, and subclasses that can be defined in terms of word order, are therefore useful for disambiguation[140].

A	**B**	**C**	**D**	**ATTR**	**HEAD**	**ATTR**
predeterminers	demonstratives	possessives	quantifiers	ADJ	N	ADJ
<quant1>	<dem><artd>	<poss>	<quant3>	(PCP)	(PROP)	PCP
		differentiators	NUM	<fract>	(ADJ)	PP
		<diff><ident>	ADJ <num>		(PCP)	(DET)

[140] Right of the copula 'são', for instance, a quantifier like 'poucos' cannot be subject complement @<SC, if there is an adjective between the copula and the quantifier. On the other hand, if 'poucos' neighbours 'são' without an interfering adjective, it will be @<SC if followed by an article determiner (DET <art>), but @>N if followed by an adjective:

são perigosos @<SC poucos @>N animais no mundo.
são poucos @<SC os @>N animais perigosos no mundo.
perigosos @SC> são poucos @>N animais no mundo.

AB <quant2><arti>						
todos ambos	estes os	seus meus outros mesmos próprio DET	poucos muitos três primeiros	últimos velhos novos meros meios	livros amigos dentes paulistas conhecidos	paulistas perdidos do Brasil seus todos próprio
uns, uma, cada, nenhum, certo vários DET, diversos DET tantos, quantos, quais cujos						

The first 4 fields, A, B, C and D are determiner prenominals, the fifth is for "attributive" prenominals, the sixth is for the np-head, and the last accommodates postnominal modifiers and arguments. The 4 determiner fields, as a rule, do not allow iteration (i.e. more than 1 element per field), while the attributive fields 5 and 7 do. Note that the determiner subclasses are arrived at by permutation tests only, while traditional pronoun classification is primarily semantic. There is some co-extension of classes, though. Thus, the *demonstrative* class (<dem>) is roughly equivalent to B, and the *possessive* class is equivalent to C. The pronouns 'mesmo' and 'próprio' ('himself') are sometimes semantically classified as demonstratives, as in Almeida (1994), but seem permutationally to belong to class C:

estes mesmos/meus poucos amigos
esses meus/outros poucos amigos
o próprio/meu pai
os próprios/meus sete nanos

Perini (1989, p.153) lumps 'mesmo' with the possessives on these grounds, alongside with 'outro', noting the free order within the class:

esse outro meu amigo/ esse meu outro amigo

There is, however, yet another possibility: 'mesmo', 'próprio' and – in a way – 'outro' can't substitute for neither a demonstrative *or* a possessive:

comprou este/meu/*mesmo/*próprio/?outro carro

The explanation is that 'mesmo' and 'próprio' can't be the *first* element of an np. Rather, they function as a kind of "identity modifiers" that *follow* certain definite determiners (i.e. demonstratives and possessives). When following 'o' or 'este', identity modifiers can be placed left of a possessive, which explains the apparent free order co-occurrence with possessives (class C):

o/este mesmo/outro (meu) livro	(the/this **same/other** book of mine)
meu mesmo/outro livro	(the **same/other one** of my books)
o próprio (meu) pai	(my father **himself**)
Meu próprio pai	(my **own** father)

Note that ‘próprio’, and to a lesser degree, ‘mesmo’ and ‘outro’ (?) undergo a slight change in meaning, depending on whether they follow a demonstrative or a possessive.

The traditional class of definite articles can be defined as the demonstratives ‘o’, ‘a’, ‘os’ and ‘as’, when appearing as class B prenominals (@>N):

> D**os** (<art>) bolos sobram só **os/estes** (<dem>) que fizemos ontem.

The third traditional pronoun class that permits “adjectival” (i.e. @>N) usage, is that of the so-called indefinites. The class contains different types of quantifiers, and is most easily defined *via negationis,* as all *but* demonstratives and possessives. Thus, in my field scheme, indefinites can appear in all places but B and C, including the “joint field” AB. There are three types:

A todos (‘all’), ambos
AB um, nenhum, todo=o (‘all of’, ‘the whole’), todo (every), quantas, quais, tantas, vários ...
D muito, muitos, quatro

The AB field includes ‘um’, ‘uma’ which in the singular traditionally are called indefinite articles. Like for definite articles, the DET word class can be retained, and a distributional definition crafted: Indefinite articles are ‘um’ and ‘uma’ when used in the singular and prenominally (@>N). As a numeral (NUM), of course, ‘um’ belongs in field D. The difference can be tested with a possessive (C) or with ‘cujo’, that “covers” the AB-field, and thus only allows a numeral ‘um’ to its right:

Comeu uma DET sua maçã (não uma pera).	*Cuja uma maçã comeu?
Comeu sua uma NUM maçã (não dois).	Cuja uma maçã comeu?

The six traditional pronoun classes (demonstratives, possessives, indefinites, personal, interrogative and relative pronouns), or seven, if reflexives are regarded as separate from personal pronouns, are not mutually exclusive[141]. Thus, there are relatives/interrogative that at the same time are indefinites. Those of them that allow @>N usage, fall in my DET-AB category, too (quantas, quais ...). The relative/interrogative ‘cujo’ seems to have its own distributional type, DET-ABC, since it doesn’t allow possessives (C) to its right. However, in my view the exclusion of possessives right of ‘cujo’ is a semantic clash (‘cujo’ itself expresses possession, too) rather than a syntactic rule. After all, sentences like the following are not entirely agrammatical:

A empresa, cujos meus ações já vendi há cinco mesos, andava muito mal

‘Outro’, which is often included among the indefinites, has a special distribution, too. As mentioned above, it can – like ‘mesmo’ - follow B- and C-determiners. Where semantically possible, this holds for AB-determiners, too. It is for semantic reasons that ‘outro’, but not ‘mesmo’ can follow ‘um’: both ‘outro’ and ‘um’ are indefinite, but ‘mesmo’ isn’t. Unlike the other two “identity modifiers”, ‘outro’ *can* fill the leftmost slot in an np – but it hereby acquires another meaning:

um outro livro	(‘a different book’)
outro livro	(‘yet another book’)

A few other indefinites (‘vários’, ‘diversos’) have a double distribution as either DET-AB @>N (‘some’, ‘a number of’) or as @N< (‘different’). Since the change in meaning occurs when filling the @N<, which is typical of modifier adjectives (ADJ) rather than determiners (DET), it can also be captured by tagging the words with a different word class.

Certain DET or SPEC quantifiers, as well as some NP’s with a head denoting quantity, can premodify a nominal head mediated by the preposition ’de’:

(a) DET +de +countable:	*algumas/muitas das suas maçãs*
(b) SPEC +de +mass noun	*um=tanto de esmola*
	algo da riqueza do velho
(c) NP +de +mass noun	*uma pinga de esta água*

In (a), there is agreement between the quantifier DET and the modified noun, supporting a premodifier analysis. On the other hand, the DET does not appear in its usual place, since there is an article to its right. The only slot left for a @>N constituent

[141] Not to mention the striking fact that the closed classes of relatives and interrogatives comprise exactly the same words, making these categories purely syntactic-semantic.

’algumas de’ would be that of DETA (<quant1> or predeterminer), which also seems to be fitting for ’um tanto de’, ’algo de’ and ’uma pinga de’ in (b) and (c). However, this analysis is *syntactically* awkward in that it strands a preposition in a constituent of which it is not head, but dependent. In the CG-terminology used here, ’de’ would have to be tagged as adverbial post-adject (@A<) of the quantifying adnominal @>N (constituent bracketing added):

((algumas @>N de @A<) as @>N maçãs @NPHR)
(((uma @>N pinga @>N) de @A<) esta @>N água @NPHR)

Not to speak of the unorthodoxy of this analysis, it is in conflict with ordinary CG-rules trying to establish PP dependencies. A preposition’s ”need” for an argument is very strong in the parser’s rule set, and without major grammar surgery, ’maçãs’ and ’água’ are bound to receive the @P< tag. Therefore, I have opted for an analysis with the quantifier as head postmodified by a PP in which the ”semantic head” *functions* as argument of preposition:

(algumas @NPHR (de @N< (as @>N maçãs @P<)))
(uma @>N píngua @NPHR (de @N< água @P<))

A similar problem arises with adjective modifiers. Again, the parser sticks to surface syntax, tagging the *semantic* modifier as *syntactic* head, and the *semantic* head as argument of preposition:
(O @>N estúpido @NPHR (de @N< rapaz @P<))
rather than:
(O (@>N estúpido @>N de) @A< rapaz @NPHR)

The remaining, non-DET, fields of an np, 5, 6 and 7, could be called for pre-attributive, head and post-attributive, respectively. Though many adjectives and nouns can appear in several of these three slots, a few are restricted to *one* slot and can be used for substitution testing. ‘Mero’ and ‘meio’ define the pre-attributive field (5), ‘seu’, ‘tal’, ‘assim’ and pp-attributes define the post-attributive field (7), and concrete object nouns like ‘árvore’, ‘faca’ or ‘sol’ define the head position (6).

When disambiguating nouns and adjectives, it is important for the CG-rules to be able to “trust” the distributional fact that determiners come in a certain order, and especially that determiners come left of nouns, and adjectives and participles right. Therefore, the few exceptions are marked in the lexicon as <post-det> (possessives, ‘próprio’, ‘todo’) and <pre-attr> -adjectives (‘grande’, ‘novo’, ‘velho’ ...), respectively. The <pre-attr> tag does not mean that @>N is the *preferred* function of these adjectives. Rather, it implies that all *other* adjectives are *not* <pre-attr>, but *post*-nominal (@N<) when modifying a noun. Only very few adjectives (the above mentioned ‘mero’ and ‘meio’) are *obligatorily* pre-attributive in their distribution, and receive a special tag, <ante-attr>. A fourth lexicon-tag that helps place words in the last 3 fields of an np with regard to each other, is the <attr> tag for nouns (‘ateista’, ‘iconoclasta’, ‘apoiador’, ‘caboclo’)[142], which means that the nouns concerned can appear with postnominal function (@N<), i.e. to the right of other nouns.

From the above, CG-rules like the following can be crafted:

* Discard adjective in favour of noun reading, if the word is not <pre-attr> and if the word to the left is a DET
* Select adjective instead of noun reading, if the word is not <attr>, and the word to the left is a noun
* Map a DET word as a prenominal (@>N), if it is surrounded by nouns and is not <post-det>
* Discard @>N function, if the word is ADJ and not <pre-attr>
* Discard @N< function, if the word is N and not <attr>
* Discard np-head function (i.e. @SUBJ, @ACC etc.) in favour of @N<, if the word is N <attr> and the word to the left is N
* Discard @N< function in favour of np-head function even in an N <attr> word, if the word to the right is N <attr>, too
* Select head function over both @>N and @N< for adjectives not <pre-attr> that are flanked by a determiner to the left and another adjective to the right

Postnominal pp-adjects (CG-marked as PRP @N<) can be both modifiers (that permit repetition) and arguments (that don’t):

@N<ARG	**@N<MOD**
a execução do revolucionário	a execução da véspera
o medo da crise	o medo da criança
a confiança no governo	um espião no governo

Typically, valency bearing nouns are deverbals in ‘-ação’, ‘-mento’ or nouns from the semantic field of cognition, and thus, traditionally, the valency link between an N and a pp-argument is tested “etymologically” (cf. Perini, 1989, p. 180): If a corresponding verb-argument relation can be found, the pp is valency-bound, if the pp can be replaced by an etymologically corresponding adjective, it is an adjunct. Sometimes, the correspondence is close (confiança/confiar em, execução/executar), but often it is remote (medo da criança - ?medo infantil, medo - ?temer), nonexistent (da véspera - ?) or plain wrong, as in ‘?um espião governamental’ (‘a government spy’), since ‘governamental’ only corresponds to ‘do governo’, not to ‘no governo’.

[142] The N <attr> class is an open class, comprising especially ‘-ista’, ‘-or’ and profession nouns, or, to put it in other words, “things a human being can be”. Conversely, there are adjectives that are especially likely to fill the head slot of an np, like ‘-ês’ adjectives (‘um francês’, ‘os burgueses’).

Because of these problems, I would like to propose a co-ordination test. Since valency bearing nouns govern a specific preposition, which cannot be exchanged and bears no literal meaning, co-ordination with nouns where the PRP @N< would keep its literal meaning, should not be possible. In other words, a postnominal pp can only be shared by two co-ordinated noun heads if both valency-bind that preposition, or if both don't. In the examples, valency governing nouns are in bold script:

*a **execução** e a idade do revolucionário	(ARG-MOD clash)
*o **medo** e o tamanho da crise	(ARG-MOD clash)
*a **confiança** e a crise no governo	(ARG-MOD clash)
a execução e a festa da véspera	(MOD-MOD co-ordination)
o medo e a febre da criança	(MOD-MOD co-ordination)
um espião e um comunista no governo	(MOD-MOD co-ordination)
a **cativação** e a **execução** do revolucionário	(ARG-ARG co-ordination)

Another argument/modifier test for postnominal pp adjects can be based on whether the introducing preposition can be replaced by another "literal" preposition, like substituting 'sem' for 'com', or one place preposition for another. In modifier pp's, such substitution is usually *syntactically* possible (barring some semantic oddities), while in argument pp's, any preposition has to obey selection restrictions dictated by the preceding np-head.

On a corpus basis, ambiguous cases like 'o medo da criança', where a noun allows the same pp both as argument and modifier, seem to be rare, and given enough valency and semantic information from the lexicon, Constraint Grammar could probably be made to handle the distinction in most cases. As a first step, CG rules would remove argument adject readings if the head noun bears no valency tag (<+PRP>) for the preposition concerned. As a second step, rules should check whether the preposition's argument (@P<) semantically matches the preposition's literal meaning. For instance, place nouns as @P< increase the chances of 'em'-headed adject pp's being modifiers, while +ANIM nouns as @P< point towards argument-status for 'em'-headed adject pp's.

However, since the parser for the time being (with its present tag set) does not distinguish between argument and modifier @N<, nominal valency is used to distinguish between pp *adjects* (@N<, @A<) and pp *adverbials* (@ADVL) instead, a process which is described in the chapter on adverbial function.

4.2.2 The adpositional group

	head	*argument adjects*	*modifier adjects*
AP	ADJ adjective PCP participle (participles are *attributive* participles, which - unlike verbal participles after *ter/haver* - have gender and number inflexion) ADV adverb DET determiner	PP as @A< (post-adject) *cheio* ***de*** *parecido* ***com*** *antes* ***da ceia*** *depois* ***da festa*** *relativamenta* ***à lei*** NP as @A< *inclusive* ***os alunos*** FS-KOMP comparative subclauses as @A< *mais velho* ***que*** AS-KOMP comparative small clauses as @ A< *tão velho* ***como*** *tanto dinheiro* ***quanto***	ADV as @>A (mostly pre-adject intensifiers, but also @A<) ***muito*** *rico* *agressivo* ***demais*** *fala* ***muito*** *depressa* *fala depressa* ***demais*** ***muito*** *poucos* SPEC or "NP" as @>A (rare) ***um=tanto*** *devagar* ***nada*** *religioso* AS-KOMP comparative clauses as @A< *forte* ***como*** *um urso*

Structurally, the traditional form categories of adjective group (adjp), adverb group (advp) and what I would call determiner group (detp), have much in common. They all allow a probably closed class of intensifier pre-adjects, and prepositional groups as modifier or argument post-adjects. Pre-adject intensifiers co-vary with one post-adject intensifier, 'demais'.

@>A	@HEAD	@A<
muito/pouco bem/mal completamente/nada algo/um tanto incrivelmente	velho ADJ visível ADJ contente ADJ derrubado PCP socialista N <attr> iconoclasta N <attr> bem ADV depressa ADV poucos DET	de corpo e ânimo na cidade com a situação na época pelos comunistas de vocação como mais ninguém demais
mais/menos tão		do que eu como aqui

Ignoring semantic incompatibilities (*pouco poucos, *menos poucos), the intensifier pre-adject function is so basic to all the above groups, that it can be used to *define* a common umbrella-group for all of them on formal grounds. Better still, certain intensifier adject words *(nada, algo)*[143] never appear as prenominals (@>N), so the criterion cannot only be used to define adjp's, advp's and detp's as *one* group category, but also to distinguish it from the np category. I will call the new group type for ***adpositional group (ap)***, since its prototypical functional distribution covers "adposed" elements, adjects, adjuncts and the attributive function of predicative. Prototypical heads and substitution word classes are **ad**-words, i.e. **ad**jectives and **ad**verbs, which also explains the CG-icons for the groups adject dependents, @>A and @A<. In order to distinguish between adjects in np's and ap's, I will use the terms adnominal adjects and adverbial adjects, respectively.

The three fields in an ap are fairly easy to define in a formal way, since the @>A field allows only adverbs and certain quantifier pronouns or "pronominal np's" ('algo', 'um tanto'), and the @A< field allows only prepositional groups and – in the special comparative structures – comparandum ACL's and FS's. The head field can be defined negatively, as the position that has or allows adject intensifiers *(nada, algo)* to its left.

Note that it is the co-occurrence or grammaticality of certain types of pre-adjects that defines np's and ap's, *not* the heads word class. Thus, articles remain @>N even with adjective or pronoun heads, building np's, and intensifier adjects remain @>A even with noun heads, building ap's. In the examples, np-heads are in bold face, ap-heads are in italics.

(o	**azul**	(tão	*claro*)	do céu)
@>N	ADJ	@>A	@N<	@N<

(o	**bem**)	e	(o	**mal**)
@>N	ADV	@CO	@>N	ADV

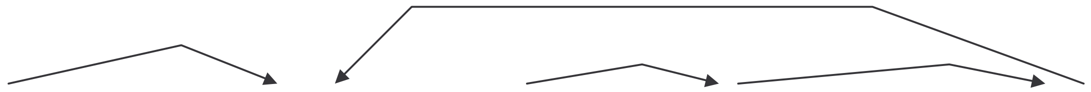
(o	**pouco**	(que sobra))
@>N	DET	@N<

(um	**comunista**	(muito	*ateista*))
@>N	N <attr>	@>A	N <attr> @N<

In the flat dependency notation used here, a chain of prenominals (@>N) is regarded as sisters pointing to an np-head to the right of the whole chain, rather than to each other.

[143] But not *muito* or *pouco*, which both can appear preonominally.

However, adnominals *can* point to a determiner head, if the latter is part of an adverbial adject (@>A) and does not function as adnominal itself:

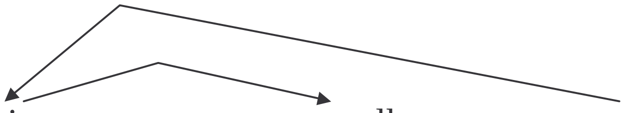


(um	**professor**	((um	**tanto**)	*iconoclasta*)))
DET	N	DET	DET <quant>	N <attr>
@>N	@NPHR	@>N	@>A	@N<

The distinction between modifier and (valency bound) argument adjects in ap's can be made visible by an isolation test. Modifier adjects can, argument adject cannot be isolated when substituting the ap by an interrogative dummy, 'o que' or 'como':

o que era de corpo e ânimo?	velho
o que era de vocação?	socialista
o que era como mais ninguém?	iconoclasta
?o que o pai era da guerra?	receoso
*o que a região era em ouro?	rico
*o que foi do que a última vez?	pior
o que ela era demais?	tímida
o que era mais do que eu?	velho

Note that the last example tests for modifier status of 'mais do que eu' with regard to the ap-head 'velho', *not* against the argument status of the KOMP< constituent ('do que eu'), which is an argument of 'mais', not 'velho', as can be seen from the fact that 'velho do que eu' *without* 'mais' is agrammatical. What makes the case difficult is the fact that 'mais do que eu' is a *disjunct pre*-adject modifier of 'velho'. The @KOMP< constituent itself, then, is an adverbial post-adject, of 'mais', *inside* the larger pre-adject:



mais	velho	do que eu
ADV<komp> @>A	ADJ @HEAD	@KOMP<

4.2.3 The prepositional group (PP)

	head	*argument adjects*	*modifier adjects*
PP	PRP preposition	@P< N or NP nominal phrase *sem* ***o amigo*** ADV adverb *para* ***lá****, até* ***hoje*** FS finite subclause *depois* ***que*** *...* ICL non-finite clause *para* ***ajudar a velha***	@>P intensifiers: ***muito*** *sem graça* operators: ***até*** *no Brasil*

The third kind of group advocated here is the prepositional group (pp). A pp is not hypotactic like np's and ap's, but katatactic, a fact which makes it more difficult to decide on which constituent to count as head of the group. However, valency-wise it is the preposition that links the group to a head on the next syntactic level. Thus, it is a specific preposition that is governed and ”asked for” when a verb, noun or adjective allows pp-arguments. One could say that - though being able to replace the whole group - it is the preposition that is outward ambassador of the group. Therefore, in dependency grammar, the preposition counts as head of the pp, with the rest of the pp rolling as the preposition's [dependent] argument (@P<).

The argument slot of a pp group can be filled by almost any type of word class, group or clause, though most typically so by np's and those word classes that easily qualify as np-heads, including infinitives and infinitive clauses[144].

(a) passeava com **a mãe** (NP)
(b) discutiram sobre **você** (SPEC)
(c) gostava de **ler na cama** (ICL)
(d) andava com medo de **magoá-la** (ICL)
(e) tem chovido até **hoje** (ADV)
(f) os amigos tinham se casado sem **que o soubesse** (FCL)

PP's in general do not allow ordinary modifiers like NP's and AP's, but only a kind of ”set operators” that can precede most groups (g), and – in a few cases – premodifying intensifiers (i) or “time operators” (h). In all three cases one could argue that what is modified is not the preposition head, but rather the PP as a whole:

[144] In Portuguese, infinitives and infinitive clauses even allow preposing a definite article, like ordinary nominal material: ***o começarmos cedo*** *vai ajudar muito.*

(g) isso existe (até/nem (nos Estados Unidos))
(h) lutava com o inimigo, (ainda/já (com a energia da raiva))
(i) se retirou (muito (sem querer))

Such an analysis would make the resulting, larger, group hypotactic, and it would thus no longer qualify as a PP. Since intensifier adjects project AP-hood (according to the definition used in the last chapter), the group in (i) could be called an AP with adverbial function, with a complex (PP-) head and an adverbial pre-adject (@>A) as modifier. Given the predicative function of the group in (h), a similar solution might work here. Both the enlarged groups in (h) and (i) can be fronted and focused as whole constituents:

(h’) Era ainda com a energia da raiva, que lutava com o inimigo
(i’) Foi muito sem querer que se retirou

‘Até’ and ‘nem’ in (g), however, are different. They can operate on constituents outside the “adpositional” range, too, like subjects and objects (‘confiava <u>até nos Estados Unidos</u>’), and appear to be oddly “transparent” with respect to their supposed PP head. Thus, PP’s cannot be focused *together* with modifiers like ‘até’ or ‘nem’:

(g’) ? é até/nem nos Estados Unidos que isso existe.

One possible explanation for the agrammaticality of (g’) is that ‘até/nem’ is a clause level constituent (@ADVL), and doesn’t attach to the PP at all. In this case, however, there should be no difference in meaning whatever the adverbial precedes the verb or the PP:

(j) ele escreve livros **até** em francês. (rather than in English)
(k) ele **até** escreve livros em francês. (rather than just *speaking* French)

Another solution is to assign to “operator adverbs” the function of focus markers, which accounts both for why they have to immediately precede their head and why they can’t be moved along into the focus bracket of ‘é/era/foi .... que’. And since focusing is neutral with respect to focused form, a focused PP would still be a kind of “meta-PP”.

For reasons of notational clarity, my parser needs to dependency-attach all of the above PP-modifiers to a *word*, i.e. the preposition head (@>P), rather than the whole PP. However, if we assume that *all* pre-adjects in PP’s behave in the same way, dependency-wise (i.e. modify the whole PP), then there is nothing in the way of *interpreting* the @>P tag differently from ordinary word-to-word dependency tags, or filtering all intensifier and time operator @>P tags into @>A tags.

To make things even more complicated, focus markers can be attached not only to the groups types sketched above, but also to complementisers (or, arguably, whole clauses), as the following examples indicate:

(a) Veio **só @>A quando** nada sobrava do jantar.
(b) Convidou **só @>N quem** quisesse ajudar.
(c) Pagou **só @>S porque** temia um proceso civil.

In the case of “quando” (a) and “quem” (b), existing group types can be used, ap and np, respectively, with the focus marker tagged as @>A or @>N. However, to cover the rare case of a focused conjunction (c), I was forced to introduce a new attachment tag, @>S (subordinator modifier), arguably creating a distinct group type too - that of subordinator group. If proven useful, the concepts of @>S and subordinator group could eventually be enlarged to cover cases like (a) and (b), too.

4.3 The verb chain

4.3.1 The predicator: Constituent group or clause hierarchy?

@FAUX	finite auxiliary (cp. @#ICL-AUX<)
@FMV	finite main verb
@IAUX	non-finite auxiliary (cp. @#ICL-AUX<)
@IMV	non-finite main verb
@PRT-AUX<	verb chain particle (preposition or "que" after auxiliary)
@#ICL-AUX<	argument verb in verb chain, refers to preceding auxiliary

(the verb chain sequence @FAUX - @#ICL-AUX< is used, where both verbs have the same subject , @FMV - @#ICL-<ACC is used where the subjects are different)

VERB CHAINS (MATRIX VERB STRUCTURES)

From the point of view of dependency grammar, a maximal VP in Portuguese subsumes all adjuncts and arguments of a verb, including the subject, turning such a VP into a form category very close to that of *clause*. But what does count as head of this maximal VP if there is more than one verb in the clause. Do auxiliaries count as dependents of the main verb, and - if so - are they dependents at the same constituent level as subject and objects, or at a lower level? Or is it always the first (finite) verb that functions as head, - as, for instance, number agreement with the subject and selection restrictions on the form of the following verbs suggest.

Depending on the notational convention, I do not believe that the above alternatives have to be contradictory: Both main verb and first auxiliary are heads, only on different levels, and both contribute features to a complex constituent which heads the clause.

In my Portuguese CG, the surface-syntactic solution is that the first verb in a verb chain valency-binds the second one (in the same way a constituent clause is governed), the second verb governs an eventual third, and so on, suggesting the outermost auxiliary as the head of the whole structure and the trailing (non-finite) main verb as head of the rest of the VP, creating a multi-layered clause-hierarchy where every additional auxiliary wraps a new onion-layer around the VP-main-verb kernel:

(1)	*ele*	***continua***	*querendo*	*ser*	*eleito*	*presidente*
	@SUBJ	***@FAUX***	*@#ICL-AUX<*			
			@IAUX	*@#ICL-AUX<*		
				@IAUX	*@#ICL-AUX<*	
					@IMV	*@SC*

While this is an intuitive way to handle verb chains in flat dependency grammar[145], it does not do justice to the fact that the relation between the members of the verb chain is another than the one between the verbal head and its complements in one-verb sentences. The verb chain itself is less hierarchical and more "holistic" in its feature sharing than a clause. I would therefore like to argue that the head of the VP is a complex unit in its own right, a group-like structure which I will call VC (verb chain), or - functionally - the predicator[146]. This creates a distinction between the (higher) "clausal" VP-level (maximal VP) and the (lower) "phrasal" VC-level (minimal VP). While dependency links within the VC are preserved, it will then be the VC as a whole that arguments like direct object and subject attach to. One advantage of this concept is, that features like number and person are shared by the whole VC and not only attributed to the finite verb, and that complex features like the *ter* PC and MQP tenses or even aspect have a place to be, and need not be arbitrarily attached to a single word. This way counterintuitive dependency discrepancies can be avoided, like attaching the subject to the finite auxiliary (for agreement reasons), but the ACC, DAT and PIV objects to the non-finite main verb (for valency reasons).

The following are examples of the functional uses of verb chains in Portuguese, with the complex VC feature given in square brackets:

(2)

	@AUX	@PRT-AUX<	@IMV	
*** complex tenses**				
	ter/haver PR		+ *PCP*	[perfeito composto]
	ter/haver IMPF		+ *PCP*	[mais-que-perfeito composto]
	ter/haver COND		+ *PCP*	[condicional II]
	ter/have FUT		+ *PCP*	[futuro II]
	ir	+ *a*	+ *INF*	[near future, 'to be going to']
	vir	+ *de*	+ *INF*	[recent past]
*** passive voice**				
	ser		+ *PCP*	["action passive"]
	estar		+ *PCP*	["state passive"]

[145] The solution originally proposed in (Karlsson, 1995), evades part of the problem by not using dependency markers - the members of the verb chain are juxtaposed without suggesting a hierarchy, making the notation compatible with both a reading that sees auxiliaries as dependents of the main verb, and one that attaches non-finite main verbs to auxiliaries and auxiliaries to preceding auxiliaries.

[146] The *predicator* unit is recognizes in many German grammars as *Prädikat*, whereas English (generative) grammars often define *predicate* as a VP consisting of the main verb and its dependents, minus the subject, leaving auxiliaries to form their own group as a constituent of the clause.

*** aspect**

estar/andar/continuar/seguir		+ *GER*	[durative]
estar/andar/continuar/seguir + a		+ *INF*	[durative]
começar	+ *a*	+ *INF*	[inceptive]
acabar	+ *de*	+ *INF*	[conclusive]
deixar	+ *de*	+ *INF*	[cessative]

*** modals**

ter	+ *de/que*	+ *INF*	[obligation]
dever		+ *INF*	[probability, obligation]
poder		+ *INF*	[posssibility]
saber		+ *INF*	[capacity]
querer		+ *INF*	[optative]

In all of the above cases both verbs in the VC have the same subject, and can be analysed in the same way as in (1). In the matrix verb structures in (3), however, there are two (different) subjects, which is why I prefer to read the first verb as a main verb, and the second as the head of a non-finite subclause functioning as clausal object of the first.

*** perception verbs and "ACI"[147]**

(3a) *Do quarto, ouvi* *os outros sair da casa*.
@ADVL> @FMV @#ICL-<ACC
@SUBJ> @IMV @ADVL

*** causatives**

(3b) *O rei mandou* *o delegado chamar os assaltantes*.
@SUBJ> @FMV @#ICL-<ACC
@SUBJ> @IMV @<ACC

This object-ivity can even be expressed morphologically: when the second subject is pronominalised, it takes the accusative case and is hyphen-attached to the matrix verb (3c).

(3c) *O rei mandou-* *o chamar os assaltantes*.
@SUBJ> @FMV @#ICL-<ACC
@SUBJ> @IMV @<ACC

[147] Latin for "Accusative with infinitive"

One might be tempted to conclude from morphology to syntax and assign a direct object (@<ACC) reading to the pronoun in (3c), instead of the @SUBJ> reading, making the reduced - subjectless - infinitive clause either an adverbial (@<ADVL) or an object complement (@<OC), cf. (3d)[148].

(3d)	*O rei mandou-*	*o*	*chamar*	*os assaltantes*.
	@SUBJ> @FMV		@#ICL-<ADVL/<OC*	
		@<ACC*	@IMV	@<ACC

However, like German, but unlike English, Portuguese can - in causative structures - omit the subclause subject altogether (3e)[149]. Since subject omission is normal in Portuguese, but object omission is not, this is an argument in favour of the "subject in subclause" reading for the "accusative" (pro)noun in concatenations of type (3c).

(3e)	*O rei mandou*	*chamar*	*os assaltantes*.
	@SUBJ> @FMV	@#ICL-<ACC	
		@IMV	@<ACC

Without verbal valency information, this sentence is, of course, ambiguous: With an intransitive/ergative verb in the non-finite position, the trailing NP becomes a ("leftward") subject.

(3f)	*O rei mandou*	*entrar*	*os assaltantes*.
	@SUBJ> @FMV	@#ICL-<ACC	
		@IMV	@<SUBJ

Since the object-pronoun subject (3c) of the non-finite clause can be fronted in both ACIs and causatives, two other types of ambiguity can be created - the first syntactic and the second notational.

(3e)	*O rei*	*o*	*mandou*	*chamar*.
	@SUBJ>		@FMV	@#ICL-<ACC
		@ACC>>		@IMV
		@SUBJ>>		

[148] These alternative @#ICL-OC readings are especially tempting, when the accusative is located *before* the matrix verb: *O rei **o** mandou chamar os assaltantes*, since two adjacent @SUBJ> tags (for both *rei* and *o*) appear somewhat awkward. In the case of perception verb ACI's (3a), the accusative pronoun or noun phrase can fill the matrix verb's transitive valency slot on its own, providing a further argument in favour of the V - ACC - OC reading: 'Ouvi os outros sair da casa.' - 'Ouvi os'. This is not true of of causatives - in the very least, there is a meaning change in the matrix verb: 'fez a filha obedecer.' - *'fez a filha.' For further examples, cp. chapter 4.4.2 and the manual "Portuguese Syntax" (Bick, 1999).

[149] If the subclause verb has both transitive and intransitive valency, subject omission as in (2b2) opens for a new ambiguity in the subclause *object* (here: *os assaltantes*): in theory, it can now also be read as left-attaching subject.

Ordinarily, *'o'* is understood as *subject*[150] of the non-finite clause in sentences of type (3e). In literary language, however, *'o'* might also be *object* of the non-finite clause, this being one of the dubious cases in the application of the clitic fronting test.

The other ambiguity concerns notation. How is it possible to know where the dependency marker of *'o'* attaches - at the first, finite, or second, non-finite, main verb? Since the default definition is attachment to the *nearest main verb* (which is *not* the correct choice in this case), I have opted for a special notation in similar sentences: A double dependency marker (>>) refers to the *second* main verb to the right.

With a reflexive object pronoun in the same construction, the subject/object ambiguity can be resolved by means of the valency class of the second verb: Intransitive/ergative verbs (3f) favour a subject-reading, transitive (3g) or transobjective (3h) ones favour the object-reading.

(3f)	*O rei*	*se*	*deixou*	***cair***	*na cama*.
	@SUBJ>		@FMV	@#ICL-<ACC	
		@SUBJ>>		@IMV	@<ADV

(3g)	*O rei*	*se*	*deixou*	***levar***.	
	@SUBJ>		@FMV	@#ICL-<ACC	
		@ACC>>		@IMV	

(3g)	*O rei*	*se*	*fez*	***eleger***	*presidente*.
	@SUBJ>		@FMV	@#ICL-<ACC	
		@ACC>>		@IMV	@<OC

4.3.2 Verbal matrices: How to define an auxiliary?

Concatenating verbs: Subclass criteria

Portuguese auxiliaries are much harder to define in a consistent way than their English cousins, and Portuguese grammars are unclear and diverging on the matter. One minimalist position would acknowledge only *ter/haver* and *ser/estar* with participle or gerund main verbs. This, however, would include Brazilian duratives (which are constructed with *estar + GER*), and exclude European Portuguese duratives constructed with *estar + a + INF*. Furthermore, analogous forms exist for *andar + GER/a*,

[150] This is also the reading my parser is set to prefer.

continuar + GER/a and so forth. Also, participles after *ser* and *estar* behave structurally the same, but *estar* alternates with *ficar* (diluting its auxiliary-status), and its participles alternate with adjectives, thus being syntactically equivalent to subject predicatives. After removing *estar* from the auxiliary list, only a kernel of *ter/haver* with tense readings, and *ser* with passive voice readings would be left in the auxiliary camp.

A more liberal view would allow modals and aktionsart markers, maybe also ACI-constructions and causatives. These cover all kinds of direct and preposition mediated infinitive chains. How to draw a formal line? Inspired by traits of the core auxiliaries, a number of tests is proposed in the literature:

(a) leftmost position in a verb chain
(b) transclausal subject identity
(c) no selection restrictions for the subject
(d) no imperative
(e) no (semantic) selection restrictions on the number 2 verb
(f) allows object pronoun fronting (clitic fronting)
(g) exclusion of interfering "não"
(h) finite subclause substitution test
(i) passivisation test for clause coherence

While (a) obviously delineates the pool of verbs from which to choose auxiliaries, it doesn't *define* them. (c), (d) and (e) are really about auxiliaries not having semantic lexical content, a criterion that would exclude all but the tense and voice auxiliaries. Some modals, for instance, violate (c), since they select +HUM in the subject (*dever*, *saber*), the imperative criterion (d) asks for +CONTROL in the subject and splits the otherwise coherent group of perception verbs (-CONTROL) and causative (+CONTROL) in two. Similarly, some causatives (*mandar*), like some cognitives (*prometer*), but again unlike perception verbs, violate (e) by selecting for +CONTROL in the *second* verb's subject. Tests (f) and (g) are about "transparency": real auxiliaries are expected to attach to their main verbs in an unseparable way. The stricter of the two is the negation test (g), with only *ser*, *ter* and *ir* (!) passing, while the clitic fronting test[151] (f) works well and coherently for most auxiliary-candidates that directly "govern" non-finite verb forms. For these verbs the subject identity test (b), comparing the main clause subject to the (often unexpressed) subject of the non-finite clause, yields very similar results. The fact that two different tests, one morpho-syntactic, the other semantico-syntactic, agree on the same list of words, strengthens both tests' legitimacy. ACI-structures and causatives are excluded by both tests (the clitic to be fronted is the

[151] non-nominal pronominal material is moved from a position between matrix-verb and non-finite verb to a fronted position immediately to the left of the matrix verb.

second verb's object[152]), with some exceptions and ambiguities in the causative group (treated below).

Sadly, the clitic fronting test is negative or unclear for some preposition mediated infinitive constructions expressing aktionsart and modal functions by same-subject verb chains:

(1) Object pronoun fronting in preposition mediated auxiliaries (AUX+PRP+INF)

object pronoun fronting test **mediating preposition/particle**	**positive** *'já **o** acabou de fazer.'*	**negative or dubious** *?'nunca **o** negou de fazer.'*	**negative** **'sempre **o** sonhava com fazer.'*
"a" (<a^xp>)	aprender, aspirar, começar, continuar, desandar, desatar, entrar, estar, ficar, passar, propor, tornar voltar	botar, chegar, dar, deitar, destampar, falhar, faltar, ir, vir *(maybe only with [non-female] pronouns not starting in 'a' ?)*	
"de" (<de^xp>)	acabar, deixar, desistir, evitar, falhar, faltar, haver, intentar, largar, necessitar, parar, planejar, precisar, projetar, prometer, ter, tratar, vir	assentar, escusar, folgar, negar, pegar	começar, continuar, cuidar, determinar, dever, entrar, ficar
"em" (<em^xp>)	aceder, pensar	aspirar	coincidir, confiar, contar, cuidar, desandar, destampar, entender, espraiar, estar, sonhar, timbrar, vacilar
"com" (<com^xp>)			sonhar
"para" (<para^xp>)	estar		
"por" (<por^xp>)	começar, estar		acabar, anelar, ansiar, trabalhar
"que" (<v+que>)	ter		

The subject identity test, while excluding ACI-constructions and causatives, includes *all* modals and aktionsart markers for semantic reasons.

Test (h) tests for "clausality" and against "auxiliarity" by trying to replace the non-finite structure by a finite *que*-subclause. The test confirms the above ACI- and causative groupings, but is somewhat stricter than the subject identity test in other areas,

[152] The object of the matrix verb is, of course, the subject of the second, non-finite verb, and *can*, if pronominal, always be fronted.

removing e.g. *querer* from the modal list, as well as *precisar* and some others from the list of preposition mediated auxiliaries.

Another clausality test is the passivisation test (i), as proposed in Perini (1989) for detecting an interfering clause boundary in the verb chain. Again, *ir* passes the test (1a), *querer* fails it (2a):

(1) Pedro vai comer o frango. (1a) O frango vai ser comido por Pedro.
(1b) *Comer o frango é ido por Pedro.
(2) Pedro quer comer o frango. (2a) *O frango quer ser comido por Pedro.
(2b) Comer o frango é querido por Pedro.

The passivisation test is also a transparency test like (f) and (g): For *ir*, the verb chain is transparent, suggesting auxiliarity, and ’comer o frango’ cannot be isolated as @ACC and made the subject of a corresponding passive clause (1b). With *querer*, the verb chain is not transparent, and ’comer o frango’ *can* be made subject of passive (2b). The passivisation test subsumes a number of other tests:

- it tests for patient case role (PAT) in the subject, since this would disallow another (object) PAT in the *same* clause, and contradict a ’por X’ agent of passive constituent in the passivised clause.
- it implies lack of selection restrictions (test c), since in the passivised clause the same verb has to ”tolerate” a different subject. Many concatenating verbs are cognitive verbs (admitir, adorar, decidir, negar) select for +HUM subjects creating a passivisation conflict with –HUM objects.
- it implies lack of imperative (test d), since PAT subjects imply lack of the control (CONTR) feature.

Between the extremes of accepting *all* concatenating verbs as auxiliaries (a) or restricting the category to *ser*, *ter* and *ir* (g), we have now 2 sets of tests that come up with 2 more or less coherent lists of auxiliary candidates:

1. **subject identity test**, backed by the pronoun fronting criterion, the two of which yield the same results for chains without prepositions (GER, PCP, INF), but differ somewhat in the case of preposition mediation, where the pronoun fronting criterion is ”soft”.

2. **passivisation test**, backed by the finite subclause substitution test and +PAT, –CONTR and lack of selection restrictions for the subject.

As can be seen from the overview of concatenating verbs in the parser’s lexicon (end of chapter), the auxiliary set 2 is a subset of auxiliary set 1. The reason is, of course, that different subjects of matrix and subclause imply two clauses and make one-clause-passivisation impossible:

O rei mandou matar a ovelha.
*A ovelha mandou ser morta pelo rei.

For the same reason, both test-sets sharply exclude ACI- and causative constructions, and the verbs concerned will here be kept outside the auxiliary camp.

A problem with the subject identity test is that some of the preposition-mediated and a few other auxiliary candidates have a double status - they can sometimes appear as full verbs governing preposition phrases with ICL-arguments substituting for NP-arguments. In this case, the subject of the infinitive clause must be expressed, it must be in the nominative if pronominal, and there would be inflexion agreement between the infinitive (thus personal) and its subject:

Gostaria de eles me visitarem
(as opposed to the auxiliary reading in *eles gostam de viajar*)
Temo de (eles) cairem
(as opposed to the ”auxiliary” reading in *temo de cair)*

I tend to think that the reason for, for instance, *gostar* and *supor* allowing a different subject in its ICL complement, and *tencionar* not allowing it, is semantico-lexical rather than proof of these verbs’ membership in to different *syntactic* classes. Opting for the passivisation criterion for auxiliarity, we could hold that subject identity across matrix and subclause is just one of three possible semantic permutations for the subjects of ICL-complements (same subject, different subject or either), and that the criterion of same subject is a necessary but not a sufficient condition for auxiliarity.

A problem with the passivisation test is that some candidate verbs (*começar a/de, continuar a/de, deixar de, parar de*) can appear both with a PAT/-CONTR subject (1c, 2c) or a AG/+CONTR subject (1a, 2a), yielding two different meanings and conflicting results with the passivisation (1c, 2c) and imperative tests (1d, 2d). And while their complements fail the finite subclause substitution test, they *do* take direct np-objects without apparent change in meaning, unlike all other verbs in the set: *“começou a aula”*, or*“parou o cavalo”*.

(1a) A empresa continua a produzir o antigo modelo. ‘.. continues to produce ...’
(1b) O antigo modelo continua a ser produzido. ‘.. continues to be produced’
(1c) A inflação continua a crescer. ‘.. keeps increasing.’
(1d) Continue a produzir o antigo modelo! ‘go on producing ..!’
(2a) Pára de molestar a irmã. ‘(he) stops molesting ...’
(2b) ?A irmã pára de ser molestada. ‘.. stops being molested.’

(2c) Parou de chover. ‘(it) stopped raining.’
(2d) Pare de molestar a irmã! ‘stop molesting ..!’

The meaning shift from AG/+CONTR to PAT/-CONTR ”hurts” more in (2b) than in (1b), but the fact, that – formally – all 4 verbs pass the passivisation test and fail the subclause substitution test, seems enough to include them in the auxiliary set.

In Portuguese, there is a large group of reflexive matrix-verbs (*acostumar-se a*, *lembrar-se de*, *negar-se a/de*) most of which pass the same-subject test, making them auxiliary candidates. However, all of them fail the passivisation test due to the syntactic object status of ”se”. Since most don’t pass the pronoun fronting, imperative and selection restriction tests, either, they will here be excluded from the auxiliary set.

In all, the set then comprises of 22 verbs that all express either tense, voice, modality or aktionsart (cp. bold faced verbs in list at end of chapter).

Those concatenating verbs, that according to the above criteria do not qualify as auxiliaries, but – unlike full verbs governing ICLs – do permit non-nominative pronouns as subject of the ICL (*and* can front those object pronouns), can be grouped as transobjective constructions, and here, some intuitive semantic subclasses can be distinguished also in more formal ways. The ACI-verbs *ver, ouvir, sentir* (3a) demand accusative case, permit infinitive inflexion with NP subjects, and can govern ÷CONTROL verbs (”processes”), whereas causatives permit (3a1) or demand (3a2) dative, or a mediating preposition before the infinitive (3c), and govern mostly +CONTROL verbs (”actions or activities”).

The different classes of auxiliaries and other concatenating verbs are shown in table (2) below. Classification criteria are (a) whether the subject of the infinitive clause is the same as for the matrix verb, (b) what kind of verb-complements are allowed in between the verb of the matrix clause (MC) and the verb of the infinitive clause (ICL), (c) the (simple or double) case function of such interposed complements, (d) whether or not fronting of (ICL-) object pronouns to the left of the matrix verb is possible, and (e) whether or not the infinitive in the ICL is person-inflected.

(2) Typology of Portuguese auxiliaries and other concatenating verbs

@SUBJ of ICL	**same as for matrix verb**	**other than for matrix verb**	
complements allowed in between MC-verb and ICL	**pronouns**	**pronouns**	**NP**

case function		**ACC/DAT** @OBJ of ICL	**ACC** @OBJ of MC & @SUBJ of ICL	**DAT** @OBJ of MC & @SUBJ of ICL	**NOM** @SUBJ of ICL	**none** @OBJ of MC & @SUBJ of ICL
± fronting	÷ INF-inflexion	<x+PCP> ***1*** <x+GER> ***1*** <x> ***2a*** <a/de^xp> ***2b1***	<xt-ACI> ***3a*** <xt/xd> ***3b1*** <xtp> ***3c1***	<xt/xd> ***3b1*** <xd> ***3b2*** <xdp> ***3c2***		
	± INF-inflexion					<xt-ACI> ***3a*** <xtp> ***3c1*** <xdp> ***3c2***
÷ fronting	÷ INF-inflexion	<em/por^xp> <xrp> ***2b2***				<xd> ***3b2*** <xt/xd> ***3b1***
	+ INF-inflexion				+ICL ***4***	+ICL ***4***

1. Functional auxiliaries for tense, voice or aktionsart demanding +PCP or +GER

<x+PCP>	ter/haver +PCP (PC-tense)
	ser +PCP (passive voice)
<x+GER>	estar/andar/continuar/seguir etc. +GER (durative aktionsart)

2. Other "auxiliaries" with anaphoric subject in the ICL (÷inflexion)

a) modal auxiliaries and others, +INF: <x> dever, poder, querer, saber

b) preposition mediated auxiliaries, +PRP+INF

b1) tight concatenations (+ fronting of object pronouns)

<xp> +a/de + INF: acabar de, gostar de, prometer de, etc.

b2) loose concatenations (÷ fronting of object pronouns)

<xp> +em/por + INF: acabar por (?), anelar por, ansiar por

<xrp> +PRP +INF: lembrar-se de, recusar-se a, esforçar-se por, comprazer-se em

3. Transobjective constructions with different subject in the ICL

a) ACI-constructions, +ACC+INF (ICL ±CONTR, ±inflexion after NP-ACC)

<xt-ACI> ver, ouvir, sentir

b) Causative constructions (÷inflexion)

b1) with accustive/dative-choice: +ACC/DAT+INF

<xt><xd> mandar, deixar, fazer

b2) only dative (or +PRP-A +PIV/NP)

<xd> aconselhar, permitir, possibilitar

c) preposition mediatied transobjective constructions, +ACC+PRP+INF

c1) with accusative

<xtp> acostumar alg. a, estimular alg. a, lembrar alg. de, ...
convencer alg. a, decidir alg. a, pôr alg. a ...

c2) with dative (or +PRP-A) +PIV/NP

<xdp> dizer-lhe de/para, proibir-lhe a/de, permitir-lhe a

4. Accusative ICL's after full verbs, with (mostly?) different subject, +NOM+INF

<+ICL> dizer, possibilitar, julgar, supor, detestar, ...

In an actual CG-parse, verb chain hierarchies are chains of (governing) auxiliaries (@AUX) and (governed) non-finite auxiliary complement clauses (@#ICL-AUX<). In the examples, sublcauses (including auxiliary complement clauses) are tab-indented, and group level constituents are space-indented:

```
O             [o] <*> <art> DET M S                        @>N
instituto     ”instituto” N M S                           @SUBJ>
de            ”de” <sam-> PRP                              @N<
o             <-sam> <art> DET M S                             @>N
impeachment   ”impeachment” N M S                            @P<
nunca         ”nunca” <dei> ADV                           @ADVL>
havia         ”haver” <x+PCP> V IMPF 1/3S IND VFIN  @FAUX
sido          ”ser” <x+PCP> <ADJ> V PCP M S   @#ICL-AUX<   @IAUX
testado       ”testar” <de^vp> <ADJ> V PCP M S                   @#ICL-AUX<  @IMV
$.
```

(The institution of impeachment had never been tested.)

```
Deixou        ”deixar” <*> <de^xp> V PS 3S IND VFIN  @FAUX
de            ”de” PRP                               @PRT-AUX<
ser           ”ser” <vK> V INF 0/1/3S                @#ICL-AUX<  @IMV
uma           ”um” <quant2> <arti> DET F S                   @>N
tendência     ”tendência” <+para> N F S                     @<SC
congressual   ”congresso” <DERS –al [ATTR]> ADJ M/F S        @N<
para          ”para” PRP                             @<ADVL @N<
se            ”se” <refl> PERS M/F 3S/P ACC                  @ACC>
tornar        ”tornar” <vrK> V INF 0/1/3S              @#ICL-P<  @IMV
um            ”um” <quant2> <arti> DET M S                     @>N
grande        ”grande” ADJ M/F S                               @>N
partido       ”partido” N M S                                 @<OC
$\,
...
```

([It] stopped being a tendency in congress in order to become a great party, ...)

```
Quando          ”quando” <*> <rel> ADV                          @#FS-ADVL>  @ADVL>
estava          ”estar” <x+GER> V IMPF 1/3S IND VFIN                  @FAUX
lutando         ”lutar” <vi> V GER                                    @#ICL-AUX< @IMV
em              ”em” PRP                                                    @<ADVL
Stalingrado     ”*stalingrado” <*> <HEUR> PROP M/F S/P                       @P<
em              ”em” PRP                                                    @<ADVL
$1942           ”1942” <cif> <card> NUM M/F P                                @P<
$\,
o               ”o” <art> DET M S                                @>N
general         ”general” <+n> N M S                            @SUBJ>
alemão          ”alemão” ADJ M S                                 @N<
Friedrich       ”*friedrich” <*> <HEUR> PROP M/F S/P             @N<
Paulus          ”*paulus” <*> <HEUR> PROP M/F S/P                 @N<
$\,
a               ”a” <sam-> PRP                                  @<ADVL
o               ”o” <-sam> <art> DET M S                          @>N
descobrir       ”descobrir” <vq> V INF 0/1/3S                    @#ICL-P<  @IMV
que             ”que” KS                                              @#FS-<ACC  @SUB
estava          ”estar” <vK> V IMPF 1/3S IND VFIN                           @FMV
cercado         ”cercar” <vt> <ADJ> V PCP M S                               @<SC
por             ”por” <sam-> PRP                                             @A<
o               ”o” <-sam> <art> DET M S                                        @>N
Exército        ”exército” <*> N M S                                           @P<
Vermelho        ”vermelho” <*> ADJ M S                                          @N<
$\,
deve            ”dever” <x> V PR 3S IND VFIN                    @FAUX
ter             ”ter” <x+PCP> V INF 0/1/3S                      @#ICL-AUX< @IAUX
constatado      ”constatar” <vq> <ADJ> V PCP M S                      @#ICL-AUX< @IMV
que             ”que” KS                                                    @#FS-<ACC  @SUB
Hitler          ”*hitler” <*> <HEUR> PROP M/F S/P                                 @SUBJ>
não             ”não” <dei> <setop> ADV                                           @ADVL>
era             ”ser” <vK> V IMPF 1/3S IND VFIN                                   @FMV
um              ”um” <quant2> <arti> DET M S                                       @>N
estrategista    ”estrategista” N M/F S                                            @<SC
tão             ”tão” <dem> <quant> <KOMP> ADV                                      @>A
genial          ”genial” ADJ M/F S                                                 @N<
como            ”como” <rel> <komp> <igual> ADV                                  @#FS-KOMP<  @COM
parecia         ”parecer” V IMPF 1/3S IND VFIN                                          @FMV
$.
```

(When he was fighting at Stalingrad in 1942, the German general Friedrich Paulus, on discovering that he was surrounded by the Red Army, must have discovered that Hitler not was an ingenious strategist as it [had] seemed.)

List of concatenating verb currently used in the PALAVRAS lexicon

Legend:

x concatenating verb

<x> concatenating verb directly governing infinitive: *querer*

<x+PCP> concatenating verb directly governing past participle: *ter, haver, ser*

<x+GER> concatenating verb directly governing gerund: *estar, continuar*

<xp> concatenating verb governing infinitive with mediating preposition:
e.g. <a^xp> <de^xp> <por^xp>

<xt> concatenating verb with accusative object functioning as subject for the infinitive: *ver, ouvir, mandar*

<xd> concatenating verb with dative object functioning as subject for the infinitive: *permitir, aconselhar*

<xtp> concatenating verb with accusative object functioning as subject for preposition mediated infinitive
e.g. <a^xtp> <de^xtp> <por^xtp> *acostumar alg. a, lembrar alg. de*

<xdp> concatenating verb with dative object functioning as subject for preposition mediated infinitive
e.g. <a^xdp> <de^xdp> *proibir-lhe de*

<xr> concatenating reflexive verb directly governing infinitive: *propor-se*

<xrp> concatenating reflexive verb governing preposition mediated infinitive
e.g. <a^xrp> <de^xrp> <por^xrp> *recusar-se a, cansar-se de, esforçar-se por*

vt monotransitive full-verb governing direct object (@ACC)

vp monotransitive full-verb governing prepositional object (@PIV)

# (<x> column) does occur with simple infinitive (<x>)

* (<xp> column) fronting possible, at least in Brazil [(*) unsure]

xd/A dative pronoun can be replaced by PP with ”a”

ICL non-finite clause

P can be transformed into passive without major change in meaning (valid for first 3 columns)

In the internal table cells, where only a preposition-particle of a verb chain are given, pronouns and infinitives have to be “imagined” according to the valency pattern given at the top of each column. The default in the <xtp><xdp> column is <xtp>.

	+PCP +GER	<x>	<xp>	<xt> <xd>	<xtp> <xdp>	<xr> <xrp>	<vt/vp +ICL>	
@SUBJ	**same**	**same**	**same**	**other**	**other**	**same**	**other**	
ICL-initial pronouns, **fronting:** **role:**	**+,** **ICL-obj.**	**+,** **ICL-object**	**±,** **ICL-object**	**+** **ICL-subj.**	**+,** **ICL-subj.**	**+,** **reflexive**	**÷,** **ICL-subj.**	
pronoun form	**ACC** **DAT**	**ACC** **DAT**	**ACC** **DAT**	**ACC** **DAT**	**ACC** **DAT**	**se**	**NOM**	

abalançar						a		kaste sig ud i at
acabar P	GER		de*, por					holde op med at afslutte ved at
aceder			em*					gå med til at
aconselhar				xd/A	a			råde ngn til at
acostumar					a	a		vænne ngn til at
acusar					de			anklage ngn for at
admitir		# +ter/ser						indrømme at have
adorar		#						kunne lide at
afazer					a			elske at
agachar						a		begynde at
ajudar					a			hjælpe ngn at
ameaçar		#						true med at
andar P	GER							gå og
anelar		#	por					længes efter at
animar					a	a		opmuntre til at
ansiar		#	por					længes efter at
aparelhar						a		forberede sig på at
apreciar		#						værdsætte at
aprender		#	a*					lære at
apressar						a		presse til at
arriscar						a		vove at
arrojar						a		driste sig til at
aspirar		#	a*, ?em					stræbe efter at
assentar			de(*)					aftale at
assistir					a			hjælpe med at
atrever						a, com		turde
autorizar					a			bemyndige til at
botar			a(*)					begynde at
buscar		#						søge at
cansar						de		blive træt af
chegar P			a(*)					komme til at, opnå
coagir					a			tvinge til at
coincidir			em					INF'e samtidigt
começar P	GER		a*, de, por*					begynde at begynde ved at
compelir					a			tvinge til at
comprazer						em		finde fornøjelse i
condenar					a			dømme til at
confessar		#						indrømme at
confiar		#?	em		em			tro på,have tillid til
conformar						a		tilpasse til
conseguir		#						opnå at
contar		#	em					regne med at
continuar P	GER		a*, de, sem					fortsætte med at
convencer					a, de			overtale til at

convidar					a			indbyde til at
costumar P		#			a	a		pleje at
crer		#?					em^vp	tro (på) at
cuidar			em, de					tage sig af at
dar			a(*)				para ^vp	begynde
decidir		#			a, de	a		beslutte at, få til at
deitar			a(*)					give sig til at
deixar P			de*	xt, xd				holde med at, lade
deliberar		#				a		beslutte at
desandar			a*, em					begynde at, ende med at
desatar			a*					begynde at
desculpar					de			undskylde for at
desejar		#						ønske at
desistir			de*					afstå fra at
destampar			a(*), em					begynde at, udbryde i
determinar		#	de		a	a		beslutte, overtale til
dever P		#	de					måtte
dissuadir					de			fraråde ngn ngt
divertir						a		more sig ved at
dizer					de/para ^xdp		vt	give ngn ordre til at
empecer					de			hindre ngn i at
empenhar						em		tage sig af at
encarregar					de	de		tage på sig at
ensinar					a			undervise i at
entender			em					overveje at
entrar			a*, de					begynde at
envergonhar						de		skamme sig over at
escusar			de(*)		de	a, de		ikke behøve at, dispensere fra at, undlade, nægte
esforçar						por, em		anstrenge sig for at
esperar		#						håbe at
espraiar			em					gøre sig umage ved
esquecer						de		glemme at
estar P	GER (PCP)		a*, em, para*, por*					være ved af
estimular					a	a		tilskynde til at
evitar		#	de*					undgå at
excitar					a	a		opildne til at
expor					a	a		udsætte for at
falhar			a(*), de*					forsømme at
faltar			a(*), de*					undlade at
fartar						de		køre træt i

fazer				xt, xd		xdr		få til at, lade sig ...
ficar P	GER (PCP)		a*, de					blive
fingir		#						foregive at
folgar			de(*)					glæde sig at
guardar						de		beskytte ngn mod
habilitar					a			uddanne til at
habituar					a	a		vænne til at
haver P	PCP		de*					skulle
impedir					de			hindre i at
incitar				xt	a			opildne til at
induzir					a			overtale til at
intentar		#	de*					have til hensigt at
ir P	GER	#	a(*)					gå og, FUT, ville
isentar					de	de		fritage for at
lançar						a		kaste sig ud i at
largar			de*			a		opgive at, gå i gang med at
lembrar					de	de		huske at (gøre)
levar					a			få til at
lograr		#						have held med at
mandar				xt, xd			vt	give ordre til, lade
merecer		#						fortjene at
meter						a		sætte til at
necessitar			de*					have brug for at
negar		#	de(*)			a, de		nægte at
obrigar					a			forpligte til at
ocupar						a		beskæftige sig med
oferecer		#						tilbyde at
olvidar						de		glemme at (gøre)
opor							a^vrp	være imod at
ousar		#						vove at
ouvir				xt	a		vt	høre ngn ACI
parar P			de*					holde op med at
parecer P		#						se ud til at
passar P			a*					gå over til at
pegar			de(*)					begynde at
pensar		#	em*					tænke på at
permitir				xd/A	a^xdp		vt	tillade at
persuadir					a			overtale til at
planejar		#	de*					planlægge at
poder P		#						kunne
pôr					a	a		sætte til at
possibilitar				xd/A			vt	muliggøre at
precisar		#	de*					have brug for at
preferir		#						foretrække at
preparar						para		forberede sig på at

pretender		#						foregive at
pretextar		#						foregive at
procurar		#						føge at
proibir					de, a/de^xdp		vt	forbyde (ngn) at
projetar		#	de*					planlægge at
prometer		#	de*					love at
propor		# =+ICL	a*, de*			xr	vt	foreslå at
querer		#						ville, ønske at
recordar						de		mindes at
recusar						a		vægre sig ved at
resignar						a		affinde sig med at
resolver		#				a		beslutte sig til at
saber		#						kunne (viden)
seguir	GER							fortsætte med at
sentir				xt				føle ngn ACI
ser P	PCP							blive (PASSIV)
soer P		#						pleje at
sonhar		#	com, em					drømme om at
temer		#						frygte at
tencionar		#						have til hensigt at
tentar		#						forsøge at
ter P	PCP		de*vque*					skulle, måtte, PC
timbrar			em					lægge det an på at
tornar P			a*					INF'e igen
trabalhar			por					anstrenge sig for at
tratar			de*				de^vUr	prøve, handle om
vacilar			em					tøve med at
ver				xt				se ngn ACI
vir P	GER	#	a(*), de					komme til at, lige have
visar		#						sigte efter at
voltar P			a*					INF'e igen

4.4 Clause types and clause function

The parser distinguishes between three (sub)clause form types, *finite* (FS), *non-finite* (ICL) and *averbal* (AS), depending on whether the clause in question features a finite or non-finite head verb, or no verb at all. Function tags for subclauses are hyphen-attached to the subclause form tags tags (e.g. FS-<ACC for a finite *direct object* subclause). AS and FS tags are attached to the (obligatory) complementizer of these clause types (conjunction, relative or interrogative), while ICL tags are attached to the head verb of the clause. This way, clause tag bearing words will have a minimum of two tags, one relating to (intra-clausal) word/group function (@), the other to (higher level) clause function (@#).

In the following, the different clause types and their syntactic potential will be discussed and exemplified individually.

4.4.1 Finite subclauses

@#FS- finite subclause
(combines with clausal function and intraclausal word tag,
e.g.@#FS-<ACC @SUB for "não acredito *que* seja verdade")
@#FS-S< sentence anaphor
(refers back to the whole preceding clause '...., *o que* era novo para mim')

Finite subclauses cover a wide range of constituent functions, both free and valency bound. Many verbs allowing subclause arguments have semantico-syntactic selection restrictions concerning which clause types they allow. Most "cognitive" verbs, for instance, allow or even demand a *que*-clause or a finite interrogative subclause as direct object:

(1)

		A noiva não	*acreditava*	***que*** *ele a amasse*.
<vq>	[cognitive]	SUBJ (human)	V+que-"that"	FS-ACC (completive)
		A mãe	*perguntou*	***quando*** *viria*.
<v+interr>	[cognitive]	SUBJ (human)	V+qu-word	FS-ACC (interrogative)

Using a traditional - word class analogous - typology, one can distinguish between finite subclauses that cover the prototypical functions of nouns, adjectives or adverbs, respectively:

• Nominal FS, valency bound in clause/VP or PP, or as apposition

with absolute relative pronoun or adverb:

***Quem** cedo madruga* .... (SUBJ)

*Molesta **quem** apareçer*. (ACC)

*Seja **quem** for* (SC)

*Mostrava a pedra a **quem** quisesse ver*. (P<)

*O pai não veio para o aniversário dele, **o que** não o surpreendeu*. (S<)

with interrogative pronoun or adverb:

*Quis saber **quem** lhe mandara o presente*. (ACC)

*Não sei **quando** ele chegou.* (ACC)

with conjunctional *que*:

*Nem lhe parece estranho **que** o Pedro tenha comprado o sítio*. (SUBJ)

*Soube **que** foi o único candidato*. (ACC)

*Só foi avisado depois **que** o seu jatinho levantou vôo*. (P< or A<)

• Attributive FS (relative postnominal clauses), adject in NP

1. as modifier, with postnominal relative pronoun or adverb:

*O homem **que** encontrei ontem* (N<)

*A amiga com **a qual** apareceu na festa* (N<)

*O ano **quando** se casaram* ... (N<)

2. as argument, with *que* or interrogative pronoun or adverb

*A proposta **que** ele venha para aqui não me parece realística.*

• Clause level adverbial FS

1. as adjunct, with relative adverbial or subordinating conjunction

*João não fiz nada **para que** ela voltasse*. (ADVL, purpose)

*Entraram na vila **quando** amanheceu*. (ADVL, time)

*Desliga, amor, **que** tem gente na linha!* (ADVL, cause)

*Faz **como** quiseres!* (ADVL, måde)

2. as argument, with relative adverb

*O avô mora **onde** o mato começa*. (ADV-argument)

4.4.2 Non-finite subclauses (infinitives , gerunds, participles)

@#ICL- non-finite subclause
(combines with clausal function and intraclausal word tag,
e.g. @#ICL-SUBJ> @IMV in "*consertar* um relógio não é fácil")

In my dependency grammar non-finite subclauses (ICL) appear both as part of a hierarchically organised predicator in the verb chain (VC) and as ordinary constituents in clauses and groups. In the VC-case the ICL is functionally tagged as @ICL-AUX<, referring back to an auxiliary (which can itself be non-finite and @ICL-AUX<). In this chapter I will be concerned with ICL-functions outside the predicator. The most common cases are infinitive arguments:

• **Infinitive** as argument in VP

(1a) ***Retomar*** *o controle foi difícil*. (SUBJ)
(1b) *Manda o filho* ***comprar*** *leite*. (ACC, causative)
(1b’) *Manda o filho* ***comprar*** *leite*. (OC, causative)
(1c) *Viu o marido* ***bater*** *na mulher*. (ACC, perception verb "ACI")
(1c’) *Viu o marido* ***bater*** *na mulher*. (OC, perception verb "ACI")
(1d) *Julgo o carro* ***ser*** *caro demais*. (ACC)
(1e) *Não temos onde* ***morar***. (ACC)
Não tem quem ***perguntar.*** (ACC)
(1f) *O problema era* ***acabar*** *com os bandidos*. (SC)
(1g) *O problema é não* ***sermos*** *bastante fortes*. (SC)
(1h) *Disse ao amigo onde* ***comprar*** *um bom vinho.* (ACC)
(1i) *... se nao permitir a si mesmo* ***ser*** *apenas gente.* (ACC)
(1j) *Chama isso* ***fazer*** *tábua rasa.* (OC)

As a standard, the parser tags ICL’s in causative and ACI constructions (1b and 1c) as @#ICL-<ACC, with the nominal “accusative” element (‘filho’, ‘marido’) as subject (@SUBJ>) of the infinitive. There is, however, another possibility (1b’ and 1c’), with a matrix clause level @<ACC nominal and a - smaller - @#ICL-<OC (object complement). The existence of a clause level @<ACC nominal gives justice to the pronoun substitution test, that yields *accusative* pronouns:

o *manda comprar leite*
o *viu bater na mulher*

The @#ICL-<OC itself can also be substantiated by substitution with other OC material:

o manda ***para Brasil*** *(PP @<OC)*

*o manda **sozinho** (ADJ @<OC)*
*o viu **com outra mulher** (PP @<OC)*
*o viu **furioso** (ADJ @<OC)*

Since object complements in all other valency patterns are optional constituents, it should be possible to test the viability of a V - ACC - OC analysis by judging the grammaticality of the "naked" V - ACC string:

o manda comprar leite	- *?o manda.*
o faz comprar leite	- **o faz.*
o viu bater na mulher	- *o viu.*

The test suggests a difference between causatives and sense verbs, the latter testing positive, the former negative. 'mandar' superficially seems to pass the test, but changes its meaning underway ('to send' instead of 'to order'). Other causatives or ordering verbs, like 'fazer', 'permitir' etc., fail more obviously. In terms of parsing notation, one way of showing this structural difference between causatives/ordering verbs and sense verbs (ACI-verbs) would be choosing @SUBJ> @#ICL-<ACC in one case, and @<ACC @#ICL-<OC in the other[153].

• **Infinitive** as argument in NP
(2a) *Tem muito que **estudar***. (N<)

• **Infinitive** as argument in PP
(3a) *Era uma proposta difícil a **entender*** (P<)
(3b) *a possibilidade de eles não **aparecerem*** (P<)
(3c) *Para lhe **ajudar**, propôs outra solução.* (P<)
(3d) *Para o amigo lhe **ajudar**, bastava uma palavra só.* (P<)
(3e) *Pede para você **ficar** com ele.* (P<)

• **Infinitive** as sentence adjunct adverbial
(4a) *Veio lhe **agradecer** pessoalmente.* (ADVL)
(4b) *Foi à televisão **recitar** o documento.* (ADVL)

• **Infinitive** as complement in AS
(5a) *.. do=que **sucatear** suas próprias esperanças*. (AS<)

[153] For a detailed discussion of Portuguese transobjective constructions, see also "Portuguese Syntax", chapter 7.4 (Bick, 1999).

While it is one of the clausality tests for infinitive structures, that the subject of the ICL be other than that of the main clause, this is not an obligatory feature, as the pair (3c) - (3d) shows. In the case of different subjects, Portuguese can use personal (inflected) infinitives (1g, 3b), making it easier for the CG-rules to see what the subject is. Most of the above ICL-functions are subject to selection restrictions. Some cannot freely alternate with other nominal material (NP and FS), like (4a) and (4b), that allow only alternation with *adverbial* material, (1e) that only alternates with NPs, and the completly idiosyncratic (3a). Also the matrix verbs allowed in each case, are restricted classes, like perception, causative and cognitive verbs for ACC, and movement verbs for ADVL (4). ICLs headed by relatives seem to occur only with *'ter'* (1e, and, indirectly, 2a). For preposition argument ICLs, the lexical matrix restrictions reside not in the preposition (which is "transparent"), but in the next higher dependency level, if the PP is a postnominal complement (3a, 3b) or prepositional object (3e).

The prototypical usage of the gerund is in adverbial adjunct ICLs (1), while complement function is rare (2), literary (3), or fixed (4):

• **Gerund** as sentence adjunct adverbial

(1a) *<u>**falando** do João</u>, não quero convidá-lo.* (ADVL)

• **Gerund** as clause level argument (of verb)

(2a) *Como imagina-lo <u>**partilhando** à vera a administração com outros</u>*,...(OC)
*Mostrou gangues e organizações da extrema direita <u>**entoando** uma cantilena neonazista</u>* (OC)

• **Gerund** as argument in PP

(3a) *<u>Em **comendo**</u>, podes ir brincar.* (P<)

Again, *'ter/haver'* has its own, fixed construction[154]:

• **Gerund** as clause level argument (of verb)

(4a) *Tem <u>gente **morrendo** de fome no Brasil</u>.* (ACC)
(4b) *Tem <u>o motorista **esperando**</u>.* (ACC or ACC OC)

[154] Superficially, the @ACC gerund structures in (4) resemble the @OC gerund constructions in (2), and object complement is indeed an alternative tagging possibility. However, object complements demand direct objects to refer to, and ‘gente’, ‘motorista’ or ‘um garçon’ in (4) cannot be isolated as direct objects the same way ‘lo’ and ‘gangues’ can in (2):

Imagina-lo.
Mostra gangues e organizações da extrema direita.
*Tem gente no Brasil.
*Tem o motorista.

The difference is reminiscent of the one between ACI-sense-verbs (‘ver’, ‘ouvir’) and causatives or ordering verbs (‘permitir’, ‘fazer’), where the former allow a V + <ACC + ICL-<OC interpretation, while the latter only allow V + SUBJ> + ICL-<ACC.

(4c) *Há sempre um garçon* ***discutindo*** *um outro***.** (ACC)

Another fixed, clausal construction occurs with the preposition *com* and *sem*.. These two prepositions can function as a kind of "complementiser" in creating clause-like adverbials where what ordinarily would be the nominal complement of the preposition, is predicated in a clausal way by gerunds, place or time PPs, or APs expressing state. The gerund case can be treated in a consistent way, as an ICL:

• Prepositional "complementisers"

(5a) *... com um ator étnico* ***estrelando****.* (P<)

(5b) *... com 15.000 homens do Exército* ***patrulhando*** *a cidade.* (P<)

The semantic origin for this construction may be the aspectual use of *estar +GER*, as the gerund-alternation with *a+INF* suggests, analogous to the European Portuguese *estar +a+INF.*

(5c) *... com os @>N olhos @P<* ***a*** *@N<PRED* ***flamejar*** *@#ICL-P<.*

Here, the ICL is *within* a lower level PP, and not directly dependent on *com.* As in (5d-g), it is difficult to assign subclause status to a structure without a verbal constituent on the highest level. Though the construction is reminiscent of averbal clauses (AS) of the type *'While in Rome, ..'*, there are two important differences: (a) An AS is headed by an ordinary complementiser (never a preposition), legitimised as such by its appearance in ordinary, verb-containing subclauses, and (b), the body of an AS is "pure predication", i.e. a prepositional, adverbial or adjectival predicative, while there is an interfering nominal in the *com/sem* structure. Tagging this nominal (*os olhos* in 5c) as subject (@SUBJ>) *without* a main verb to attach the dependency marker to, is notationally problematic. Also, with the nominal @P< argument gone, no clear candidate would be left to bear the (now "clausal") @P< tag, stranding *com/sem* without a marked complement. For all these reasons, I prefer to introduce a special predicational linking tag, @N<PRED (postnominal nexus predicative), meaning that the tag-bearer predicates the preceding nominal in a "clausal", but verbless, way.

(5d) com todo=mundo @P< **seminu** @N<PRED

(5e) sem ela @P< **na** @N<PRED *casa* P<

(5f) com o @>N joelho @P< **fincado** @N<PRED *no* @A<ADV *morto* @P<

(5g) com o @>N dinheiro @P< *já* @>A **fora** @N<PRED *da* @A< *bolsa* @P<

Incidentally, the @N<PRED tag comes handy in a few other cases as well, as when a "sentence apposition" (@S<) itself is predicated:

(6) ..., tudo @S< **pago** @N<PRED *com* @A<PIV *meros* @>N *40.000* @P<

Like gerunds, past participles can appear in adverbial ICLs. The structures are equivalents of the Latin *ablativus absolutus*:

• **Participle** as sentence adjunct adverbial (ablativus absolutus)
***Feito** o trabalho, temos tempo para ...* (ADVL)

In analogy with the clausal gerund usage in (4), contrasted with the com/sem-predication in (5f) one might also expect (7a) or (7b), respectively:

(7a-4') Tem a @>N mão @SUBJ> **machucada** @#ICL-<ACC.
(7b-5f') Tem a @>N mão @<ACC **machucada** @N<PRED.

The clausality of such sentences is, however, syntactically ambiguous with an ordinary NP reading, and very hard to disambiguate, cf.:

(7c) Tem a flor @<ACC machucada @N<,mas esqueceu as outras provas do crime.

For a more detailed discussion, including the postnominal and predicative functions of participle structures, see chapter 4.4.4.

4.4.3 Averbal subclauses (small clauses)

@#AS- 'averbal' (i.e. verbless) subclause
(combines with clausal function and intraclausal word tag,
e.g. @#AS-<ADVL @ADVL> in "ajudou *onde* possível")
@AS< argument of complementiser in averbal subclause

The classical concept of a clause, which was advocated in the previous sections, is built around the notion of a main verb and its complements and adjuncts. Here the syntactic unit of *predicate* (bracketed in the examples in (1)) conveys information (the *predication*) concerning a state-of-affairs (SOA). A full-blown predicate would normally contextualise this predicational information by relating it to a subject (what the information/predication is about), either explicit (1a), anaphoric (1b) or - in certain Romance languages like Portuguese - implied by inflexion and valency (1c).

(1a) ***O seu amigo*** *[comprou um carro].* ('His friend bought a car.')
(1b) ***Ele*** *[trabalhava].* ('He' worked.')

(1c) *[Cheg**ou**]*. ('He/she/it arrived.)

A predication without a subject refers directly to the world:

(1d) *[Chove].* (‘it is raining’)
(1d') *[Faz frio].* (‘it is cold’)
(1e) *[Leia] !* (‘Read!’),

or even a quick warning like:

(1f) *Atenção, [quente] !* (‘Attention, hot!’).

If there *is* a subject to refer to, then what relates the predication to it, is usually - where present - the predicate's verbal part (underlined in 1a). Therefore, the verb (or verb chain) can be called *predicator*[155] (i.e. what predicates). Intransitive (1b) and ergative (1c) verbs can completely subsume the functions of predication and predicator, and most content verbs (action, activity, event and process verbs) are at least part of the predication (1a). Further, modals and certain other auxiliaries (*dever*, *querer*, *saber*, *ir +INF*, *começar+a+*INF, ...) might then be seen as predicators predicating a modality of a predicate (1g), creating a new, complex predicate[156]. Copula-verbs (*ser, estar, ficar,* ..), however, have nearly no semantic weight of their own, and are thus pure predicators. In these cases, the predication is averbal (verbless), as in (1h) where the predicator *estava* predicates the predication *doente* (in the post-copula adjective/noun case traditionally called a *predicative*) of the subject *criança.*

[155] To make a distinction between what predicates and what is predicated, also adresses the problem of how to mediate between syntactico-functional on the one side and semantico-functional distinctions on the other.

The syntactico-functional concept of predicator is primarily syntactic - it is a useful dependency hook for other clause level constituents like subject and object, and has been recognized as what I would like to call "the small (complement-free) VP" in traditional *English* PSG (as opposed to the object incorporating English “enlarged” VP and the Romance "big VP" embracing *all* verb-complements - including the subject). The term predicator is inspired by the usage in (Bache et. al., 1993), but my hierarchical left leaning dependency treatment of verb chains assigns complex predicators an *internal* structure different from the one advocated by Bache et. al, based on constituent analysis with the chain's head to the *right*. In particular, the dependency notation does not assign complex predicators constituent status in the same way, and hierarchical bracketing would - without transformation - turn main verb + complements into a complex (ICL-) dependent *"inside"* the auxiliary clause (cp. 4.3.1, (1))..

The concept of predication, on the other hand, is not about syntax proper, but about information structure. Thus syntactic elements can be seen as vehicles for a predication, allowing for both "information-free" purely syntactic units (like the copula case 1h) followed by information-bearing elements (like the predicative) and for multifunctional elements, where there is no one-to-one relation between syntactic and semantic function (like in 1g, where the predication *jointly* resides in two syntactic elements, the predicator and the direct object). My semantico-functional use of the term *predication*, must be distinguished from a syntactic definition like the one employed in (Bache, 1996, p.25), where a predication is a meta-constituent equalling the predicate minus (modal) operators, and can be identified by syntactic tests like co-ordination, fronting and substitution (by a pro-form).

[156] Notationally my parser captures this complexity by describing verb chains as a multi-layered clause hierarchy, using a new @#ICL-AUX< tag for each new inner layer, and regarding the first verb in the verb chain, finite or not, as the dependency head of the whole structure (cp. 4.3.1, (1)).

(1g) *Não [queria [**comer outro bolo**]]*. (He wouldn't eat another piece of cake.')
(1h) *a criança [estava **doente**]* ('the child was ill'),

Since it is semantically independent of the verb, predicational material that can appear in copula-clauses, like adjectives, attributive nouns (*comunista*), place- and time adverbials, is the most likely to appear in verbless predications. Predications like *quente* in (1f) or (2b) might in fact be regarded as elliptical predicates, where the copula has been omitted.

(2a) *Bebe o chá quente!* ('Drink the tea hot!')
(2b) *Bebe o chá quando quente!* ('Drink the tea while [it is] hot!')
(2c) **Bebe o chá quando ele quente!* ('Drink the tea while it [is] hot!')

The agrammaticality of (2c) shows that omission of the predicator entails obligatory omission of the subject - which could otherwise be added exploiting the verb's number and gender information in an anaphoric way. Therefore, *quente* has to be predicated directly of a main clause entity (or of the world), - here the direct object *chá*. In (2a) this is no problem, since we have the clause-level function of object complement (@OC) to account for this phenomenon. Both @OC and the similar @SC (subject predicative complement) and @PRED (free predicative) are obvious cases where predications are clause-level constituents, the difference being that @OC and @SC are usually valency-bound, while @PREDs are not.

In (2b), however, *quente* is isolated from the main clause by - a complementiser. Though both predicator- and subjectless, one could argue that *quando quente* is still a kind of clausal entity, since it boasts a subordinator (the relative adverbial *quando*), that can help establish a contextualised SOA even without the help of a predicator. Rather than, for instance, solving the problem by calling *quando* in (2b) a preposition (something which is quite common in the analogous case of the comparative *como*), I would like to argue that it is still a complementiser, and that its semantic content, temporal, spatial or comparative, is what turns *quente* into a contextualised SOA. Here, the SOA is not predicated of a grammatical subject, but yet of part of its intensional potential (conditioned by a when, where or how). Interestingly, most instances involve adverbial complementisers and conditional conjunctions (like *embora* 'though'), while the "pure" (completive) complementiser *que* ('that'), which is void of semantic content, can *not* occur in this kind of (averbal clausal) construction - unless the construction is comparative with *que* featuring at least anaphoric semantic content[157] (3a).

[157] This is not at all a contradiction, since comparative 'que' is not completive, and could, in fact, be treated as a different lexeme, a classification for which there is other "circumstantial evidence", both diachronic and translational (cp. 4.5.2). Thus the two meanings 'that' (completive) and 'than/as' (comparative) of Portuguese *que* can be etymologically traced to two different Latin origins, 'quod' and 'quam', respectively. The view that comparative *que* anaphorically "borrows" semantic

(3a) *O filho é mais alto que o pai [é].* (The son is taller than the father [is].)
(3b) *O filho é tão alto como o pai [é].* (The son is as tall as the father [is].)
(3c) *O filho é um homem como um urso [é].* (The son is a man like a bear [is]).

The comparison structures in (3) lack a predicator, too, but they do have a subject, and the "internal" function of the relative adverbial comparator (3b-c) can be regarded as a predication, or at least a dummy substitute for a predication: 'A bear is *like that* ' (3c) or 'the father is *as/less tall* ' (3b). Even the otherwise semantically "empty" conjunction *que* (3a) and the special polylexical comparator *do=que* (3a') gain semantic content from their comparative function, "borrowing" the missing predication from the comparative kernel they measure against.

An argument for not reading comparators as prepositions - in Portuguese - is the fact that they do not demand prepositional ("oblique") case of personal pronouns, but rather plain nominative case:

(3a') *O filho é mais alto do=que ele* (The son is taller than he [is].)
(3b') *O filho é tão alto como ele [é].* (The son is as tall as he [is].)

Finally, a few comparisons are structured much like (2), with predications and without subject:

(3d) *Ele fala como [ele] pensa.* (He talks like he thinks).
(3e) *É tão avaro como [?ele][é] rico.* (He is as .. as [he is] rich.)
(3f) *É mais avaro do=que [?ele][é] rico.* (He is more .. than [he is] rich.)

While (3d) has a predicator (which is at the same time the predication), (3e-f) match (2) quite nicely, with no obvious predicator. However, subject omission does not seem as obligatory in the face of a missing predicator, indicating, perhaps, that the comparators here do not (like *quando* in (2)) contextualise the subject, but merely provide a link to yet another predicator. Since comparative constructions include contrasting, there are cases where subject omission is altogether impossible for semantic reasons, as in (3g), where gender inflexion of the predicative adjective forces a reading with two different subjects:

(3g) *(Ele) é tão avaro como ela rica.*

content from a hooked comparative kernel in much the same way as *como* does, could, in fact, be used to make it a member of the relative adverbial class (<rel> ADV) in my system, which is exactly where Latin 'quam' would belong ...

Only for "hook-less" comparisons there is a slight difference - 'que' *can* be used to replace 'como', but onlywith at least a comparative kernel to "borrow" from*:* "Ela é **linda** que nem um anjo" ('She is beautiful as not even an angel [is]'), Without an adjective like *linda*, i.e. after non-attributive nouns like in(3c), 'qu*e'* can *not* replace 'como'.

Apart from the presence of predications or subjects, yet another argument, motivated by descriptional consistency, can be cited in favour of elliptic clausal readings: - the fact that, in Portuguese (unlike English and Danish), the co-occurrence of complementisers (conjunctions and relative or interrogative adverbs) and finite subclauses is 100%: There are no finite subclauses without complementisers, so why should there be complementisers without (at least incomplete) subclauses?

In my system, I call this kind of predicator-elliptic, but subordinated clauses for **averbal subclauses (@#AS).** ‘Averbal small clauses’ would be another possibility, but might perhaps create confusion, since the term ‘small clause’ has been suggested, with another meaning, for English clauses *with* verbs and *without* complementisers (Radford, 1988).

In terms of CG-notation, marking such averbal subclauses (@#AS) is both unproblematic and logical: Clause form and function has so far been marked on the main verb for non-finite subclauses (@#ICL), and on the (obligatory) complementiser for finite subclauses (@#FS), so averbal subclauses, featuring a complementiser but lacking a verb, belong naturally in the second group, together with @#FS-clauses.

Having accepted the notion of averbal clause, one then finds that every AS has two functional obligatory parts, (a) the complementiser, that bears tags for both internal function (usually, that of adverbial or subordinator), and external function, and (b) the predication. Due to the implicit notation, in CG, for complex constituents, complementiser and predication have to be linked by a dependency relation in order to “assemble” the clause. Though the predication/subject unit is doubtless the semantic kernel of the whole (and can even sometimes replace it, as seen in the @OC example 2a), I have chosen the same technique as in PPs, where the linking element (the PRP) is regarded as dependency-head. For AS-predications, this link can then be expressed as @AS<, reducing function to almost pure dependency.

In terms of word class material, predications have to be “predicative” when the ellipsed verb is a copula (4a-b, 5a-b) and the construction is not comparative, but may even here belong to different word classes (ADJ, “attributive” N, locative ADV or PP). For comparative constructions (3c-d), the range of permissible material is very wide, since it is not necessarily predicational and its type depends on the comparative base it has to match.

(4a) Quando [quando] <rel> <ks> ADV @ADVL @#AS-ADVL> 'when'
jovem [jovem] <h> <ante-attr> **ADJ** M/F S @AS< 'young'
, Inocêncio integrou a corrente política Ação Popular.

(4b) Quando [quando] <rel> <ks> ADV @ADVL @#AS-ADVL> 'when'
garoto [garoto] **N** M S @AS< 'a boy'
, não gostava de brigar.

(4c)	Como	[como] <rel> <ks> ADV @COM @#AS-ADVL> 'as'
	sempre	[sempre] <dei> <atemp> <setop> **ADV** @AS< 'always'
	, só concorreram candidatos filiados ao Partido Comunista.	

(4d)	.....,	
	bem=como	[bem=como] <ks> <prp> ADV @COM @#AS-<ADVL 'just like'
	em	[em] <sam-> **PRP** @AS< 'in'
	as	[a] <-sam> <art> DET F P @>N 'the'
	epístolas	[epístola] N F P @P< 'epistles'
	de	[de] <+top> PRP @N< 'of'
	São=Paulo	[São=Paulo] PROP M S @P< ‘Paulus’

Though relative adverbials, like in (4) and all comparative constructions (6) are the predominant complementiser class for averbal clauses, ordinary conjunctions do occur:

(5a)	Se tiver uma parte,	
	ainda=que	[ainda=que] <+SUBJ> KS @SUB @#AS-<ADVL 'even if'
	pequena	[pequeno] **ADJ** F S @AS< 'small'

(5b)	Tiveram de enfrentar a concorrência de similares estrangeiros,	
	se=bem=que	[se=bem=que] KS @SUB @#AS-<ADVL ‘though’
	associados	[associar] <a^vtp> <jn> <ADJ> **V** PCP M P @AS< ‘affiliated’
	a	[a] <sam-> PRP @A< ‘with’
	a	[a] <-sam> <art> DET F S @>N ‘the’
	tecnologia	[tecnologia] <am> N F S @P< ‘technology’
	de	[de] <sam-> PRP @N< ‘of’
	aqui	[aqui] <-sam> <dei> ADV @P< 'here [this country]'

As can be seen from the examples in (4) and (5), the functional distribution of averbal subclauses in my corpus includes primarily adverbials (@#AS-ADVL) and comparatives. Of the latter, “hooked” comparisons (i.e. with a <KOMP> antecedent like in 6a, cp. section 4.5.2) receive a special dependency-tag (@#AS-KOMP<), while “absolute” comparisons can attach directly to a nominal comparative base (@#AS-N< in 6b, @#AS-A< in 6c), - or to the main verb, predicate (6d) or clause (5e), in which case they will be treated as adverbials (@#AS-ADVL).

(6a)	com	[com] PRP @<ADVL ‘with’
	um	[um] <arti> DET M S @>N ‘a’
	distanciamento	[distanciamento] N M S @P< ‘distancing’
	maior	[grande] <KOMP> <corr> <jn> ADJ M/F S @N< ‘bigger’
	que	[que] <komp> <corr> KS @COM @#AS-KOMP< ‘than’
	o	[o] DET M S **@AS<** ‘that’
	de	[de] <sam-> <+hum> PRP @N< ‘of’
	os	[o] <-sam> <art> DET M P @>N ‘the’

	demais	[demais] DET M/F S/P @>N ‘other’
	brasileiros	[brasileiro] <N> N M P @P< 'Brazilians'

(6b)	historiadores	[historiador] <prof> N M P @<SC 'historians'
	como	[como] <rel> <ks> <prp> ADV @COM @#AS-N< ‘like’
	os	[o] DET M P **@AS<** ‘those’
	de	[de] PRP @N< ‘of’
	hoje	[hoje] <dei> <atemp> ADV @P< ‘today’

(6c)	ambicioso	[ambicioso] <h> ADJ M S @<PRED 'ambitious'
	como	[como] <rel> <prp> ADV @COM @#AS-A< ‘like’
	um	[um] <arti> DET M S @>N ‘a’
	César	[César] PROP M S **@AS<** 'Cæsar'
	de	[de] <sam-> <+top> PRP @N< ‘from’
	o	[o] <-sam> <art> DET M S @>N ‘the’
	sertão	[sertão] <top> N M S @P< ‘wilderness’
	de	[de] <sam-> PRP @N< ‘of’
	o	[o] <-sam> <art> DET M S @>N '-'
	Arkansas	[Arkansas] <HEUR> PROP M/F S/P @P< ‘Arkansas'

(6d)	vende entre os roqueiros	
	assim=como	[assim=como] <rel> <ks> ADV @COM @#AS-<ADVL ‘like’
	o	[o] <art> DET M S @>N ‘-’
	jazz	[jazz] <ll> N M S **@AS<** 'jazz'
	entre	[entre] PRP @N< @<ADVL ‘among’
	os	[o] <art> DET M P @>N ‘the’
	jazzófilos	[jazzófilo] <attr> N M P @P< 'jazzophiles’

(6e)	Segundo	[segundo] <rel> <prp> ADV @COM @#AS-ADVL> ‘according to’
	sua	[seu] <poss 3S/P> DET F S @>N ‘his’
	estimativa	[estimativa] N F S **@AS<** ‘estimate’
	$,	

In all of (6), the clause-internal function of the complementiser is that of comparator (@COM), embodying the juxtaposition of what is compared and what it is compared to in much the same way as a copula predicator (which is one of the arguments in favour of a clausal analysis). Like copula verbs, however, such juxtaposing complementisers can either predicate a “how” or a “what”. Compare (7a) an (7b), where the English translation, ‘like’ and ‘as’, respectively, shows the difference. For Portuguese, CG-rules can exploit article clues: como/like usually governs a non-definite NP, while como/as usually governs an NP without an article.

(7a) Ele trabalhava como @COM um escravo. (‘He worked like a slave’)
- Ele é como um escravo. (‘He is like a slave [“slavish”]’)

(7b) Ele trabalhava como @PRD escravo. (‘He worked as [a] slave’)
- Ele é um escravo. (‘He is a slave’)

In my system, I mark this distinction by using a special @PRD tag (“role predicator” or “predicative subordinator”) for cases like (7b). Though both AS’s will be tagged as functioning adverbially (@#AS-ADVL), the second contains a significant element of “predicativity”. The AS in (7b) is much like @PRED or @SC, since it predicates something of the subject. To take the analogy even further, consider (7c), where the AS is predicated of the *object* in much the same way as an @OC (object complement) would.

(7c) A mai tratava a filha @<ACC como @PRD @#AS-<ADVL escrava @AS<.

In contrast to comparisons, role predication of both subjects and objects is lexically more restricted, especially for the object predicating cases. I have therefore chosen to treat the resulting disambiguation problem in terms of valency, and have established corresponding valency information in the lexicon:

(8a) trabalhar <como^va> ‘to work as’
(8b) lembrar <como^vta> ‘to remember (s.o.) as’

Only transobjective valency of type (8b) permits @PRD arguments after participles. These will “clause-externally” be tagged as adverbials in passive constructions (9a), but as post-adjects (@#AS-A<) after attributive or predicative participles (9b).

(9a)	sera	[ser] <x+PCP> V FUT 3S IND VFIN @FAUX ‘will be’
	lembrado	[lembrar] <vt> <v-cog> V PCP M S @IMV @#ICL-AUX< ‘remembered’
	como	[como] <rel> <prp> ADV @PRD @#AS-<ADVL ‘as’
	um	[um] <quant2> <arti> DET M S @>N ‘a’
	período	[período] <per> N M S @AS< 'period'
	inigualável	[inigualável] <n> ADJ M/F S @N< ‘incomparable’

(9b)	ficou	[ficar] <vK> <v-cog> <ink> V PS 3S IND VFIN @FMV ‘became’
	conhecido	[conhecer] <vt> <PA> <ADJ> V PCP M S @<SC ‘known’
	como	[como] <rel> <prp> ADV @PRD @#AS-<ADVL ‘as’
	um	[um] <quant2> <arti> DET M S @>N ‘a’
	professor	[professor] <prof> N M S @AS< ‘teacher’
	um=tanto	[um=tanto] <quant> <det> ADV @>A ‘a little’
	iconoclasta	[iconoclasta] N M/F S @N< ‘iconoclastic’

The averbal small clause concept allows for an elegant nesting analysis of chains of complementisers with only one finite verb to share among them. Usually, the first element is a comparator and the second a subordinating conjunction (10a) or absolute relative (pronoun, in 10b, or adverb, in 10c).

(10a)	discutiram	[discutir] <vt> <vH> V IMPF 3P IND VFIN @FMV 'discussed'
	a	[a] <art> DET F S @>N ‘-’
	privatização	[privatização] N F S @<ACC 'privatisation'
	como	[como] <rel> <ks> ADV @COM @#AS-<ADVL ‘as’
	se	[se] **KS** @SUB **@#FS-AS<** ‘if’
	não	[não] <dei> <setop> ADV @ADVL> ‘not’
	conhecessem	[conhecer] <vt> <IA> <vH> V IMPF 3P SUBJ VFIN @FMV ‘[they] knew’
	o	[o] <art> DET M S @>N ‘the’
	varejo	[varejo] N M S @<ACC ‘control’
	de	[de] <sam-> PRP @N< ‘of’
	o	[o] <-sam> <art> DET M S @>N ‘the’
	programa	[programa] N M S @P< 'program'
	econômico	[econômico] ADJ M S @N< ‘economic’

(10b)	Camargo	[Camargo] <HEUR> PROP M/F S/P @SUBJ> 'Camargo'
	saiu	[sair] <ve> <sH> V PS 3S IND VFIN (B) @FMV 'left'
	como	[como] <rel> <prp> ADV @COM @#AS-<ADVL ‘like’
	quem	[quem] **<rel>** <hum> **SPEC** M/F S/P @SUBJ> **@#FS-AS<** ‘[one] who’
	toma	[tomar] <vt> <v-cog> <vH> V PR 3S IND VFIN @FMV ‘takes’
	uma	[um] <quant2> <arti> DET F S @>N ‘a’
	decisão	[decisão] N F S @<ACC ‘decision’
	pessoal	[pessoal] <n> ADJ M/F S @N< ‘personal’

(10c)	Em outras ocasiões Wilson é mais convincente	
	$,	
	como	[como] <rel> <ks> <prp> ADV @COM @#AS-<ADVL ‘like’
	quando	[quando] **<rel>** <ks> **ADV** @ADVL> **@#FS-AS<** ‘when’
	chama	[chamar] <vt> <para^vp> <vH> V PR 3S IND VFIN @FMV ‘[he] draws’
	a=atenção	[a=atenção] <acc> <chamar+> <+para-piv> VNP @<ACC ‘attention’
	para	[para] PRP @<PIV ‘to’
	os	[o] <art> DET M P @>N ‘the’
	mistérios	[mistério] <ac> N M P @P< ‘mysteries’
	contidos nos episódios da paixão e morte de Jesus	

Less common, the first element of the nesting construction is a concessive conjunction (10d), or the second element a non-finite clause (gerund based in 10d, infinitive-based in 10e):

(10d)	conta tudo,	
	embora	[embora] <+SUBJ> KS @SUB @#AS-<ADVL ‘though’
	pulando	[pular] <vi> <vH> V **GER** @IMV **@#ICL-AS<** ‘skipping’
	pedaços	[pedaço] <er> <r> N M P @<ACC ‘bits and pieces’

(10e)	é mais fácil	
	que	[que] <komp> <corr> KS @COM @#AS-KOMP< ‘than’
	a	[a] <art> DET F S @>N ‘-’
	gente	[gente] <HH> N F S @SUBJ> ‘we’
	recomeçar	[recomeçar] <vt> <vH> V **INF** 0/1/3S @IMV **@#ICL-AS<** ‘recommence’

In the above, I have argued in favour of a third type of clause, neither finite nor non-finite, *averbal small clauses*, and tried to define this category as predicator-less, but complementised elliptic clauses, retaining - as a kind of semantic clause trunc - either the predication or (in comparisons) some other focused clause constituent. Still, predicator ellipsis also occurs in clauses with neither predicatives nor comparison complementisers. Consider the following two uses of *qual*, one as a relative adverbial (11a), one as an interrogative determiner (11b):

(11a)	nada	[nadar] <vi> <vH> <ink> V PR 3S IND VFIN @FMV ‘[he/she] swims’
	qual	[qual] <rel> <Rare> ADV @COM @#AS-<ADVL ‘like’
	um	[um] <quant2> <arti> DET M S @>N ‘a’
	peixe	[peixe] <ich> N M S **@AS<** ‘fish’

(11b)	não	[não] <dei> <setop> ADV @ADVL> ‘not’
	consigo	[conseguir] <x> <vH> <ink> V PR 1S IND VFIN @FAUX ‘[I] succeed’
	saber	[saber] <vt> <+interr> V INF 0/1/3S @IMV @#ICL-AUX< ‘get to know’
	qual	[qual] <interr> DET M/F S @#FS-<ACC ‘which [is]’
	a	[a] <art> DET F S @>N ‘the’
	diferença	[diferença] <ac> <+entre> N F S **@<ACC** ‘difference’
	de	[de] <+feat> PRP @N< ‘in’
	qualidade	[qualidade] <feat> <featc> N F S @P< ‘quality’

While (11a) receives the same analysis as the comparatives in (6) and (7a), (11b) does not match the AS-pattern, since *qual a diferença de qualidade* lacks the support of syntactic parallels in the main clause that comparative ellipses would normally enjoy. Also, the function of the subclause as a whole is different, - in contrast with all the above AS-constructions, it is valency bound by the main clause main verb, *saber* (as a direct object). Therefore, the CG rule grammar has no ready made solution for the copula predicator ellipsis in this case. (11b) is, in fact, a linguistically interesting case of "unforeseen" parser input - and a chance to judge its robustness[158].

Since a Constraint Grammar, due to its "subtractive" nature, can't break down, some analysis is always given. In this case, the interrogative forces a clause reading, even without a verbal constituent, and the valency projection of *saber* even allows assignment of the correct clause function, direct object (@#FS-<ACC). However,

[158] Ideally, 'qual' would receive a @SC> tag in (11b), and 'diferença' would be @<SUBJ, but the absence of a head-verb to anchor these functional tags leads to their removal by Constraint Grammar rules.

without a predicator's valency information, no clause internal function can be assigned to the complementiser, and *diferença* orients itself towards saber, the only verb present, though isolated by a clause boundary word. Interestingly this verb-complementation yields more or less the same sentence meaning with *qual* excluded as the real sentence would with *qual* included ('I can't see the difference in quality). All other words' readings, since they rely on local context only, are both complete and correct.

4.4.4 Past participle constructions

Past participles, though clearly deverbally derived, obey my morphological class criteria for adjectives in all but the tense-constructions with *ter* , that is, like adjectives they inflect for both number and gender, none being a lexeme category. I have therefore chosen, where possible, to assign analyses analogous to APs[159], with the postnominal @N< and the predicative @<SC and @<PRED being the leading syntactic functions. Dependents of past participles, whether complements or not, must then logically be tagged as adverbial adjects, @>A or @A<. There is, however, a case, where a participle functions neither adnominally nor predicatively, but *adverbially:* the so-called "ablativus absolutus". Here, a clausal analysis is compelling, both because the participle - being a direct constituent of the clause - has no clear head to attach to, and because a subject is provided in the form of the "ablative" nominal.

4.4.4.1 Ablativus absolutus

In an *ablativus absolutus* construction, the - obligatory - subject is *preceded* by the past participle. Both sentence initial and sentence final position are allowed, the function being that of adverbial. CG-disambiguation is helped by near-obligatory comma-separation.

(1a) ..., **dado** @#ICL-<ADVL o @>N caráter @<SUBJ dos @N< dois @P<

(1b) **Feito** @#ICL-ADVL> o @>N trabalho @<SUBJ, temos tempo para ...

One might argue that the semantic role of *o trabalho* is that of patient, and that a direct object tag (@<ACC) should therefore be assigned, as in the finite full clause *'tem feito o trabalho'*. However, such an interpretation of the past participle as active cannot be maintained in the face of number and gender agreement relations between participle and nominal, typical of passive constructions *('os trabalhos são feitos')*, where the surface syntactic reading of the patient role is subject, not object.

[159] In the Helsinki CG of English, both past and present participles are assigned their own word classes on purely morphological grounds, creating an '-ed' (PCP2) and an '-ing' (PCP1) word class, respectively.

Another argument in favour of the @SUBJ reading in absolute participle constructions is that they resemble @SC participles after ’estar’. Both types of participles allow only transitive and ergative verbs, not intransitive inergative verbs. And both predicate something of a *patient* subject. As a matter of fact, for *ergative* verbs, the @SUBJ of an absolute participle construction corresponds to the *subject*, not the object, of the corresponding finite clause:

(2) **sumido** @#ICL-ADVL> o bandido @<SUBJ, as vítimas se consolaram
o bandido @SUBJ> sumiu, e as vítimas se consolaram

Superficially, the participle construction in (3a) resembles that in (3b), with a PP replacing the NP to the right of the participle. However, the participle in (3b) has a clear dependency relation, corroborated by agreement, to the main clause's subject, *a pintura*. The semantic role patient of the participle is situated *outside* the participle structure itself!

Within my range of function tags, the most sensible reading in (3b) is that of free subject predicative (@PRED), listing this case under the heading of "adjective-like" participles.

(3a) E @CO **arrancada** @#ICL-ADVL> a @>N chave @<SUBJ do @<PIV cofre @P<, ...

(3b) **Comprada** @PRED> em @A< Londres @P<, a pintura parecia autêntica.

4.4.4.2 Participle valency and arguments

Apart from the ablativus absolutus case, inflecting participles - the ones I would like to call "adjective-like" due to their morphological categories - have about the same syntactic distribution as adjectives, preferring post-nominal and predicative positions. Furthermore, they take intensifying modifiers (@>A, @A<) in much the same way adjectives do:

(1a) um estilo **muito** original/apurado/evolvido
(1b) um estilo chato/exagerado/subdesenvolvido **demais**

In fact, many participles of the type (1a, 1b) are dictionary-listed as adjectives as well. Participles can, however, be combined with "heavier" modifiers than most adjectives, expressing for example temporality or locality, increasing in weight through (2). The adverbial adjects in question project a certain degree of "verbality" onto the participle, and I will therefore use the complex tags @A<ADVL and @ADVL>A in these (non-quantifying) cases, indicating adverbial adjunct function within the participle structure (“participle adjuncts”).

(2a) um livro **antigamente** @ADVL>A muito @>A *apreciado* @N<
(2b) um mar **nunca** @ADVL>A **antes** @ADVL>A *navegado* @N<
(2c) *publicado* **este** @>N **ano** @A<ADVL ***na*** @A<ADVL***revista*** @P<***VEJA*** @N<
(2d) o processo **contra** @ADVL>A **mim** @P< *movido* @N<

The need for clause level function tags becomes even more compelling, when participles retain their root-verbs' valency in a visible way, adding complements. For example, with the exception of the perfective form of ergatives (e.g. *chegado em Londres*), most participles can be viewed as passivisation of transitive verbs, allowing attachment of the active verb's agent-subject as argument of a *por*-PP, functionally tagged as @A<PASS (3a,b,c), in analogy with the "real" passive (3d):

(3a) o país *transformado* @N< **pela** @A<PASS **campanha** @P<
(3b) um candidato *apoiado* @N< **por** @A<PASS **Farias** @P< e @CO **pela** @A<PASS **primeira-governadora** @P<
(3c) a medida *anunciada* @N< **pela** @A<PASS **fabricante** @P< **de** @N< **eletrodomésticos** @P< **Westinghouse** @N< há @A<ADVL duas @>N semanas @P<
(3d) o país foi @FAUX transformado @#ICL-AUX< **pela** @<PASS **campanha** @P<

Other clause-level complements may include predicative complements (4a), adverbial objects (4b) or prepositional objects (4c).

(4a) um recurso *chamado* @N< **agravo** @A<SC **regimental** @N<
(4b) a cantiga dolente @N< e @CO rouca @N<, *atirado* @N< **aos** @A<ADV **ramos** @P<
(4c) o navio estava *pintado* @<SC **de** @A<PIV **ouro** @P<.

Note that the parser views participle constructions like the above as structurally similar to *passive* clauses, with passive agents (@A<PASS) replacing the (active clause) subject, and what would be object complements (@OC) - in an analogous finite active clause - becoming subject complements (@A<SC) in the passive participle “clause”. Argument adverbials are not marked for subject or object relation, so the tag @ADV remains the same as in the finite active clause, yielding @A<ADV in (4b).

The combination of adject dependency tag and clause-level function tag nicely captures the distinction between "state passive" (4c, with *estar*) and "action/event passive" (4d, with *ser*), the latter featuring only *one* clause level, because the participle binds to an auxiliary within a verb chain constituent:

(4d) o navio foi pintado **de** @<PIV **ouro** @P<.

In (4c), on the other hand, a distinction could be made between the main clause, with a copula verb and a participle subject complement, and a participle subclause with the participle verb, not the VC as a whole, governing the prepositional object.

Finally hybrid tags may be a solution for the adverbial PP version of the ablativus absolutus case in (2) in the last chapter, which is here repeated in its original form as (5a) and “PP-ised” in (5b).

(5a) E @CO **arrancada** @#ICL-ADVL> a @>N chave @<SUBJ do @<PIV cofre @P<, ...

(5b) E com @ADVL> a @>N chave @P< arrancada @N<PRED do @A<PIV cofre @P<, ...

In (5b), short of tagging *chave* as subject and *arrancada* as ICL-P< (which was implemented in an earlier version of the parser), the best compromise between a clausal reading and one that retains the NP-cohesion[160] of the preposition-complement seems to be the two hybrid tags @N<PRED and @A<PIV, the first ascertaining "subject-ivity" for *chave* by *predicating* it the same way a free subject complement does (<PRED), the second retaining the prepositional object function (<PIV) of the post-participle dependent *do cofre*, while at the same time marking the basic NP-pattern's modifier hierarchy:

chave @NPHR
chave @NPHR arrancada @N<
(chave @NPHR (arrancada @N< do_cofre @A<))

[160] An argument in favour of the NP-reading is the (for Portuguese) typically attributive position of the participle *after* the noun, contrasting with the ablativus absolutus case, where the participle is located *before* the noun, a position otherwise reserved for light material or determiners.

4.5 Under the magnifying glass: Special research topics

4.5.1 Predicative constructions

4.5.1.1 Copula mediated subject complements and focus assignment

In predicative constructions with copula verbs (<vK>), as I define them here, a copula predicator predicates one nominal constituent (the predicative or subject complement) of another (the subject). Since Portuguese uses two lexically different verbs to cover the lexical space of English 'to be', one as a kind of "ontological" identity predicator for mainly nominal subject complements ('ser' <vK>), and the other as a "state" predicator for both attributive subject complements and (locative) adverbials ('estar' <va>), it makes sense to uphold this distinction for copula-like constructions in general. In this chapter, I will thus discuss only the first type, @SUBJ + <vK> + @SC. Tests with prototypical material show that even Portuguese word order is not entirely free in these cases (i-vi). I have chosen a non finite subclause as subject (underlined in the example), since an ICL - in the absence of another subject candidate ICL - can not normally function as subject and will thus force an unambiguous assignment of syntactic function.

(i) *Fazer isso (não) é perigoso*. [inversion: ?*Perigoso (não) é fazer isso*]
(ii) *(Não) é perigoso fazer isso*. [inversion: **(Não) é fazer isso perigoso*]
(iii) *?Perigoso fazer isso (não) é*. [inversion: **Fazer isso perigoso (não) é*]

The inversion test failures (where the nominal elements were exchanged) document both the "non-predicativity" of infinitives, and - as a consequence - that there *is* a "normal", if not fixed, sequence for copula constructions. The regularity shows if one considers the clause as a 3-element circular chain to be broken in one position, or a continuous 3-piece-segment to be cut from an infinite string:

..... SUBJ - VK - SC - SUBJ - VK - SC - SUBJ - VK .....

If one, with Togeby (1993, p.111), information-theoretically defines as *topic* what an utterance uses as a (known) point of departure, and as *focus* that constituent which offers relevant new information in such a way that it will be affected by a logical negation of the clause, - then subject-hood, definiteness and clause-initial position suggest topic function, while complements other than subject, indefiniteness and clause-final position suggest focus function. Since what subject complements do is predicate (i.e. reveal information) in a negatable way, and what they relate to is subjects, they seem to be ideal candidates for focus function, and the allowed positions for *perigoso* and *fazer isso* in (i)-(iii) should therefore indicate the typical topology of (information-theoretic) topic and focus in Portuguese copula constructions:

Portuguese can place subject and subject complement on either side of the copula verb (i), or both on the same side (preferably right). In the first case, the topic comes to the left, in the second (both on the same side), it takes the last position. The focus position is usually directly to the right of the copula, but fronted in the rather marked case (ii), where both subject and complement precede the copula..

The nominal material involved in these constructions can be divided into 6 groups:

A definite NP, including names: *a Moura, o número, o chefe, Maria,*
3 dos melhores, um dos problemas
B personal pronouns in the nominative case and demonstrative pronouns:
ele, eu, isso, esse
C absolute relative nominal subclauses *quem fala, que recebi*
D infinitive subclauses, interrogative and que-subclauses:
fazer isso, retomar o controle
E attributives (adjectives, participles, demonstrative determiners and certain "attributive" nouns without a determiner):
famoso, casado, chefe, presidente
F indefinite NP: *um ladrão, amigos*

These groups can be placed in the following ways:

with the subject in topic position and the predicative as focus

	topic		**focus**
	subject		**predicative**
1.	ABCD	<vK>	ABCDEF
	Maria	*é*	*bonita.*
	Quem falta	*são*	*os irmãos.*
	O chefe	*sou*	*eu.*
	Retomar o controle	*foi*	*difícil.*
	3 dos melhores	*eram*	*doentes.*
	O problema	*era*	*acabar com o bandido.*
	Essa [regra]	*é*	*a regra da democracia americana.*
	O importante	*é*	*que a sociedade saiba ...*

	O assunto	*eram*	*as ditaduras*
2.	F	<vK>	EF
	Um dinamarquês	*não é*	*malandro.*
	Cachorros	*não são*	*perigosos.*

		focus	**topic**
		predicative	**subject**
3.	<vK>	AB	AC
	Sou	*eu*	*quem fala.*
	Sou	*eu*	*que o fiz.*
	Sou	*eu*	*o chefe da fazenda.*
	É	*esse*	*o número de alunos.*
	Não foi	*Maria*	*quem bateu no cachorro.*
	São	*três*	*os erros de ...*
4.	<vK>	EF	ACD
	É	*comum*	*fazer isso.*
	Foi	*interessante*	*a palestra.*
	Era	*uma delícia*	*este bolo.*
	Erar	*um problema*	*quem compraria ...*
	Não foi	*um José Sarney*	*quem fez o Plano*
		Cruzado	

	focus	**topic**	
	predicative	**subject**	
5.	(B)EF	(A)B	<vK>
	Casada	*ela*	*nunca foi.*

with the predicative in topic position and the subject as focus

		topic		**focus**
		predicative		**subject**
6.		EF	<vK>	ABCD
		Idiota	*é*	*a sua avó.*
		Jovem	*é*	*quem se senta jovem.*
		Grande	*foi*	*o susto dele, quando soube que ...*
		Agradável	*é*	*passar o domingo na cama.*
7.		E		EF
	Para mim,	*perigosos*	*são*	*cachorros, serpentes, ...*

The examples show that the normal "predicative word classes" are EF, while the "subject word classes" are ABCD. This is unproblematic, as long as different pattern slots are filled in with members of different word class groups. Where, however, ABCD-material meets ABCD-material, or EF-material meets EF-material, as possible in patterns (1), (2) and (3), position is the determining factor. Thus, the position left of the verb in (1) and (2) forces a (topic-) subject reading, and the "middle" position in (3) forces a (focus-) predicative reading.

Since Portuguese allows for omission of the subject (which, in a way, is "incorporated" into the verbal inflexion ending), structures 1-4 may be collapsed into a single pattern: **<vK> focus-predicative.** All nominal material (ABCDEF) can be located in the single remaining slot. Because copula verbs can not ordinarily be intransitive, no ambiguity arises: the verb's only complement is the (focus-) predicative.

However, when collapsing structures 5-7 into **predicative <vK>**, we arrive at an ambiguity as to topic/focus for word classes EF, with the omitted subject's now empty "trace" position being located either left or right of the verb. In the first case, the predicative is to be interpreted as **focus-predicative**, in the second as **topic-predicative**. The first is typical for contrastive negative sentences with (focal) stress on the predicative, the second is used in affirmation sentences with (focal) stress on the verb.

	focus predicative	**topic predicative**	
5'.	*Casada*		*nunca foi.*
6'.		*Idiota*	*é.*
7'	*Para mim,*	*perigosos*	*são.*

In a question, focus is automatically placed on the interrogative word. However, only structure 5 provides for a fronted focus. In 1 and 6 the interrogative topic and focus are reversed, focus filling the topic position of the analogue declarative sentence. Thus, structures 1, 5 and 6 yield the following interrogative patterns:

1".	**focus-subject**		**topic-predicative**	
	O que		*é*	*novo ?*
5".	**focus-predicative**			**topic-subject**
	O que	*ele*	*é ?*	
6".	**focus-predicative**			**topic-subject**

O que *é* *isso ?*

In my parser, which is primarily syntactic in its functional annotation, I have chosen to tag nominal structures for subject (@SUBJ>, @<SUBJ) and subject predicative (@SC>, @<SC), not for topic and focus. These can, however, be derived from predicative position and word class material in most cases, as shown above, with the relatively rare subject-less predicative-fronted affirmation sentences being the only exceptions.

There are, however, cases where focusing involves non-nominal material in structures roughly analogous to cases (3) and (6). Compare the following:

3a *Sou* *eu* *que o fiz.*
3b *Era* *da Maria* *que ele gostava.*
3c *Era* *da Maria* *que ele falou.*

6a *Bom* *é* *que termina bem.*
6b *Aqui* *é* *que você deve ficar*.
6c *Aqui* *é* *que ele quer construir uma casa*.

Whereas cases 3a and 6a can be described in terms of absolute relative subject clauses[161], this is not so easy in the other cases: for 6b/c it would involve @SC adverbs (usually seen as @ADV in conjunction with 'estar' rather than 'ser') and assigning "que" word class status of relative adverbial which it traditionally doesn't have (though there are "slang"-cases like *"no ano que nasceu"*, and the problem of comparative 'que' in 4.4.3., exx. (3) and 4.5.2), - and for 3b/c even that measure would fail, since the PP in question *(da Maria)* cannot be replaced by simplex adverbs (of place, time and manner - like *onde*, *quando*, *como*). For valency reasons, a relative construction *must* contain the preposition *(Era Maria* ***de*** *quem gostava),* leaving us with a 3a-type sentence.

Therefore, it is logical to separate 3b/c and 6b/c from the predicative discussion, and I want to argue that they can be described as cleft structures, where the adverb or PP in focus has a constituent link to the clause after the *que*. The verb *gostar* (which has a <de^vp> valency) in 3b lacks a @PIV-constituent, and *ficar* in 6b lacks an @ADV-constituent to satisfy its <va> valency. Having shown cleft constructions for the obligatory constituents in 3b/6b, I can then use the same description for 3c/6c, with facultative @ADVL-constituents.

Without introducing new PoS classes (like FOC-SER and FOC-QUE), the most adequate word class descriptions are KS for *que* and V VFIN for *é/era*, which leaves the

[161] Though *que* cannot always be replaced by *o que/quem* (as onewould expect for true absolute relative clauses):
Somos nós que/*quem o queremos.

focusing function to the *syntactic* tags. *É/era* already needs a special @FOC> tag in "slang" sentences like

(8) *ele trabalha é com grande entusiasmo*. [he works IS with great enthusiasm].

Using the @FOC tag[162] in the above b/c-cases, and referring the *que* back to the focus constituent, we get:

(9)

Era	*da*	*Maria*	*que*	*gostava*.
V VFIN	PRP	PROP	KS	V VFIN
@FOC>	@PIV>	@P<	@FOC<	@MV

Aqui	*é*	*que*	*você*	*deve*	*ficar*.
ADV	V VFIN	KS	PERS	V VFIN	V INF
@ADV>	@<FOC	@FOC<	@SUBJ>	@FAUX	@#ICL-AUX<

Many CG-rules are based on clause-internal tag uniqueness. It is bad enough that morphological disambiguation thus has to cope with two VFIN in the same sentence, but this is somewhat remedied by the intervening *que* being allowed to retain its "isolating" KS-tag. And at least on the syntactic level, this way, there is only one @MV, and no @#FS-tag for the *que*, which could cause rule context problems with a missing valency bound constituent in the resulting @#FS-subclause.

[162] The @FOC tag is experimental and has not yet been introduced in the internet-version of the parser, which therefore offers the alternative predicative analysis in the cleavage-focus cases (9) and two adjacent @FMV tags in the slang-sentence case (8).

4.5.2 Comparison structures

@#....-KOMP< argument of comparative, e.g. "do que" referring to 'melhor' - “better *than*”
(always clausal: @#AS-KOMP< or @#FS-KOMP<)
@COM comparative subordinator (direct comparator), e.g. “[work] *like* a slave”
@PRD predicative subordinator (role predicator), e.g. "[work] *as* instructor"

Portuguese comparative structures are traditionally divided into three types of comparatives, of equality *(tão ... como),* superiority *(mais ... que)* and inferiority *(menos ... que)*, as well as two types of relative superlative, of superiority *(o mais ... de)* and inferiority *(o menos ... de).* Syntactically, the connection between *the comparative kernel* and the *comparandum* is established by means of relational particles: - relative adverbs *(como, segundo, conforme, quanto, quão)* or relative determiners *(quanto, qual),* the subordinating conjunctions *que* and *do=que*, and the preposition *de*. The relative particles are used for *equalitative* comparisons (tagged <igual>), while *que*, *do=que* and *de* cover both superiority- and inferiority- comparisons, permitting disambiguation by the same CG rules, which is why I have chosen a special term, *correlative* (tagged <corr>), as an umbrella term for these two constructions.

As shown in the table below, a comparandum headed by a relative or conjunction can take the form of either a finite subclause (FS) or an averbal small clause (AS), whereas *de*, obviously, heads a PP. Because they are loosely related - in a syntactic[163] way - to "real" comparisons, consecutive *(tão ... que)* and the so called conformative *(conforme, segundo)* subclauses have been added to both the table below (*) and the examples later in the chapter.

	COMPARATIVE KERNEL			COMPARANDUM		
	HOOK	BASE		HEADER	BODY	
O rei parece	mais/menos	velho	→	que	a rainha.	*correlative AS*
			↘	do=que	a rainha diz.	*correlative FS*
-	o mais/menos	velho	→	de	todos.	*correlative superlative PP*
-	tão	velho	→	como	a rainha (diz)	*equalitative AS (FS)*
-			↘	que	dorme muito.	*consecutive FS**
-	-	velho		como	a rainha.	*direct comparison AS*
-	-	-	.......,	conforme	dizem/ele.	*direct relativisation FS (AS)**

With the exception of the relative adverbs - that can perform direct comparisons (shaded) - all these comparandum header particles need a premodifying "hook" at the comparative kernel, to which they are dependency-linked. As correlative hooks function

[163] Consecutive constructions involve a comparative hook *(tão)*, and conformative subclauses use a comparative complementizer *(como, conforme, segundo)* , mimicking hooked comparisons in the first case, and direct comparisons in the second.

the quantifying adverbs *mais* and *menos* which denote the comparative degree of Portuguese adjectives and adverbs, and as equalitative hooks the adverbs *tão, tanto* and the determiners *tanto, tal*. There are restrictions as to which hook can be combined with which relational particle, for instance *mais/menos - que/do=que, tal - qual, tanto - quanto, tão - como*.

Like relative adverbs, also the conjunction *que* can, in certain contexts, head a direct comparandum, as in (i) and (ii). Though traditionally classified as a comparative subordinating conjunction, this type of *que* could, in fact, be classified as a relative adverb itself, and there are other constructions deserving this analysis, with adverbial "hooks" for the relative n-comparative) link, cp. (iii) and (iv).

(i) *forte que nem um urso* ('strong as not even a bear').
(ii) *Bom que seja o rapaz não é nenhum santo* ('Good as he might be, the boy is no saint')
(iii) *no ano que nasci* ('in the year when I was born)
(iv) *Ainda que* ('though')
(v) *Desliga, amor, que tem gente na linha* ('Hang up, dear, as there's somebody on the line')

Etymologically, such a word class distinction would make sense, since *que* in these examples is derived from Latin 'quam' [how], itself a relative adverbial, while "ordinary" completive, NP-producing, *que* is rooted in a Latin pronoun, 'quod' [what]. Like the relative adverb *como,* adverbial *que* occurs in causal constructions, too (v), with yet another Latin relative adverb etymology, 'qua' [where][164].

When an equalitative hook *(tão, tanto, tal)* is combined with the conjunction *que* as relational particle introducing a finite subclause, this subclause will have a consecutive meaning, sometimes - when used with the subjunctive mood in the subclause - implying finality. This construction, though not a comparison as such, does measure the "degree" of a predication or modifier, affecting the mental image of the comparative base in much the same way a real comparandum does. The fact that *que* here can be replaced by *de tal maneira/modo que* ('such as to'), with *que* relating back to an adverbial PP expression, again suggests a relative adverb reading.

Expressions like ..., *como dizem* ('according to what they say'), *como já disse* ('like I said') or ..., *conforme vi* ('as [far as] I have seen'), involve what I call relative adverbs, too, and are vaguely related too (hook-less) comparisons, though the relativisation expressed, relates to a statement (and most likely its truth-value), not to a measure.

[164] A last kind of adverbial *que* is the insensifier, as in *que lindo!* ('How beautiful!'), going back to Latin 'quid'.

In my parser, comparative hooks are tagged <KOMP>, and the comparandum header particles <komp>. Correlatives receive <corr> and equalitatives <igual>. Though originally secondary tags, all these tags are now disambiguated by CG-rules, mainly for semantic reasons. Thus, the conjunction *que* will be treated (a) as 'that' without the <komp> tag, (b) as 'than' if left with both the <komp> and <corr> tags, and (c) as 'such that' (consecutive) with a <komp> tag only.

The functional difference between direct and hooked comparisons is shown by marking the latter with the external (@#AS/FS) tag KOMP< - for their dependency relation to a <KOMP> tagged hook, whereas the former (direct comparisons) receive external dependency tags linking them to nominals (N<), post-adjects (A<) or main verbs (ADVL, ADVL>, <ADVL). The "comparison agreement" CG-rules, that for the comparandum header select those <komp>, <corr> or <igual> tags that match its hook (correlative <komp><corr> in (2a-d) and equalitative <komp><igual> in (2g-k)), will remove all three tags in the case of a direct comparison (cp. ex. (2l) and (2m); instead, the <prp> ("prepositional") tag is chosen for averbal clause (@#AS) comparandum headers (cp. ex. (2n)) and the <ks> ("conjunctional") tag for finite subclauses (@#FS) comparandum headers (cp. ex. (2o)).

As to clause internal function (for the @#AS and @#FS comparandum cases), the umbrella function "complementiser" is differentiated in the following way:

1. pure comparator, @COM

Ela é **mais** *<KOMP><corr> ADV @>A jovem*
do=que *<komp><corr> KS* **@COM** *@#AS-KOMP< a amiga.*

@COM is used for correlative structures (where the comparandum header is a conjunction without argument or adjunct function), for equalitative small clause structures (where the "adverbiality" of *como* and *quanto* as well as the "determinerhood"[165] of *qual* and *quanto* is reduced to a minimum and can not be recovered without unfolding the AS into a hypothetical deep structure clause) and for direct comparisons (where, without a hook and an external KOMP< tag to share the task, the complementiser's internal tag is the only element marking the comparison)

2. argument/adjunct comparator, @ADVL>, @SUBJ>, @ACC>, @SC>, @ADV>

Ficou **tal** *<dem><KOMP><igual> DET @<SC*

[165] Especially assigning more function to *determiner* comparandum headers is problematic, since the superficial function is much more "preposition-like" than "argument-like". Without a verb in the subclause, an object reading for *quanto* in *'ele comeu tanto quanto eu.'* doesn't make much sense, though this is suggested by the parallel object function of *tanto* in the main clause. It would be easier to assigne ADVL function to relative adverbs in the same position, since this reading has already been introduced for AS structures of the type *when in Rome, ...* However, in the latter case, the adverbial retains a certain predicational quality, usually of time ('when') or location ('where'). This argument and the fact that I favour a homogenous reading for the equalitative AS-structures, have lead to my choosing the "neutral" @COM tag for the whole group in question.

***qual** <rel><komp><igual> DET **@SC>** @#FS-KOMP< antes era.*

Verb complement and adjunct tags are used for equalitative FS-structures (where the verbal valency of the subclause permits and demands clause internal functional differentiation)

3. subordinating conjunction, @SUB

*Comeu **tanta** <KOMP><igual> DET @>N comida*
*   **que** <komp> KS **@SUB** @#FS-<ADVL nada sobrou para a irmã.*

@SUB is used for consecutive structures (where the equalitative hook regains its primary intensifier/quantifier function, and the comparandum header *que* reverts to its function of "pure" subordinator)

4. role predicator, @PRD

*Apontou o amigo **como** <rel> <prp> ADV **@PRD** @#AS-<ADVL como seu advogado.*

This function is quite different from the first three in that it does more than compare, it does not *measure* the comparison base, but *changes* it semantically, having an intensional rather than extensional effect. However, I include @PRD in the list of comparison structures, because, like direct comparisons, it uses the relative adverb *como*, and the ambiguities @COM @#AS-<ADVL vs. @PRD @#AS-<ADVL as well as PCP @COM @#AS-A< vs. PCP @PRD @#AS-A< are semantic rather than syntactic, making them quite hard to resolve[166]. On the other hand, not making the distinction would raise problems at higher levels of parsing, since the difference in semantics necessitates different translations, cf. 'He works *as* a slave' (@PRD) vs. 'He works *like* a slave' (@COM).

In table (1), all the different comparandum head functions are listed together with the words with which they can appear, as well as an English translation highlighting the semantic differences.

(1) **Table: Comparandum header types**

	PP	AS	FS
<komp> <corr>		@COM @#AS-KOMP< mais/menos..que	@COM @#FS-KOMP< mais/menos..do=que 'than'

[166] One structural difference is, that the noun following a @PRD complementizer usually is used without an article, whereas @COM triggers the indefinite article. Sadly, both rules aren't iron cast, and definite articles appear in both cases. Another way of making the distinction is via valency, since the @PRD small clauses are much more likely to appear after a fairly limited number of verbs. In my lexicon, I list the "intransitive" <como^va> (e.g. *trabalhar como*) and the "transitive" <como^vta> (e.g. *propor alg. como*).

		mais/menos..do=que 'than'	
<komp> <igual>		@COM @#AS-KOMP< tão..como/quanto, tanto..como/quanto, tal..qual DET, tão..quão 'as..as'	@ADVL> @#FS-KOMP< tão..como/quanto, tanto..como/quanto, tão..quão 'as..as' @SUBJ> @#FS-KOMP< tanto..quanto DET 'as..as' @ACC> @#FS-KOMP< tanto..quanto DET 'as..as' @SC> @#FS-KOMP< tanto..quanto, tal..qual DET 'as..as' @ADV> @#FS-KOMP< tanto..quanto ADV 'as..as'
<prp> (AS) <ks> (FS)		@COM @#AS-(<)ADVL(>) como, segundo, conforme, assim=como, qual ADV 'like' @PRD @#AS-(<)ADVL(>) como 'as'	@COM @#FS-(<)ADVL(>) como, segundo, conforme, assim=como, qual ADV 'like'
		@COM @#AS-N< como 'like' @PRD @#AS-N< como 'as'	
		@COM @#AS-A< como 'like' @PRD @#AS-A< como 'as'	
<komp>	@KOMP< mais/menos..de 'more..than' o/os/a/as +mais/menos..de 'most..of' (superlative)		@SUB @#FS-KOMP< tão/tanto ..que ('such that', consecutive)

The comparative degree of Portuguese adjectives and adverbs is synthetic (2a) only for a small closed class of lexemes *(grande - maior*, *pequeno - menor*, *bom/bem - melhor*, *mal - pior)*, and only exists for the superiority correlation; in all other cases, the comparative is analytical (2b), using the intensifiers *mais* ('more') for superiority, *menos* ('less') for inferiority and *tão*, *tanto*, *tal* ('as') for equality. In my grammar, the first two appear in so called correlative structures, the others in equalitative structures.

(2a)

Mas	[mas] KC @CO ‘but’
acostumou-	[acostumar] <hyfen> <a^xrp> <vH> V PS 3S IND VFIN @FAUX ‘[he] got used’
se	[se] <refl> PERS M/F 3S/P ACC/DAT @<ACC '-'
a	[a] PRP @PRT-AUX< ‘to’
analisar	[analisar] <vt> <vH> V INF 0/1/3S @IMV @#ICL-AUX< 'analysing'
o	[o] <art> DET M S @>N ‘the’
país	[país] <inst> <HH> N M S @<ACC ‘country’
com	[com] PRP @<ADVL ‘with’
um	[um] <quant2> <arti> DET M S @>N ‘a’
distanciamento	[distanciamento] N M S @P< ‘distancing’
maior	[grande] **<KOMP> <corr>** ADJ M/F S @N< **‘greater’**
que	**[que] <komp> <corr> KS @COM @#AS-KOMP< ‘than’**

o [o] DET M S @AS< ‘that’
de [de] <sam-> <+hum> PRP @N< ‘of’
os [o] <-sam> <art> DET M P @>N ‘-’
demais [demais] DET M/F S/P @>N ‘other’
brasileiros [brasileiro] <N> N M P @P< 'Brazilians'

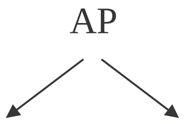
(2b)
ele [ele] PERS M 3S NOM/PIV @SUBJ> ‘he’
é [ser] <vK> <sN> V PR 3S IND VFIN @FMV ‘is’
mais [muito] **<quant> <KOMP> <corr>** ADV @>A ‘more’
bonito [bonito] <jn> ADJ M S @<SC 'beautiful’
do=que [do=que] <komp> <corr> KS @COM @#AS-KOMP< ‘than’
capaz [capaz] <h> ADJ M/F S @AS< ‘skilled’

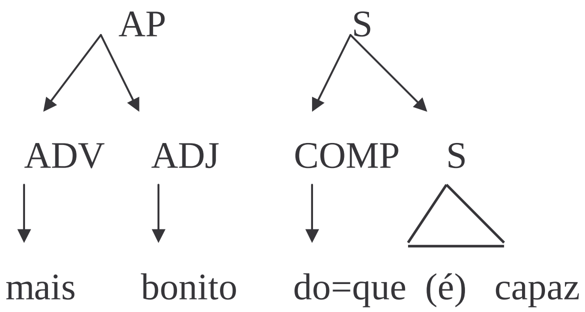
Both (2a) and (2b) are correlative structures, and the comparandum is in both cases a small clause. The difference between synthetic and analytical comparative may be treated as merely morphological, yet it spawns a difference in syntactic complexity, since comparative and comparandum are adjacent in (2a), but isolated by the head of the comparative kernel in (2b). The syntactic break becomes more palpable where there is no synthetic parallel - as when the comparative kernel is not an adjective or adverb, but a noun, and the comparator hook not an intensifier adverb but a determiner. Even adjuncts (2c) or relative subclauses can interfere.

(2c)
Tinha [ter] <vt> <ink> <rH> V IMPF 1/3S IND VFIN @FMV ‘[he] had’
menos [pouco] <quant3> **<KOMP> <corr>** DET M/F S/P @>N ‘less’
dinheiro[dinheiro] <cm> N M S @<ACC ‘money’
para [para] <+INF> PRP @<ADVL 'for'
gastar [gastar] <vi> <rH> V INF 0/1/3S @IMV @#ICL-P< ‘spending’
do=que [do=que] <komp> <corr> KS @COM @#AS-KOMP< ‘than’
o [o] <art> DET M S @>N ‘-’
seu [seu] <poss 3S/P> DET M S @>N ‘his’
irmão [irmão] <fam> N M S @AS< 'brother'

The flat dependency notation of the parser retains the link between comparative hook and comparandum whatever the distance, using the dependency marker KOMP<, pointing leftward (back) to the <KOMP> tag at the hook. Due to the disjunct constituent 'menos .. do=que', tree structures as in traditional generative grammar yield much more awkward analyses. For instance, (2b) might be tree-analysed in the following way:

(2b')

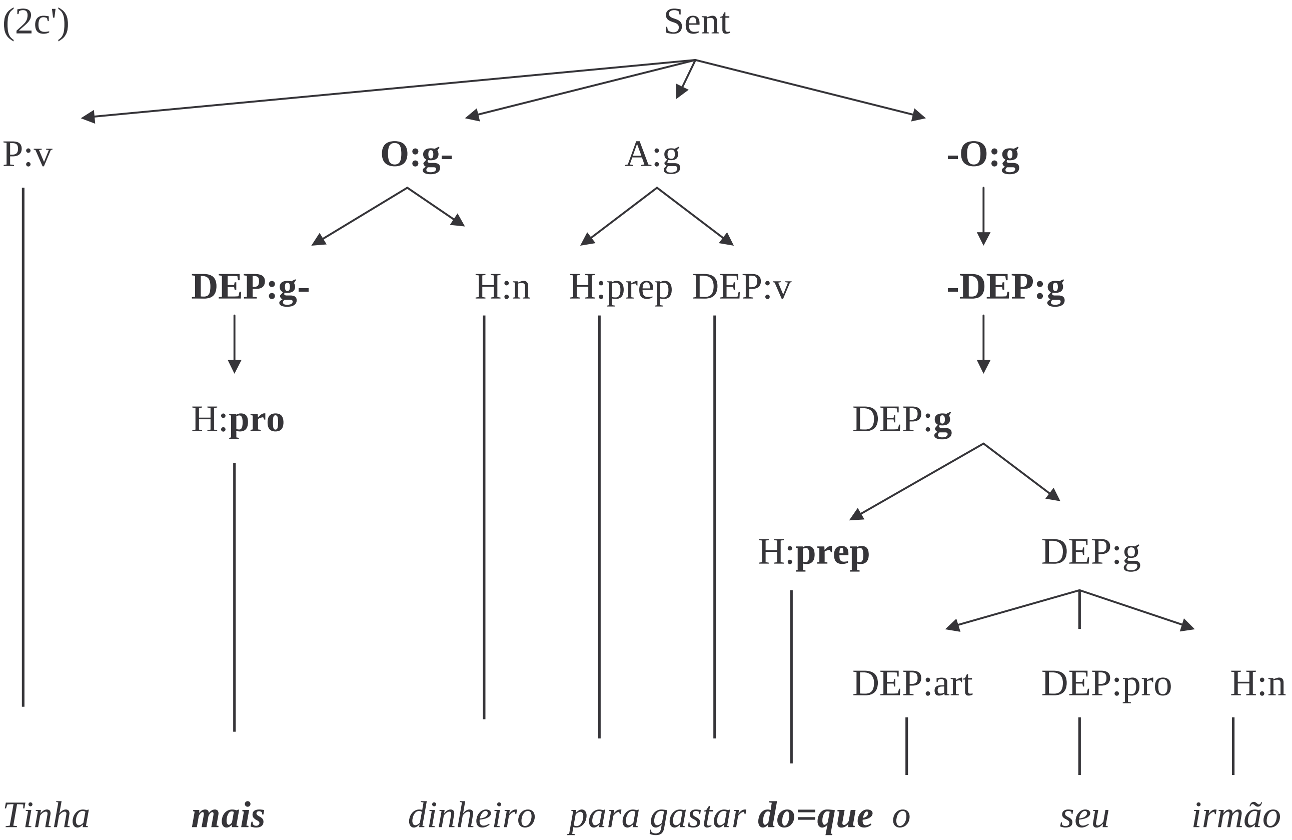




In this analysis, in order to avoid a disjunct constituent, the semantic dependency link between 'mais' and 'do=que' is ignored, and the comparandum attaches to the AP 'mais bonito' as a whole, as would be the case in a synthetic comparative. Also, the comparandum is analysed as a (deep structure) subclause rather than the small clause analysis used in my parser, which works without hypothetical or zero constituents. Besides being unnecessarily complicated, I find it illogical in one aspect to *add* information (the zero/hypothetical predicator) while at the same time *removing* dependency information that is lexically present in the surface string (the pairing of *mais* - *do=que, tal* - *qual, tão* - *como*).

Alternatively, one might introduce disjunct constituents, as in the Odense English syntax model (Bache et. al., 1993), and analyse the comparandum as a (comparative) group rather than a clause:

(2c')



[P=predicator, O=object, A=adverbial, H=head, DEP=dependent, Sent=sentence, v=verb, n=noun, pro=pronoun, prep=preposition, art=article, cl=clause, g=group]

In this analysis, in order to match surface linearity while at the same time preserving dependency relations, two disjunct constituents are necessary, i.e. the direct object, and - beneath it - the dependent group that contains the comparative hook and its dependent argument, the comparandum group. Note that in order to avoid a clausal analysis and the ensuing zero constituent, the AS is interpreted as a group, and the complementiser *do=que* as a preposition.

Seemingly, either comparative dependency (2b') links or constituent continuity (2c') must be sacrificed in a tree structure notation, compromising either information content or - since constituents don't *look* very much like constituents when disjunct - descriptive elegance, while a CG-style flat dependency notation captures all of the information *without* awkward modifications to its core conventions[167].

In Portuguese, the superlative is not part of the degree paradigm of an adjective in the same way as in English (synthetically-morphologically '-er' - '-est', analytically 'more' - 'most'). Rather, synthetic superlatives only appear as absolute forms (i.e. without comparandum), and are built by derivation with the suffix '-íssimo' (with a few irregular exceptions), and analytical superlatives - that do allow comparandum dependency - are formed by adding the definite article in front of the comparative: *o maior* ('the biggest'), *o mais rápido* ('the fastest').

(2d)

Estamos	[estar] <x+GER> <ink> V PR 1P IND VFIN @FAUX ‘[we] are -ing’
nos	[nos] PERS M/F 1P ACC/DAT @ACC> '-'
tornando	[tornar] <vrK> <rH> V GER @IMV @#ICL-AUX< ‘turning [into]’
o	[o] <art> DET M S @>N ‘the’
mais	[muito] <quant> **<KOMP> <corr>** ADV @>A ‘most’
pobre	[pobre] <h> ADJ M/F S @<OC ‘poor’
de	**[de] <sam-> <komp> PRP @KOMP< 'of'**
os	[o] <-sam> <art> DET M P @>N ‘the’
países	[país] N M P @P< ‘countries’
urbanos	[urbano] ADJ M P @N< 'urban'
industriais	[industrial] <n> ADJ M/F P @N< 'industrialised'

In the example sentence (2d), the superlative *o mais pobre* is complemented by a comparandum, - the PP introduced by *de*. For superlative structures, no other comparandum header is possible. Note that the comparative hook is still the intensifier

[167] Of course, since the difference between (2c) and (2c') is mainly notational, any comparison should take non-linguistic arguments into account as well, like the pedagogical value of a model within a given teaching frame work (cp. chapter 7.2 for a pedagogical discussion of CG).

mais, which by itself does not denote a superlative. The superlative reading resides in the combination of definite article + comparative. One might say that *o* really is an intensifier adject (@>A) of *mais*, and not an article at all (since it modified the comparativeness of *mais* into "superlativeness"), but I have chosen to follow traditional morphology in this case, and tag *o* as prenominal dependent @>N (i.e. retaining the article/determiner reading).

As shown in table (1), the comparandum in (non-superlative) correlative structures can also be finite subclauses (FS):

(2e1)

é	[ser] <vK> <ink> V PR 3S IND VFIN @FMV ‘[she] is’
mais	[muito] <quant> **<KOMP> <corr>** ADV @>A ‘more’
bonita	[bonito] <jh> ADJ F S @<SC ‘beautiful’
do=que	**[do=que] <komp> <corr> KS @COM @#FS-KOMP< ‘than’**
dizem	[dizer] <v-cog> <ink> <vH> V PR 3P IND VFIN @FMV ‘[they] say’

the syntagmatic idiom *'por mais ADJ que SUBJUNCTIVE'* can be analysed in the same way (2e2). Since two of the words in the expression aren't fixed lexically, the pattern can not be entered into the lexicon as a whole[168], and an analytic approach is mandatory.

(2e2)

por	[por] PRP @ADVL> ‘as’
mais	[muito] <quant> <KOMP> <corr> ADV @>A ‘-’
contraditório	[contraditório] <n> ADJ M S @P< ‘contradictory’
que	[que] <komp> <corr> KS @COM @#FS-KOMP< ‘as’
pareça	[parecer] V PR 1/3S SUBJ VFIN @FMV ‘[it] might seem’

Sometimes, the whole of a comparandum is in premodifier position (2f1), syntactically isolating the pertaining head. A very common case is *mais/menos de + NUM* .

(2f1)

Nazaré	[Nazaré] <HEUR> PROP M/F S/P @SUBJ> 'Nazareth'
era	[ser] <vK> <ink> V IMPF 1/3S IND VFIN @FMV ‘was’
uma	[um] <quant2> <arti> DET F S @>N ‘a’
vila	[vila] <by> N F S @<SC ‘village’
de	[de] <+hum> PRP @N< 'of’
não	[não] <dei> <setop> ADV @>A ‘not’

[168] Lexicon entries are the preferred solution for "frozen polylexicals", like complex prepositions (e.g., *em=vez=de* 'instead=of'). This way, the structure can be recognized as a whole already at the preprocessor level, and assigned *one* (integrated) word class and syntactic function. Somewhat more difficult is the case of "lexical idioms" (complex nouns and incorporating verbs), where compound-internal inflexion (at the first element in a complex noun, or at the verb in incorporations) complicates the situation. Such polylexicals are entered into the lexicon, but the preprocessor alone can't recognize them without asking the tagger for help - i.e. for morphological/inflexion analysis of the potential parts of such polylexicals. The above is one example of level interaction in the parsing system.

mais=de	**[mais=de] <quant> ADV @>A ‘more than’**
$2.000	[2.000] <cif> <card> NUM M/F P @>N '2.000'
habitantes	[habitante] <N> N M/F P @P< ‘inhabitants’

Even the whole of the comparative structure, including its nominal base, may function as a premodifier, making the syntactic break even more gross (2f2). One solution is to give *mais/menos* a nominal reading and assign it the functional tag that would otherwise belong to its head (here: @P<). In the case of (2f1), this yields the reading *uma @>N vila @NPHR de @N< não @>A* ***mais DET @P<*** *de @KOMP< 2.000 @>N habitantes @P<*. Though this analysis is semantically unsatisfying, it would be quite hard to make the comparative structure "transparent" for those CG-rules that would have to link *habitantes* as preposition-argument to the postnominal *de* 5 words and several constituent borders to the left. Therefore, in order to improve syntactic transparency, an "inter-processor" - situated between the morphological and the syntactic module - recognises *mais/menos de + NUM* strings and replaces them by *mais=de (menos=de) <quant> ADV + NUM*, allowing for a simple adverbial pre-adject tag (@>A) for *mais=de (menos=de).*

(2f2)

mais=de	**[mais=de] <quant> ADV @>A ‘more than’**
dez	[dez] <card> NUM M/F P @>N ‘ten’
dias	[dia] <dur> <per> <num+> N M P @>A ‘days’
depois	[depois] ADV @ADVL ‘afterwards’

Also in equalitative comparison structures both adjects (2g) and NPs (2h) can be modified, the first typically by the intensifier adverb *tão..como/quanto*, the second by the quantitative determiners *tanto..quanto* or *tal..qual*. A special ambiguity problem arises from the fact that both *tanto, quanto* and *qual* also have an ADV reading and can function as intensifiers, too.

(2g)

para	[para] <+INF> PRP @<ADVL ‘in order to’
criar	[criar] <em^vtp> <sH> V INF 0/1/3S @IMV @#ICL-P< ‘create’
confusão	[confusão] <am> N F S @<ACC ‘confusion’
tanto	[tanto] <quant> **<KOMP> <igual>** ADV @<ADVL ‘as much’
em	[em] <sam-> PRP @<PIV 'in'
o	[o] <-sam> <art> DET M S @>N ‘the’
julgamento	[julgamento] <+por> N M S @P< ‘trial’
por	[por] PRP @N< ‘for’
crime	[crime] N M S @P< ‘crime’
de	[de] PRP @N< 'of'
responsabilidade	[responsabilidade] <am> N F S @P< ‘responsibility’
como	**[como] <rel> <komp> <igual> ADV @COM @#AS-KOMP< ‘as’**
em	[em] <sam-> PRP @AS< 'in'

o	[o] <-sam> <art> DET M S @>N ‘the’
processo	[processo] <+por> N M S @P< ‘law suit’
por	[por] PRP @N< ‘for’
delito	[delito] N M S @P< ‘crime’
comum	[comum] <jn> ADJ M/F S @N< ‘common’

(2h)

comeu	[comer] <vt> <vH> <ink> V PS 3S IND VFIN @FMV ‘[she] ate’
tanto	[tanto] <quant2> **<KOMP> <igual>** DET M S @<ACC ‘as much’
quanto	**[quanto] <rel> <komp> <igual> DET M S @COM @#AS-KOMP< ‘as’**
ele	[ele] PERS M 3S NOM/PIV @AS< ‘he [did]’

(2g) and (2h) feature AS-comparanda, but of course, both adjects and NPs can also be modified by comparison structures involving FS-comparanda:

(2i)

Não	[não] <dei> <setop> ADV @ADVL> 'not'
sou	[ser] <vK> V PR 1S IND VFIN @FMV ‘[I] am’
tão	[tão] <dem> <quant> **<KOMP> <igual>** ADV @>A ‘as’
perfeito	[perfazer] <ADJ> V PCP M S @<SC ‘perfect’
quanto	[quanto] <rel>**<komp><igual><quant> ADV @ADVL> @#FS-KOMP<** ‘as’
dizem	[dizer] <v-cog> <vH> V PR 3P IND VFIN @FMV ‘[they] say’

(2j)

Ainda	[ainda] <setop> ADV @ADVL> ‘still, furthermore’
se	[se] <refl> PERS M/F 3S/P ACC/DAT @ACC>-PASS ‘it’
discutia	[discutir] <vt> <vH> <ink> V IMPF 1/3S IND VFIN @FMV ‘is discussed’
se	[se] KS @SUB @#FS-<ADVL ‘if’
o	[o] <art> DET M S @>N ‘the’
projeto	[projeto] N M S @SUBJ> 'project'
aprovado	[aprovar] <vH> <ADJ> V PCP M S @N< ‘approved’
ia	[ir] <x> V IMPF 1/3S IND VFIN @FAUX ‘was going to’
colocar	[colocar] <vt> <vH> V INF 0/1/3S @IMV @#ICL-AUX< 'place'
tanto	[tanto] <quant2> **<KOMP> <igual>** DET M S @>N ‘as much’
dinheiro	[dinheiro] <cm> N M S @<ACC ‘money’
quanto	**[quanto] <rel> <komp> <igual> DET M S @ACC> @#FS-KOMP< ‘as’**
se	[se] <refl> PERS M/F 3S/P ACC/DAT @SUBJ> ‘they’
previa	[prever] <vt> <vH> V IMPF 1/3S IND VFIN @FMV ‘foresaw’

Since the comparandum in (2i) and (2j) is a real clause, its header must be assigned some clause internal argument or adjunct function. In the adverb case, this may be either adjunct adverbial (@ADVL>) or adverbial object (@ACC>), and the (relative) determiner may function as subject (@SUBJ>), direct object (@ACC>) - as in (2j) -, or subject complement (@<SC) - as in (2k).

(2k)

Ficou	[ficar] <vK> <v-cog> <ink> V PS 3S IND VFIN @FMV ‘[it] became’
tal	[tal] <dem> **<KOMP> <igual>** DET M/F S @<SC ‘[such]’
qual	**[qual] <rel> <komp> <igual> DET M/F S @SC> @#FS-KOMP< ‘as’**
antes	[antes] ADV @ADVL> ‘before’
era	[ser] V IMPF 1/3S IND VFIN @FMV ‘[it] was’

A different class of comparisons are direct comparisons, without a hook. Here, the external function will not be @#..KOMP<, but - for ordinary comparisons - that of postmodifier (@#.N<, cf. (2l) and @#..A<, cf. (2m)), or - for the loosely related, statement modifying conformatives - adjunct adverbial (@#..ADVL>, cf. (2n), and @#..<ADVL, cf. (2o)). Semantically, (2l) and (2m) resemble the equalitatives above, and the typical header, the relative adverb *como*, belongs to the group of equalitative headers. Still, the difference becomes clear in bilingual contrasting - in English, for instance, the direct comparison *como* translates as 'like', the hooked *como* as 'as'.

(2l)

tomou	[tomar] <vt> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] established’
contato	[contato] <am> <+com> N M S @<ACC ‘contact’
com	[com] PRP @N< ‘with’
vários	[vários] <quant2> <quant3> DET M P @>N ‘several’
casos	[caso] <ac> N M P @P< ‘cases’
difíceis	[difícil] <n> ADJ M/F P @N< ‘difficult’
como	**[como] <rel> <prp> ADV @COM @#AS-N< ‘like’**
o	[o] DET M S @AS< ‘that’
de	[de] <+hum> PRP @N< ‘of’
Collor	[Collor] PROP M S @P< ‘Collor’

(2m)

ambicioso	[ambicioso] <h> ADJ M S @<PRED 'ambitious'
como	**[como] <rel> <prp> ADV @COM @#AS-A< ‘like’**
um	[um] <quant2> <arti> DET M S @>N ‘a’
César	[César] PROP M S @AS< 'Cæsar'
de	[de] <sam-> <+top> PRP @N< ‘from’
o	[o] <-sam> <art> DET M S @>N ‘the’
sertão	[sertão] <top> N M S @P< ‘wilderness’
de	[de] <sam-> <+hum> PRP @N< ‘of’
o	[o] <-sam> <art> DET M S @>N ‘-’
Arkansas	[Arkansas] <HEUR> PROP M/F S/P @P< 'Arkansas'

Besides *como*, a group of other relative adverbs, for instance *conforme*, *segundo* ('according to') or *assim=como* ('as well as', 'like'), is allowed in these structures, where the function of the subclause is conformative[169] . It is not at all easy to assign a word class to these particles, and traditional grammars often use as many word classes as there are syntactic functions. Typically, when heading an AS (2n), they would be called a preposition, when heading a FS (2o), they would be termed conjunctions. I find it most logical, in analogy with the handling of *onde* (place) and *quando* (time) to add *como* (manner) and others, and then map different syntactic functions onto <u>one</u> word class, that of relative adverbial. However, for reasons of both notational flexibility and semantic-translational[170] differentiation, I disambiguate the *secondary* tags <prp> ("prepositional") and <ks> ("conjunctional") for these words, so the tag chain can easily be filtered into the traditional word classes of preposition and conjunction, respectively, while in the first case (<prp>) also filtering clause function labels (@#) into group function labels (@) and replacing the @AS< tag by @P<[171].

(2n)

Segundo	**[segundo] <rel> <prp> ADV @COM @#AS-ADVL> ‘according to’**
sua	[seu] <poss 3S/P> DET F S @>N ‘his’
estimativa	[estimativa] <+de+interr> N F S @AS< ‘estimate’
$,	
a	[a] <art> DET F S @>N ‘the’
política	[política] <pp> N F S @SUBJ> 'policy'
seletiva	[seletivo] <n> ADJ F S @N< ‘selective’
de	[de] PRP @N< ‘of’
desenvolvimento	[desenvolvimento] <cI> <CP> N M S @P< ‘development’
será	[ser] <vK> <sN> V FUT 3S IND VFIN @FMV ‘will be’
capaz	[capaz] <h> <+de> <+de+INF> ADJ M/F S @<SC ‘capable’
de	[de] PRP @A< ‘of’
elevar	[elevar] <vt> <vH> V INF 0/1/3S @IMV @#ICL-P< ‘raising’
o	[o] <art> DET M S @>N ‘the’
produto	[produto] <mon> N M S @<ACC 'BNP'

[169] From an information-theoretical point of view, the main difference between a conformative and other direct comparison subclauses (AS/FS) is that the latter directly replaces an old (preconceived, prototypical) 'how' with a new 'how', while the former adds a meta-'how' (source, interpretation) to the 'how' targeted. One can, however, imply the other, which can make it difficult to draw a clear border-line. Consider the following cline of hook-less comparison/conformative constructions:

(i) A world like an orange
(ii) The world is like/as Ptolemy sees it.
(iii) The world like/as Ptolemy sees it
(iv) The world according to Ptolemy
(v) According to Ptolemy (as Ptolemy sees it), the world is like an orange.

(i) clearly add a 'how' (orange), and (v) clearly adds a meta-'how' (Ptolemy's view) since there already is another 'how' (orange). (iv), however, implies a 'how' (orange) by providing a meta-'how', and (iii) even uses a typical comparandum header ('like/as') to this end. (ii), finally, shows, that what normally should be a conformative, can syntactically fill a typical "primary" how-slot (the subject complement). In my system, the relation between direct comparatives and conformatives can be seen from the common "clause-internal" function tag, @COM, and the word-class lumping as relative adverbs, while the differences are expressed by the "clause-external" function tag (@#), suggesting modifier readings for direct comparisons, and adverbial readings for conformatives.

[170] *Conforme*, for instance, translates 'according to' when <prp>/AS, but 'according to what' when <ks>/FS.

[171] In the case of (2n) the transformation would yield *segundo PRP @ADVL> sua DET @>N estimativa N @P<, ...*

interno	[interno] <n> ADJ M S @N< '-'
bruto	[bruto] <jn> ADJ M S @N< '-'

(2o)

trouxeram	[trazer] <vt> V PS/MQP 3P IND VFIN @FMV '[they] brought'
cerca=de	[cerca=de] <c> ADV @>A ‘about’
$4	[4] <cif> <card> NUM M/F P @>N '4'
bilhões	[bilhão] <num+> N M P @<ACC ‘billion’
de	[de] PRP @N< '-'
dólares	[dólar] N M P @P< 'dollars'
$,	
conforme	**[conforme]<rel><ks>ADV @ADVL> @#FS-<ADVL ‘according to [what]’**
se	[se] <refl> PERS M/F 3S/P ACC/DAT @SUBJ> ‘they’
estima	[estimar] <v-cog> <vr> <vH> V PR 3S IND VFIN @FMV ‘estimate’
$,	

Even *qual*, which mostly occurs as a prenominal or relative *(o=qual)* determiner, can appear as relative adverbial and head a direct comparison (2p).

(2p)

ele	[ele] PERS M 3S NOM/PIV @SUBJ> ‘he’
nada	[nadar] <vi> <vH> V PR 3S IND VFIN @FMV ‘swam’
qual	**[qual] <rel> <Rare> ADV @COM @#AS-<ADVL ‘like’**
um	[um] <quant2> <arti> DET M S @>N ‘a’
peixe	[peixe] <ich> N M S @AS< 'fish'

A very special case of direct comparison are *como se* ('as if') constructions, where an analysis as direct comparative AS strands two complementisers on top of each other, since the AS's body is itself a clause, introduced by the subordinating conjunction *se*.

(2q)

A	[a] <art> DET F S @>N ‘the’
imagem	[imagem] N F S @SUBJ> ‘picture’
foi	[ser] <x+PCP> V PS 3S IND VFIN @FAUX ‘was’
considerada	[considerar] <v-cog> <vtK> V PCP F S @IMV @#ICL-AUX< ‘considered’
um	[um] <quant2> <arti> DET M S @>N ‘an’
ícone	[ícone] N M S @<OC ‘icon’
de	[de] <sam-> <+top> PRP @N< ‘for’
a	[a] <-sam> <art> DET F S @>N ‘the’
abertura	[abertura] <hul> <am> N F S @P< ‘opening’
política	[político] <jn> ADJ F S @N< ‘political’
$,	
como	**[como] <rel> <ks> ADV @COM @#AS-<ADVL ‘as’**
se	**[se] <v+interr+> KS @SUB @#FS-AS< ‘if’**
ela	[ele] PERS F 3S NOM/PIV @SUBJ> ‘it’

devesse	[dever] <x> V IMPF 1/3S SUBJ VFIN @FAUX ‘had to’
começar	[começar] <ve> <vH> V INF 0/1/3S @IMV @#ICL-AUX< ‘start’
logo	[logo] ADV @<ADVL ‘at once’
por	[por] <+top> PRP @<ADVL '-'
ali	[ali] <dei> <top> ADV @ADVL> 'there'

Many grammars solve this puzzle by assigning a label of "complex conjunction" to *como se* as a whole *(*translating as a lexicon entry *como=se* in my system), avoiding the double clausal analysis. However, apart from notational coherence, a computational parser has another, more technical reason to opt for the more analytical analysis - the handling of ambiguity.

(2r1) Como se discute na assembleia, ..   (‘as is discussed in the Council’)
(2r2) Como se discutisse na assembleia, .. (‘as if he were discussing in the Council’)

In (2r2), *se* is a conjunction (heading a subclause functioning as body for a comparandum AS), but in (2r1) *se* is a reflexive personal pronoun (here semantically acting as subject for an adverbial *como*-subclause), a context dependent difference that must be resolved by disambiguation rules and is beyond the scope of a grammar-free preprocessor.[172]

Yet another case of direct comparison is the semantically quite different *role predicator* construction (e.g., 'he works *as* assistant ..'). One might argue that it is not a comparison at all, but since the difference is purely semantic, and *como* is used as header, I prefer to file this pattern together with the comparison group. Only AS-clauses are found in role predications, and for many typical verbs they appear to be valency-triggered in much the same way as adverbial/prepositional objects or subject complements. Both <va>/<vK>-like ("monotransitive" or "subject complementing") and <vta>/<vtK>-like ("transobjective") patterns exist. (2s) is an example of transobjective usage, and *passar como* ['to be considered ...'] is one of the few cases, where the role is predicated of the subject (leaving apart passivisation, of course). However, the border line to free adjunct use (2t), without a valency frame, is very fuzzy, which is why I use the adjunct tag @#AS-<ADVL, rather than the argument tag @#AS-<ADV. Also, there are no clear cases of obligatory role predicator arguments in Portuguese, though for some lexemes, the meaning difference between the role predicator valency frame and the word's other valency patterns justifies a polysemic analysis where a role predicator argument is obligatory *with regard to a certain meaning*. Interestingly, when I checked for role predicator constructions in my corpus, the 5 best candidates for this class (of semantically *obligatory* role predicator arguments) were 5 common verbs, *dar, haver, ter, tomar, tratar*, that at the same time can bind role predicator arguments by means of

[172] In the latest internet version of my parser, synthetical *como se (‘as if’)* is not a lexicon entry, but is reassembled into one “word” *como=se* after disambiguation.

a preposition, *por*, then allowing for a (valency bound) prepositional object reading (@PIV).

(2s)

havia	[haver] <x+PCP> V IMPF 1/3S IND VFIN @FAUX '[he] had'
apontado	[apontar] <vt> <vH> V PCP M S @IMV @#ICL-AUX< ‘appointed’
o	[o] <art> DET M S @>N ‘the’
antigo	[antigo] <ante-attr> ADJ M S @>N ‘old’
procurador=da=República	[procurador=da=República] N M S @<ACC ‘attorney general’
Inocêncio	[Inocêncio] PROP M S @N< 'Innocêncio'
Mártires	[Mártires] PROP M P @N< 'Mártires'
Coelho	[Coelho] PROP M/F S @N< 'Coelho'
como	**[como] <rel> <prp> ADV @PRD @#AS-<ADVL ‘as’**
seu	[seu] <poss 3S/P> DET M S @>N ‘his’
advogado	[advogado] <title> N M S @AS< ‘lawyer’
dativo	[dativo] ADJ M S @N< ‘assigned’

(2t)

foi	[ser] <x+PCP> <ink> V PS 3S IND VFIN @FAUX ‘[he] was’
descongelado	[descongelar] <vt> <vH> V PCP M S @IMV @#ICL-AUX< ‘unfrozen’
como	**[como] <rel> <prp> ADV @PRD @#AS-<ADVL ‘as’**
herói	[herói] N M S @AS< ‘[the] hero’
de	[de] <sam-> <+hum> PRP @N< ‘of’
o	[o] <-sam> <art> DET M S @>N ‘the’
empresariado	[empresariado] <HH> N M S @P< ‘employers’
$.	

Also the *tão..que* pattern is atypical in the comparison camp, the comparandum has to be an FS, and in spite of linking to an equalitative hook, it uses the subordinating conjunction *que* as header. Semantically, the comparative kernel is measured by a deduction or consequence rather than a real (static) comparandum, the result being a consecutive subclause.

(2u)

Foi	[ser] <vK> <ink> V PS 3S IND VFIN @FMV ‘[it] was’
um	[um] <quant2> <arti> DET M S @>N ‘a’
ano	[ano] <dur> N M S @<SC ‘year’
tão	[tão] <dem> <quant> **<KOMP> <igual>** ADV @>A ‘as’
ruim	[ruim] <+para> <jn> ADJ M/F S @N< ‘bad’
para	[para] <+hum> PRP @A< 'for'
os	[o] <art> DET M P @>N ‘the’
videntes	[vidente] <prof> N M/F P @P< ‘clairvoyants’
que	**[que] <komp> KS @SUB @#FS-KOMP< ‘as’**
eles	[ele] PERS M 3P NOM/PIV @SUBJ> ‘they’
estão	[estar] <x+GER> V PR 3P IND VFIN @FAUX ‘are’

temendo	[temer] <vt> <vH> V GER @IMV @#ICL-AUX< ‘fearing’
por	[por] <sam-> PRP @<ADVL 'for'
o	[o] <-sam> <art> DET M S @>N ‘-’
seu	[seu] <poss 3S/P> <si> DET M S @>N ‘their’
futuro	[futuro] <per> N M S @P< ‘future’

4.5.3 Tagging the quantifier 'todo'

The generalising Portuguese quantifier ‘todo’ [all] is both gender and number inflecting (‘todo’, ‘toda’, ‘todos’, todas’), usually in agreement with an NP-head, and does not take premodifiers. According to my morphological word class criteria, it must therefore primarily be categorised as a determiner (DET). ‘Todo’ is related to the invariable (male singular) pronoun quantifier ‘tudo’, which can not be used as a determiner, and therefore is tagged SPEC (“specifier”). Syntactically, however, the situation is much less clear, since the use of ‘todo’ is not restricted to NP-modifier or even nominal function. As a result of its syntactic variability, the word gives rise to a fair amount of semantic ambiguity relevant in a translational context, even when used as a pre- or postnominal modifier. Still, in spite of such lexemic ambiguity, 'todo' allows a distributional definition/disambiguation for most of its uses/meanings. Consider the following modifier cases:

(1a) em **toda** a minha vida ...
(1b) ... deixá-la por lá a vida **toda**.
(1c) ... **todo** o mundo sabendo, menos ele.
(2a) **Toda** célula tem seu conteúdo separado.
(2b) **Todo** mundo tem fotógrafo aqui ...
(3a) **Todas** as manhãs, depois do café, ...
(3b) ... engole as raças **todas** com a mesma graça ...
(3c) Tenho andado muito envolvida com **todos** eles.
(3d) o que eles conseguiram ? Eles **todos** ?

By applying the criteria of position, number and (added) definiteness, four main groups can be distinguished in the examples:

	Plural		Singular	
added definiteness (article, demonstrative, personal pronoun, name)	yes	no	yes	no
Prenominal position	3a, 3c enumerative [all of them]	-	1a, 1c integrative [all of it]	2a enumerative distributive [each of them]
Postnominal position	3b, 3d enumerative [all of them]	-	1b integrative [all of it]	-

As can be seen, plural forms are indifferent to position (pre- and postnominal position alternate freely), and imply added definiteness, usually, but not necessarily, provided by the definite article or a demonstrative, or by a definite expression, that doesn’t permit an article, like in ‘todo Portugal’, ‘todo ele’.

Lack of added definiteness (i.e., mostly cases where 'todo' is followed immediately by its NP-head) implies both singular and prenominal position, which is why I assign it its own classification, enumerative distributive.

The third group comprises singular forms with added definiteness, with free position alternation ('toda a vida' and 'a vida toda' are equivalent terms).

My distributional distinction between "enumerative" and "integrative" can be semantically interpreted as being about "how many" and "how much", respectively, and the tagger's lexicon provides two corresponding tags, <enum> and <integr>, for disambiguation. For 'todo', the distinction is clearly related to the feature of countability, with enumeratives being countable, and integratives not. Rather than say (for reasons of logical semantics) that enumerative distributives ('toda casa' - 'each house') are countable but, as existential quantifiers, do not have a plural form, I prefer to view them as the singular form of plural enumeratives ('todas as casas' - 'all [the] houses') - which otherwise would be countables without a same-lexeme singular. Unlike English, Portuguese - using the same lexeme for both cases - provides morphological evidence for this view.

Integrative 'todo' ('todo o bolo' - 'all of the cake'), on the other hand, when pluralised morphologically ('todos os bolos' - 'all cakes'), is no longer a mass expression ('how much'), and the concept of 'whole cakes' can only be expressed by means of a different lexeme, the adjective *inteiro* ('whole'): 'bolos inteiros'. Thus, what I mean with the term *integrative* ('how much'), is uncountable only for the lexeme 'todo', not for 'inteiro'. Interestingly, 'inteiro' not being polysemous, it is the ambiguous 'todo' that exhibits such strict morphological-distributional limitations, allowing, as a translational bonus, the automatic distinction between 'all', 'each' and 'whole'.

(2b) is a special case. According to my systematics, it should only have an enumerative distributive meaning, but is, in fact, usually employed in alternation with (1c), possibly because of semantic interference, - ‘todo o mundo’ analytically means ‘the whole world’ (<integr>), but is metaphorically used to mean ‘everybody’. Under the semantic <enum> influence of ‘every’, the expression might then have been shortened to ‘todo mundo’, *not* usually meaning ‘every world’ but ‘everybody’. As an exception and a metaphor, ‘todo mundo’ must enter the tagger's lexicon as a fixed expression.

Note also that ‘todo’ is the only Portuguese determiner allowed to modify personal pronouns (3c-d). Since it also is the only modifier allowed *before* the definite article, an explanation may be that personal pronouns replace whole *definite* NPs, and that "added definiteness" is exactly what distinguishes the two meanings of 'todo' seen in connection with personal pronouns - plural enumeratives (3c, 3d) and singular integratives ('todo ele' - 'all of him'), but not distributive enumeratives ('every', 'each').

Like all Portuguese determiners, ‘todo’ can act on its own as an NP (like the SPEC pronoun class), with the whole range of syntactic functions this implies, for example:

(4a) A menor entre **todas**, e a única a não ... (@P<, argument of preposition)
(4b) **Todos** o escutavam com atenção. (@SUBJ>, subject)
(5) se não se apagam de **todo**, pelo menos se esbatem ...(@P<)

Normally, however, this does not happen with the <integr> singular forms, since ‘tudo’ covers these cases. Only in a few fixed expressions ‘todo’ can be substituted for ‘tudo’: ‘ao todo’ [i det hele taget, by and large], ‘de todo’ [totalt, totally], ‘de todo em todo’ [i det hele taget (after all), fuldstændigt (completely)]. (5) is one of the rare real corpus examples.

Still, the ‘todo’ male singular form *does* appear regularly as *head* of an NP (6). These cases cannot directly be compared to (5) or (4), since even when acting as a SPEC-NP (in traditional terminology "independent" or "substantival" pronouns, as in 4 and 5), Portuguese determiners cannot usually take determiner adjuncts themselves. The most appropriate analysis of (6) is therefore the nominal one, with ‘todo’ read a noun heading its own NP (underlined, 6a - ‘as a whole’, 6b 'a whole').

(6a) ..., no <u>seu @>N **todo** @P<</u> pachorrento, ...
(6b) sendo o indivíduo <u>um @>N **todo** @<SC</u> em crescimento contínuo

Syntactically, matters get even more complicated, since not only can ‘todo’ leave its determiner position and replace or head an NP , - it can also leave noun phrase context altogether, assuming adverbial or predicative function:

(7a) Mas uma contradição **toda** @>A especial, que se encontra ...
(7b) A sala do trono era **toda** @>A decorada
(7c) Contemplou com ternura o homem, **todo** @>A músculos, ...
(7d) ... o estafeta deu boa noite, encolhendo-se **todo** @<ADVL sob a chuvinha e disse:
(7e) A responsabilidade será **toda** @<ADVL minha
(7f) ... e que é **todo** @<ADVL um problema político, econômico ...
(8a) ..., quase **todos** @<PRED indiferentes ao bonito, ...
(8b) Correm **todos** @<PRED para a direita, menos João ...
(8c) 17,4% - **todos** @<PRED de sexo feminino - são ...
(8d) Estamos **todos** @<PRED acuados.

(8e) Estamos **todos** @<PRED morrendo.

(8f) ... tipos de exploração a que **todos** @PRED> estamos mais ou menos sujeitos

(8g) ... moram quase **todas** @<PRED em apartamentos

In (7), ‘todo’ functions as a kind of adverb (meaning ‘completely’ in English[173]), modifying an attributive expression (@>A) or even a verb (7d-f @<ADVL) notwithstanding its agreement obligations. In fact, many grammars or dictionaries provide for a part of speech reading of ‘todo’ as an “adverb”.

In (8), ‘todo’ plays the role of free predicative (tagged as @<PRED or @PRED>), referring to the subject, or - more seldom - to the last preceding NP.

Distributionally, almost all adverbial readings occur in the singular forms, almost all predicative readings in the plural forms. The hybrid situation of a morphologically inflected determiner occurring adverbially, can be solved either by tagging ‘todo’ and ‘toda’ as ADV in these cases, or simply by expressing one aspect as form (DET) and the other as function (@>A), respectively. I have chosen the first solution for other determiners, like ‘muito’ and ‘pouco’ - where there is no inflexion retained in the adverbial cases -, but this choice is more problematic for the inflecting 'todo'. In order to achieve early (i.e. word class level) disambiguation, which then facilitates syntactic CG mapping rules, I have opted for a compromise, tagging the singular cases with adverbial function (7) as "<det> ADV" (marking the morphological word class at least with a secondary tag, <det>), and the plural cases with predicative function (8) as "enum> DET".

Even the predicative cases are functionally quite unique (for a determiner): - Though also certain other determiners (possessives) can predicate the subject, like ‘minha’ in (7e), they do so as @SC (subject complement), and can not be added as @PRED (free predicators), especially not *in the presence of* another @SC (e.g., 8d) or @PRED (e.g., 8a). The licence to appear as free predicators may still, like the singular @>A function, reside in the "adverbial potential" of 'todo', but it is another kind of "adverbiality", - reminiscent of adjectives that in English would be called subject adjuncts ('he stood tall against the sky'). In adjective-inflecting Portuguese, some adjectives like 'alto' [high, tall] and 'baixo' [low], can exhibit all degrees of adjectivity (inflected, (i)) or adverbiality (uninflected, (iii)):

(i) casas **altas** ('high houses')

(ii) os irmãos crescem **altos** ('the brothers grow tall')

(iii) os pássaros voam **alto** ('the birds fly high [in the air]')

[173] The Danish translation, ‘helt’, has the same root etymology as in Portuguese, adding the morpheme ‘-t’ for adverb inflexion.

4.5.4 Adverbial function

4.5.4.1 Argument adverbials

In this text, I understand the term *adverbial* in a functional way, and will reserve it for clause level constituents, which may either be valency bound by the main verb (argument adverbials, @ADV, 1b/c) or not (adjunct adverbials, @ADVL, 1a). The first group may then be subdivided into obligatory (1c) or optional (1b) arguments.

(1a) Ele **falou** em Londres. (Ele fala. O que aconteceu/fez em Londres?)
(1b) O livro **caiu** no chão. (O livro caiu. *O que aconteceu/fez no chão?)
(1c) Ele **morava em Londres**. (*Ele morava. *O que aconteceu/fez em Londres?)

In (1) two tests are applied: The **adverbial omission test** tests for optionality (positive in 1a and 1b), and the **predicate isolation test** tests whether the adverbial can be isolated from the VP (i.e. is part of the VP or not). This test is positive only in (1a) where the adverbial is an adjunct at sentence level, not a complement at VP level.

Dependency-wise all adverbials "point" towards the main verb, the functional level-distinction between @ADV and @ADVL being made explicit by the tags and not the attachment markers (in more precise terms, @ADVL constituents do not attach directly to the main verb, but rather to a complex head, the VP - or verb chain - as a whole).

Semantically, the most common valency bound adverbials are locatives and directives, appearing both as first and (in transobjective constructions) second complement:

(2) **Locative adverbial** (i.e. ADV or PP) **complements** (obligatory in bold face)

	1. complement		2. complement	
LOCATION	<va+LOC> **estar** *na casa* 'be' **morar** 'live' **ficar** 'be, stay' **quedar** 'stay'	A	<vta+LOC> pôr *o livro sobre a mesa* 'put' colocar 'place' botar 'place' deitar 'lay' deixar 'leave' instalar 'install'	C
	entrar 'enter' aportar 'arrive' arribar 'arrive'	B		
DIRECTION	<va+DIR> **ir** *para Brasil* 'go' **viajar** *para* 'travel'	D	<vta+DIR> mandar *o filho para Londres* 'send' enviar 'send'	F

	tornar *a* 'return' voltar *a* 'return' passar *a* 'pass' subir 'climb'	E	atirar 'throw' jogar 'throw' carregar 'carry' voltar 'return'	

Semantically a cline of transitivity can be created for verbs valency-governing place-adverbials:

A) be AT	STATE ("IMPERFECTIVE") - copula-like
B) become (be AT)	TRANSITION ("PERFECTIVE") - ergative
C) cause (be AT or become(be AT))	CAUSATIVE - STATE
D) move TOWARDS	ACTIVITY/PROCESS ("IMPERFECTIVE") - motion
E) move TO	ACTION/EVENT ("PERFECTIVE") - ergative motion
F) cause (move TOWARDS/TO)	CAUSATIVE - ACTIVITY/PROCESS

where the semantic prototypes **"to be"** (A: estar, morar) and **"to move"** (D: ir, viajar) can undergo perfectivisation/ergativisation (B 'entrar' vs. E 'voltar') or causativisation (E 'pôr' vs. F 'mandar'). In a way, the valency clines A-B-C and D-E-F are analogous to the relation between the copula verb 'to be', the change-verb 'to become' and the causative 'to make'. Portuguese makes a distinction between an "identity copula" ('ser') and a "state copula" ('estar'). The latter, derived from Latin 'stare' ('to stand'), covers, among other things, locative adverbial complements, providing a lexical argument in favour of maintaining the duality of "real" copula (belonging to the <vK> valency class) and "locative" copula (belonging to the <va+LOC> class). The lexical strength of place-argument governing verbs can further be noticed from the fact, that most Portuguese expressions denoting 'to become' (i.e. ergative copula verbs) are metaphorically derived from corresponding <va+LOC> verbs: *ficar* ("stay"), *chegar* ("arrive") *a ser*, *sair* ("leave") and (in philosophical language) *devir* ("be-come").

Interestingly, the two basic prototype groups of 'being AT' and 'moving TOWARDS' (A and D) correspond to verbs with obligatory adverbial complements, while ergatives[174] and causatives have optional adverbial complements. A possible

[174] In the context of this chapter, I define *ergative verbs* as verbs featuring an internal PATIENT/THEME subject argument, i.e. denoting affectedness of the subject, meaning result focus and a change in either state or - here - location, - implying perfectivity, action rather than activity, event rather than process. Mateus et. al. (1989, p.173) argue, that inaccusative/ergative verbs (*desmaiar* - 'to pass out', *chegar* ' arrive) on the one hand and inergative verbs (*rir* - 'laugh', *trabalhar* - 'work') on the other can be distinguished from each other and in opposition to the class of monotransitive verbs (*revir* - 'check, look through'). While the latter features what they call an internal (in terms of constituent structure) PATIENT/THEME direct object and an external AGENT subject argument, inaccusatives/ergatives have only an internal and inergatives only an external argument, both expressed as subject. For Portuguese, tests are suggested indicating that the

explanation for this may be that a PATIENT/THEME argument ranks higher on the obligatority scale than a PLACE or DIRECTION argument in much the same way as the direct object ranks higher than the object complement in a transobjective construction - it is, after all, the object that is *semantically* complemented, not the verb. In fact, both causatives and ergatives have the argument role of PATIENT, either as external (object) or as internal (subject) complement.

In Portuguese, locative and directive adverbial complements can be distinguished lexically through the heading preposition ('em', 'sobre' vs. 'para', 'a') or, to a certain degree, the head adverb ('aqui' vs. 'fora', which can also be used for a pronominal substitution test on pp-adverbials). This is why valency marking of adverbial governing verbs makes sense not only for its own (syntactic) sake, but from a lexical disambiguational point of view, too. *Ir* ('to go') and *ser* ('to be'), for instance, share many inflexion forms (e.g. all of the perfeito simples, mais-que-perfeito and future subjunctive tenses), but obligatorily valency-bind different types of complements, - subject complements in the former, and directive adverbials in the latter case.

Like other non-subject non-pronoun complements, most adverbial objects appear usually *after* the main verb (@<ADV), while the closed class of relative and interrogative adverbs is used in clause-initial position (3a).

subject of inaccusatives/ergatives indeed covers what in monotransitives would be the PATIENT/THEME role, while the subject of inergatives typically retains the AGENT role:

1. Inaccusatives, but not inergatives, allow, like monotransitives, attributive or predicative past participles, in analogy with passivisation:

	Predicative	Attributive
monotransitive:	a janela está fechada	a janela fechada
inaccusative:	o rapaz está desmaiado	o rapaz desmaiado
inergative:	*o rapaz está rido	* o rapaz rido

2. Since the participle in ablativus absolutus constructions (absolute participle constructions) predicates a PATIENT/THEME, not an agent, inaccusatives and monotransitives with their object are allowed, inergatives and monotransitives with their subject are not.

	finite clause	abl.abs. with PATIENT/THEME	abl.abs. with AGENT
monotransitive:	João reviu as provas	Revistas as provas, o João ...	*Revisto o João, ..
inaccusative:	João chegou	Chegado, o João ..	-
inergative:	João trabalha	-	*Trabalhado o João, ...

3. Only inergative and monotransitive verbs, not inaccusatives, allow the AGENT suffix -*ador* ('-ator'):

	base form	agent noun
monotransitive:	construir	constructor
inaccusative:	chegar	*chegador
inergative:	trabalhar	trabalhador

(3a)

Onde	[onde] <interr> <aloc> ADV @ADV> 'where'
mora	[morar] <va+LOC> <ink> V PR 3S IND VFIN @FMV ‘[does he] live’
$?	

With the exception of leading prepositions ('a mulher a quem amava'), clause initial position is, of course, also the normal place for complementisers (not only adverbial complementisers, but also pronouns and conjunctions), heading non-finite (3b1/2), finite (3c/d) or averbal subclauses (AS). In the AS- and FS-case (3c/d), the complementiser is obligatory, and any adverbial complemetiser will therefore - in my system - also bear the clause function tag.

Apart from valency binding within the subclause, adverbial complementisers can also help attach an FS to an "outside" (main clause) valency link, in the interrogative case usually to a cognitive or speech-verb ('saber' in 3b1, 'perguntar' in 3c), or a "cognitive noun" ('pergunta' in 3d), and in the relative case possibly to yet another <va> verb ('mora onde eu moro' - 'he/she lives where I live.'), by turning the FS concerned into the right type of adverbial (in the example, locative).

Note, that heads for both adverbial arguments and interrogative subclause arguments are marked for such valency, cp. <va+LOC> ('morar' in 3a-d) and <+interr> ('perguntar' in 3c) or <+de+interr> ('pergunta' in 3d)[175].

(3b1)

Não	[não] <dei> <setop> ADV @ADVL> 'not'
sabe	[saber] **<v-cog>** <ink> V PR 3S IND VFIN @FMV ‘[he] knows’
onde	[onde] **<interr>** <aloc> ADV @ADV> ‘where’
morar	[morar] **<va+LOC>** V INF 0/1/3S @IMV @#ICL-<ACC ‘[he] lives’
$.	

(3b2)

Não	[não] <dei> <setop> ADV @>A ‘not’
tem	[ter] <vt> <ink> <rH> V PR 3S IND VFIN @FMV ‘[he] has’
onde	[onde] **<rel> <aloc>** ADV @ADV> ‘where’
morar	[morar] **<va+LOC>** V INF 0/1/3S @IMV @#ICL-<ACC ‘to live’
$.	

(3c)

Perguntei	[perguntar] **<+interr>** <vH> <ink> V PS 1S IND VFIN @FMV ‘[I] asked’
onde	[onde] **<interr> <aloc>** ADV @ADV> @#FS-<ACC ‘where’
morava	[morar] **<va+LOC>** V IMPF 1/3S IND VFIN @FMV ‘[he] lived’
$.	

(3d)

[175] In the latter case (<+interr>) arguments may of course be other than adverbial, as long as they constitute interrogative (finite) subclauses e.g. *a pergunta de quem vai participar* ('the question of who is going to participate.')

a	[a] <art> DET F S @>N ‘the’
pergunta	[pergunta] <s> **<+de+interr>** N F S @AS< ‘question’
de	[de] <+top> PRP @N< ‘as to’
onde	[onde] **<interr> <aloc>** ADV @ADV> @#FS-P< ‘where’
mora	[morar] **<va+LOC>** <ink> V PR 3S IND VFIN @FMV ‘[he] lives’

Adverbial complements can, of course, co-occur with (same-level) adjuncts (3e) or (dependent) set-operators (3f), making contexts more complicated for the CG-rules.

(3e)

o	[o] <art> DET M S @>N ‘the’
tesouro	[tesouro] <mon> N M S @SUBJ> ‘treasure’
ainda	[ainda] **<atemp> ADV @ADVL>** ‘still’
lá	[lá] <dei> <aloc> ADV **@ADV>** ‘there’
está	[estar] <va+LOC> <sN> V PR 3S IND VFIN @FMV ‘was’
$.	

(3f)

o	[o] <art> DET M S @>N ‘the’
tesouro	[tesouro] <mon> N M S @SUBJ> ‘treasure’
nem	[nem] **<setop>** ADV **@>A** ‘not even’
lá	[lá] <dei> <aloc> ADV **@ADV>** ‘there’
está	[estar] <va+LOC> <sN> V PR 3S IND VFIN @FMV ‘was’
$.	

The @ADV tag (for valency bound adverbial) is used both for monotransitive and ditransitive constructions, though one might argue that especially PLACE-adverbs function somewhat like subject and object complements, respectively, predicating a location of an NP. *Ele está doente* and *Ele está no Rio* would then both be regarded as simple copula patterns, with an @SC tag for both the adjective *doente* and the locative adverbial *no Rio*. However, since some copula verbs, like *'ser'*, do not appear with place adverbial complements[176], and some <va> verbs, like *'ir'*, can't usually be used with nominal complements, I prefer to make a valency class distinction based on complement *material* (restricting copula class membership to *nominal* subject complementation), and uphold the distinction between @SC/@OC and @ADV.

The two types of @ADV matching <va> and <vta> valency, respectively, could also be distinguished by hybrid terms (@ADV-SC and @ADV-OC), derived - in a (to-be-written) CG mapping module - from the absence or presence of an accompanying @ACC complement:

(4)

atirando	[atirar] <vta+DIR> <vH> V GER @IMV @#ICL-ADVL ‘throwing’

[176] apart from the special focus construction with *ser*

a	[a] <sam-> <+top> PRP **@<ADV(-OC)** ‘at’
os	[o] <-sam> <art> DET M P @>N ‘the’
ramos	[ramo] <anbo> N M P @P< ‘branches’
sua	[seu] <poss 3S/P> <si> DET F S @>N ‘his’
cantiga	[cantiga] <ll> N F S **@<ACC** 'song'
dolorosa	[doloroso] <n> ADJ F S @N< ‘painful’

Besides PLACE-adverbials, also other semantic types of adverbials can be valency bound (bold face for obligatory complements):

* TIME	<vt+TEMP>	**durar**, passar
* QUANTITY	<vt+QUANT>	crescer, **custar**, **pesar**, **valer**
	<vta+QUANT>	reduzir, aumentar
* QUALITY	<va+QUAL>	**estar**, **saber**, vir

Again, verbs with non-obligatory complements are either ergative-inaccusative and have an internal PATIENT/THEME subject argument (*passar*, *crescer*, *vir*) or - for monotransitive, inergative verbs - take the adverbial as *second* complement (*reduzir*, *aumentar*). The quality adverbials are restricted to 'bem'/'mal' and very few verbs (e.g. *saber* - 'taste'). They alternate with adjective subject complements, approaching the QUALITY class to that of ordinary copula verbs (iii) - and, in fact, 'estar' is often considered a member of that class:

(i) O projeto vá <va> mal ADV @<ADV. ('The project isn't going well.)
(ii) A mãe está <va> bem ADV @<ADV. ('Mother is fine.')
(iii) A mãe está <vK> cansada V PCP @<SC. ('Mother is tired.')

Only very few verbs fit in the non-locative @ADV classes, and many (all of the TIME group and most of the QUANTITY class) have atypical, *nominal* selection restrictions on the adverbial (which is why these are valency tagged in the lexicon as <vt>, and not <va>):

(5a)

vamos	[ir] <x> <ink> V PR 1P IND VFIN @FAUX ‘[we] are going to’
passar	[passar] **<vt+TEMP>** <rH> V INF 0/1/3S @IMV @#ICL-AUX< ‘pass’
quase	[quase] ADV @>A ‘nearly’
um	[um] <card> NUM M S @>N ‘one’
mês	[mês] **<dur>** **<per>** <num+> N M S **@<ADV** ‘month’
em	[em] <+top> PRP @<ADVL 'in'
Porto=Alegre	[Porto=Alegre] <top> PROP 'Porto Alegre'

(5b)

o	[o] <art> DET M S @>N ‘the’

carro	[carro] <V> N M S @SUBJ> ‘car’
vale	[valer] **<vt+QUANT>** <sN> V PR 3S IND VFIN @FMV ‘is worth’
de	[de] PRP @>N '-'
$5	[5] <cif> <card> NUM M/F P @P< '5'
a	[a] PRP @NUM< 'to'
$7	[7] <cif> <card> NUM M/F P @P< '7'
milhões	[milhco] N F P **@<ADV** 'million'

A lone counter example for "traditional" (ADV or PP) material in a non-locative @ADV complement is the pair 'reduzir' (reduce) and 'aumentar' (increase), which can even - like 'ir' and 'viajar' - take *two* adverbial complements.

(6)

reduziu	[reduzir] **<vta+QUANT>** <ink> <rH> V PS 3S IND VFIN @FMV ‘[he] reduced’
o	[o] <art> DET M S @>N ‘the’
quadro	[quadro] <sit> N M S @<ACC ‘size’
de	[de] <+hum> PRP @N< ‘of’
pessoal	[pessoal] <HH> N M S @P< ‘staff’
de	[de] PRP **@<ADV** 'from'
$1100	[1100] <cif> <card> NUM M/F P @P< '1.000'
para	[para] <+hum> PRP @<ADV ‘to’
$600	[600] <cif> <card> NUM M/F P @>N '600'
empregados	[empregado] <prof> N M P @P< ‘employees’
$.	

4.5.4.2 Adjunct adverbials

In my system, adjunct adverbials (@ADVL> and @<ADVL) are defined syntactically, as adverbials that attach to a verbal constituent at clause level, and are not or only weakly valency-bound by the main verb. The latter condition (non-agument status) can be operationalised by the adverbial omission test (where adjunct adverbials should test positive, cf. 4.5.4.1). The former (clause level attachment) can be tested by replacing all constituent groups not containing verbs or complementisers (and therefore, neither, clauses) by pronouns, i.e. NPs with independent (SPEC) or personal (PERS) pronouns, ADJPs outside NPs with *tal* ('such'), and multi-word ADVPs outside NPs or ADJPs with *lá* ('there'), *então* ('then'), *assim* ('in this way') or *tanto* ('to this extent'). Groups are pronominalised in this order (NP, ADJP, ADVP), and bigger groups "swallow" smaller groups (as long, of course, as no verbal or complementiser material would be swallowed at the same time). After pronominalisation, all remaining, "unswallowed" lexical adverbs (i.e. what the lexicon and the tagger module call adverbs), as well as all non-argument ADVPs and PPs (i.e. ADVPs and PPs that pass the VP isolation test, cf. 4.5.4.1) will be considered adjunct adverbials. In a second round of pronominalisation, after subclauses have been checked for their adverbial adjunct content, all subclauses that are immediate constituents of groups are removed, and the remainder (i.e. clause level constituent subclauses) are pronominalised. Consider the following sentence (i), where maximal verb-and-clause-free constituents are underlined and adjunct adverbials are in bold face:

(i)
Na semana passada, enquanto **ainda** cuidava de seu novo livro,
PP1 **@ADVL>** ADV **@ADVL>** PP2
o velho poeta anunciou, ***muito* depressa**, que não contava dar
NP1 ADVP **@<ADVL**
a entrevista *antes (ADV @ADVL>A)* planejada *pela editora (PP3 @A<PASS)*.
NP2

With pronominalisation, we get (ii):

(ii)
Então, enquanto **ainda** cuidava dele,
@ADVL> **@ADVL>**
ele anunciou, **assim**, que não contava dar isto.
@<ADVL

PP1, an adjunct adverbial, can be omitted ('Enquanto ainda ...') and isolated from the VP/clause ('Na semana passada, o que fez?'), while PP2, a prepositional object, doesn't pass the second test ('*De seu novo livro, o que fez?' yielding a completely different meaning) and is awkward with the first. *Ainda* is not swallowed by any group structure, qualifying for @ADVL status. *Muito*, however, functioning as adverbial adject (@>A), is part of a larger adverbial phrase, *muito depressa*, and it is only this larger group, that receives adjunct adverbial status (it's omittable and not valency-bound). Of the two NPs not part of a PP, the second, NP2, is problematic, because one might argue that the postnominal participle is verbal, after all, calling for pronominalisation of *smaller* units (iii):

(iii)
isto *antes (@ADVL>A)* planejada *por ela (@A<PASS)*.
PP

(iii) would leave *antes* and *pela editora* (PP3) "unswallowed", and both would have to be related to *planejada*, PP3 as argument (agent of passive, unisolatable: '*e a entrevista, pela editora, o que era?), and *antes* as adjunct adverbial, being both omittable ('a entrevista planejada ...') and VP-isolatable ('e a entrevista, antes, o que era?'. As described in chapter 4.4.4.2, this ambivalence is solved by assigning the constituents of an attributive participle phrase both an adject dependency tag (adverbial adject) and a clause-level function, in the form of hybrid tags, here @ADVL>A and @A<PASS.

In order to identify subclauses as adjunct adverbials, the tests would have to be repeated as many times as there are clause layers, with progressive "upward" pronominalisation of subclauses, - infinitive clauses, interrogative clauses and finite que-clauses functioning like NPs, post-nominal relatives like APs, gerund clauses like ADVPs etc.

As the above "independence" tests show (i.e. being isolatable from the VP and not being swallowed by pronominalised groups), the category of adjunct adverbials could also be viewed as a class defined *via exclusionis* - i.e. all adverbial material that is not arguments (@ADV> and @<ADV) or adjects (@>A and @A>), since the latter would be swallowed by groups, and the former by the verbal valency frame.

Though linguistically satisfactory, operational definitions (as also discussed in 4.1.3) are not necessarily good tools for automatic disambiguation. Relying on human language competence, movement and grammaticality judgement, they cannot directly be exploited by Constraint Grammar rules. Such rules have to depend on word order, context patterns and lexical potential, and in this respect, adjunct adverbials are a very heterogeneous lot - comprising what "traditionally" or semantically might be grouped as *circumstantial* adverbials of place, time, manner etc. (all of which test positive in the adverbial omission test and the predicate isolation test, cf. 4.5.4.1), but also the adverb subclasses of operators (cf. 4.5.4.5) and intensifiers (cf. 4.5.4.3), where they aren't used as adjects. Relative and interrogative adverbs (cf. 4.5.4.4) enter the class with respect to their clause *internal* function ("external" complementiser function being expressed implicitely by their bearing the subclause function tag @#).

Though in part *semantically motivated*, these subclasses can - to a certain degree - be differentiated by their *syntagmatic* preferences, with circumstantials (characteristically) in clause-final, clause-initial or clefted position, and set- and time-operators in pre-scope position (i.e. immediately left to the predicator). Though they don't allow clefting either, meta-operators differ from the other operator types in that they can be fronted, like circumstantials, or break syntactic continuity by comma-isolation. Most circumstantials, notably not-too-heavy time & place adverbials, can appear between subject and predicator, i.e. in what seems to be pre-verbal position. In the case of positional conflict, however, there is a clear order of constituents[177]:

SUBJ – circumstantial/meta operator – time operator – set operator – VFIN
ele *hoje obviamente* *ainda* *não* *comeu*
obviamente hoje

Manner adverbials, however, that semantically refer to the main verb (i.e. can't be isolated by the *'o que aconteceu?'* test), are barred from pre-verbal position. Manner adverbials that semantically refer to the subject (and, thus, could be called *predicative adverbials*), don't pass the 'o que aconteceu?' test either (since 'acontecer' covers both subject and predicate), but they do appear between subject and predicate (just like their semantic cousins, @PRED constituents). If in positional conflict with other pre-predicate adverbials, predicative adverbials are placed left of set-operators, and right of circumstantials and meta-operators:

A menina **hoje** (TIME) comeu o bolo.
timidamente (PREDICATIVE)
devagar* (MANNER)

A menina **provavelmente hoje já timidamente também** comeu o bolo.

In contrast to operators - the semantic function of which is to modify the semantic content of other constituents -, circumstantials and adverbial complementisers do have a semantic payload of their own, most noticeable that of PLACE, TIME and MANNER, as in *"onde "*, *"quando"*, *"como"*. These semantically "loaded" adverbials have to be disambiguated into valency-bound @ADV (cf. 4.5.4.1) and adjunct @ADVL, but only for place-adverbials is there a significant number of binding verb lexemes. Examples for time ('durar uma semana') and manner ('estar bem') are few and controversial.

[177] Sequences like 'não ainda', 'até hoje' etc., that seem to violate this rule, can be explained by the fact that set-operators can modify time operators, and both can modify circumstantials at *group level*. On *clause level*, however, there is a preferred order.

Therefore, the @ADV-@ADVL distinction[178] can be made almost entirely at the "early" syntactic level - i.e., the *mappings* section of the CG-rules file, where most @ADV cases can be identified, depending heavily on lexical context (especially, lexical valency information), and less on syntactic reasoning.

Material-wise, CG will map morphological adverbs only as @ADVL, @ADV on the clause level, and mostly @>A or @A< on the group level. Prepositional phrases (PPs), however, will in addition be mapped as @PIV (prepositional object) @SC (subject complement) or @PRED (free predicative) on clause level, and @N< (post-nominal) on group level. In addition, the internal structure of PPs can be much more complex than that of most adverb-phrases, which makes relevant clause level context more distant and less visible to the CG disambiguation rules. I will therefore focus on the syntactic function of PPs throughout the rest of this chapter.

Though valency bound, and therefore marked lexically on the verb, @PIV is much more difficult to disambiguate than a PP @ADV, since it covers a wide range of prepositions and semantic roles. PP @N< is somewhat easier to handle - at least where there is relevant valency information on the head -, since group level attachment allows my CG to rule out interfering clause level complements or adjuncts. At the same time, in the case of PP @N< hierarchies, leftward attachment can be left underspecified. In analogy to the clause-level distinction between @PIV/@ADV and @ADVL, also PP @N< can be differentiated into valency-bound "objects" and circumstantial "adjuncts": *discussão sobre a Dinda* (discussion about Parliament) vs. *discussão na Dinda* (discussion in Parliament). Again, the first invites lexical solutions, while the second is a real competitor to the @ADVL reading.

Just how big the syntactic ambiguity potential of PPs is in my CG-description, can be seen in (1). A simple adverbial tag (@ADVL) would be the obvious default reading for most of the categories below, - if the tag set was to be reduced for pedagogical reasons, to ensure a very low error rate, or for transformation into or comparison with other - less detailed - tagging systems. In my discussion of the other function tag alternatives for PPs, I will therefore focus on those traits that distinguish them from the @ADVL "prototype", as well as on the disambiguation tools employed.

(1a) @ADVL VEJA, **30. Dez. 92**
(1b) @<ADVL Existem **no mundo** apenas dois fozes tão enormes ..
(1c) @ADVL> **Em 1992,** o desemprego foi recorde.
Ao retornar, **em 1986**, encontrou o país transformado ...
(1d) @<ADV Antigamente, morava **no Rio**.
(1e) @ADV> No Rio, **onde** morava antigamente.

[178] After this, one of the two classe, @ADVL, will still be ambiguous, - with regard to other adverbial classes like @>A (adverbial pre-modifier).

(1f) @<PRED .. o que explica a reação da imprensa, quase **de incredulidade**
(1g) @PRED> **Com seus lagoas, praias e dunas**, o Delta é uma maravilha
(1h) @>N Recebe **de dois a três** convites por ano ...
(1i) @N< O homem **com a bicicleta da China**
(1j) @>A Quero **pelo menos** dez cópias.
(1k) @A< Sou favorável **a @A< impostos** sobre @N< o consumo, sem @<PRED taxar a produção em todo seu processo.
... com tarifas que são reajustadas abaixo **da inflação** ...
(1l) @<SC A mídia de espectadores era **de 800**.
(1m) @SC> e ganha-se um vice que nem **da mesma corrente política** é.
(1n) @<PIV Vota-se **num prefeito**, como se acabou de fazer ...
(1o) @PIV> Desiludido com o órgão, **do qual** se demitira em 1986 ...
(1p) @N<PRED ... fez-se fotografar de cuecas, com a mão **na pélvis**
(1q) @A<ADV um livro colocado **no centro da mesa**
(1r) @A<PIV dividido **entre o bem e o mal**
com o joelho fincado **no morto**
(1s) @A<ADVL o plano de enxugamento anunciado pela Westinghouse **há duas semanas**, ...
(1t) @A<PASS encontrou o país transformado **pela campanha**
(1u) @AS< ... experiência que os homens aparentemente precisam de viver, ainda=que @SUB @#AS-<ADVL só @>P **na @AS< imaginação**, para firmar sua identidade
... uma crise como @COM @#AS-N< **nos @AS< anos trinta**

Of the above, the functions closest to the @ADVL tag proper are obviously other clause level functions, i.e. @PRED and the @PIV-@ADV-@SC group, though for different reasons. The first is - like @ADVL itself - a clause-level adjunct without a valency link, and since PPs are not, like APs (the prototypical @PRED), morphologically (agreement-) marked for their link to a nominal "head", it can be very difficult to make the distinction indeed. Consider the cline in (2) where "adverbiality" (@ADVL) increases, and "predicativity" (@PRED) decreases from (2a) to (2c).

(2a) **Com seus 70 anos**, o presidente parece muito velho.
(2b) **Com seus 70 anos**, o presidente nem atinge a idade normal para um líder chinês.
(2c) **Com seus 70 anos**, o presidente tem um verdadeiro tesouro de experiências.

The difference is almost purely semantic, but can be made visible by preposition replacement. While *com* in (2a) is hard to replace ('with' functioning as a very neutral predicative link), it can be substituted for by the more instrumental *em função de* ('by means of') in (2b), or the causal *por causa de* ('because of') in (2c).

@PIV, @ADV and @SC are - unlike @ADVL - valency bound arguments, but they, too, direct their dependency markers towards the main verb, which is one of the reasons why older grammars often treat them as adverbials. As can be seen from (1q-1s) the above distinctions are maintained in participle "clauses", hybrid constructions with the syntactic distribution of an AP, but the complement valency of the corresponding verb. A special case is the agent of the passive (1t), which, as a PP, is - form-wise - similar to the other post-participle @A< functions, but corresponds to the "original" subject[179], which, of course, is a *non*-adverbial argument of the verb[180], earning its reincarnation in the participle clause a separate tag[181].

From a constituent point of view, the least @ADVL-compatible functions are those of @N< and post-adjectival or post-adverbial @A< (rarely @>N and @>A), because they are not clause/VP constituents, but nominal group constituents, either dependent on a noun (@N<) or an adjective (@A<):

(3) **Table: PP-functions**

	valency bound	**not valency bound**
attaching to V-head	@ADV, @PIV, @SC	@<ADVL
attaching to N-head	@N<	@PRED, @N<, (@>N)
attaching to A-head	@A<	(@>A)
attaching to PCP-head	@A<ADV, @A<PIV	@A<ADVL
attaching to complementiser or noun preceded by com/sem	@AS<	@N<PRED

A special case are @AS< and @N<PRED where a PP functions as a predication without a predicator, directly attaching either to a complementiser (1u) or a noun preceded by 'com' or 'sem' (1p). From a semantic perspective, one could say that the PP in these cases is *head* of the clause, since it incorporates the predicator (which would otherwise function as head).

@A< prepositional groups (PPs) are most often valency governed (with a few time & place exceptions), whereas PP @N< could be split up in a bound and an

[179] Another non-adverbial constituent in participle clauses would be a direct object, which, however disappears in the passsivisation process, or an object complement that turns subject complement in the participle clause*: um recurso chamado* ***agravo regimental***

[180] Both subject and passive argument

[181] Of course, @A<SUBJ would be a viable alternative tag, but since this would be less consistent with the surface syntactic notation otherwise used, I prefer the special tag @A<PASS. That this tag is more akin to @A<PIV than to @A<ADVL can be seen from the negative VP-isolation test: '*e o país, pela campanha, o que era?'.

unbound subclass, mirroring the corresponding distinction @ADV - @ADVL with a similar frequency advantage on the side of the free constituent. In the present version of the parser, only the valency tag of the head noun indicates the complement function of a postnominal PP, while the dependency marker at the other end of the valency link is the same as for attributive postnominals (@N<): *uma discussão* *<+sobre>* *N* ***sobre @N< política.***

Noun based tags for valency potential are also quite useful for the CG-rules when deciding whether a PP should be @N< of @ADVL in the first place. In contrast, it is very difficult to decide by syntactic context alone, whether a PP is a free @N< or a free @<ADVL. Therefore, semantic and probabilistic tools are integrated into the relevant rules.

For example[182], time-PPs rarely attach to nouns, that are not deverbal or "occasions" (events, things that happen, lexicon-tagged as <occ>):

> REMOVE (@N<) (0 @ADVL LINK 0 PRP-TEMP) (*1 N-TIME BARRIER NON-PRE-N) (*-1 NON-NP BARRIER <occ>) ; # remove the postnominal reading for PPs in favour of a temporal adverbial reading, if the leading preposition is of the right type, and if there is a time-noun somewhere to the right with nothing but prenominals in-between, and the NP to the right does not contain an event-noun (i.e. if a non-NP word can be reached to the left without an event-noun interfering)

Place-PPs seldom attach to nouns denoting time, quantity or humans:

> REMOVE (@N<) (0 @ADVL LINK 0 PRP-LOC) (*-1 N-TIME/QUANT/HUM BARRIER NON-POST-N) ; # remove the postnominal reading for PPs in favour of a locative adverbial reading, if the leading preposition is of the right type, and if there is time-, quantity- or human noun to the left with nothing but postnominals interfering

"Thought products" (<pp>, e.g. *retrospectiva*) are likely to govern "topic"-PPs headed by 'sobre' ('about'):

> REMOVE (@<ADVL) (0 PRP-SOBRE AND @N<) (*-1 (<pp>) BARRIER @NON-N<) ; # remove the adverbial reading for PPs in favour of a postnominal reading, if the leading preposition is *sobre* ('about'), and if there is a "thought product to the left with nothing but postnominals in-between.

Some - older - heuristic rules do not even make use of semantic tags, but exploit syntactic clues with a semantic correlation. The preposition 'com', for example, is more likely to be @ADVL, if it heads an "instrumental" PP, but more likely to be @N< if it

[182] The following examples are all heuristic, so it must be stressed that it isn't too hard to find or construct exceptions, but since such rules are grouped and ordered according to how safe they are, they may still be quite useful, since they will be applied to ever smaller percentages of a text, *after* safer rules have disambiguated the majority of cases - where a heuristic rule that is wrong has nothing left to discard (since the parser automatically "protects" the last remaining reading).

heads a "material" PP. The following 2 rules draw on the fact that instruments are countables, and therefore likely to be preceded by a determiner, whereas materials can follow the preposition directly.

"<com>" REMOVE (@<ADVL) (0 @N<) (1 @P<) ; # choose postnominal over adverbial reading for 'com', if the head of its argument follows directly to the right
"<com>" REMOVE (@N<) (0 @<ADVL) (NOT 1 @P<) ; # choose adverbial over postnominal reading for 'com', if the head of its argument does not follow directly to the right

For PPs that have been entered into the lexicon as polylexicals, preferences can be given in the form of probability ordered secondary tags, exploiting the fact that tag conjunction inside a single CG rule condition is interpreted by the compiler in a linear way:

SELECT (@N<) (0 <adj> + <adv> + PP) ; # choose the attributive (postnominal) reading, if a "lexical" PP is probability marked for "adjectivity"
REMOVE (@N<) (0 <adv> + <adj> + PP) ; # choose the adverbial reading, if a "lexical" PP is probability marked for "adverbiality"

Finally, the last heuristic rules will apply pure statistical knowledge, like when the @ADVL reading is removed from the tag-line of the preposition 'de', which is very common in postnominal PPs, but rarely appears as adjunct adverbial (like when denoting the starting point of a time span)[183]:

REMOVE (@<ADVL) (0 PRP-DE LINK 0 @N</<PRED) ; # remove, for the preposition 'de', the adverbial reading in favour of a postnominal or free predicative reading

For the 6 most common prepositions, table (4) looks at the syntactic function probabilities for a preposition in the immediate right hand context of a noun. Only *left* attachment is analysed, and argument-of-verb function (@<PIV, @<SC, @<ADV) is excluded, since it can be disambiguated in a non-heuristic way. Percentages are derived from manual inspection of a 12.000 word newscorpus chunk.

(4) **Table: pp-attachment statistics**

	@N< modifier	@N< argument	@<ADVL adjunct
a	-	32 %	68 %
com	33 %	22 %	45 %

[183] Another adverbial use, that of point of departure, is usually valency bound, and, like the @PIV-object cases, fully disambiguated at this late, heuristic stage of analysis.

de	77 %	23 %	-
em	33 %	12 %	55 %
para	8 %	23 %	69 %
por	38 %	15 %	46 %

Provided full nominal valency is implemented in the lexicon, two of the cases can be resolved 100%: *a* always prefers the adverbial over the postnominal modifier reading, and *de* is, when in doubt, a postnominal modifier. Though *de* does enter in adverbial pairs with *a* (from - to/till), this was rare in the above statistics[184], and it also is more significant for *a* than for *de*, since *de* is 10 times as frequent in postnominal position. Also, in the *de...a* pair, it is usually *a* that gets a left-hand noun context, while *de* follows the verb: *todos adoram V,* ***do*** *primogênito N* ***ao*** *caçula*.. The preference of *a* for an adverbial reading is further upheld by a kind of retrograde valency, where the noun in adverbial expressions like *à distância, a disposição, à moda* is marked by a <a+> tag which helps disambiguate the preceding preposition.

For *por,* virtually all the modifier @N< instances are "frequency terms" like *horas* ***por dia****, dólares* ***por dia****, vezes* ***por semana****,* which can be identified by right hand context, too, by looking for TIME-nouns of the <dur> (duration) subclass.

Para is much more likely to be used adverbially, and - like for *a* - a common right hand context suggesting such an @ADVL reading is a verb in the infinitive, making the PP a kind of purposive "subclause"

The hard cases left, then, are *com* and *em,* which in postnominal position can be used to express a feature or the location of an object, respectively. Sadly, neither feature modifiers nor location modifiers need to be valency bound. Crude semantic rules can be fashioned to supplement the purely statistical rule of preferring the adverbial reading, after all valency information has been used. Thus, since *com + feature* is semantically unlikely to function as adverbial, semantic tags denoting features (like <feat>, <fh>) can be used to decide on postnominality for *com*-PPs. Likewise, *em*-PPs could be tagged for @ADVL if they do *not* denote place, but time, since only a restricted set of deverbal nouns or 'happenings' allows temporal modifiers.

[184] Rather, most cases would be treated as valency bound adverbial objects (@ADV) in conjunctions with MOVE-verbs.

4.5.4.3 Intensifier adverbs

Intensifier (quantifying) adverbs (ADV <quant>)

assaz, bastante, bem, cada=vez=mais, eminentemente, extremamente, igualmente, imensamente, incrivelmente, mais=ou=menos, mui, muito, muitíssimo, particularmente, pelo=menos, pouco, pouquíssimo, quanto=mais, sobremaneira, sobremodo, terrivelmente, totalmente, tremendamente, vagamente
POST-ADJECTS (@A<) **<post-adv>** demais, paca, por=demais, por=demasiado *(devagar demais)*
MORPHOLOGICAL PRONOUNS **<det>** algo, meio, nada, que, todo, um=tanto, um=pouco
CORRELATIVE COMPARATIVES **<KOMP><corr>** mais, menos, mesmo
EQUALITATIVE COMPARATIVES **<KOMP><igual>** tanto, tão

The semantico-syntactict category of intensifier and the dependency-syntactic category of @>A (@A<) adject are almost co-extensive. On the one hand, intensifiers can be defined as those adverbs (and, morphologically, quantifier pronouns) that can appear in adverbial adject position, modifying adjectives or other adverbs in AP's - where they semantically permit quantifying. On the other hand, intensifiers are instrumental in defining the umbrella form category of AP. In principle, PPs (like *sem graça*) can have intensifier pre-adjects, but this is a faily rare phenomenon (6c). One might argue that this restriction is semantic rather than syntactic, since most such PPs are circumstantial (time, place, source, goal etc.) and not "adjectival" (as *sem graça*). In fact, one could say that an intensifier @>A moulds a new AP even from a PP head.

The typical position for intensifiers is directly *before* the term modified (6a, 6b, 6c), with a few exceptions (<post-adv>, *demais*, *paca*). The <det> subclass, which might be described as what really are morphological determiner pronouns (DET) *functioning* as intensifiers, is limited to the above-mentioned pre-adject position (6a1, 6d4), whereas the other intensifiers can also modify the predicate, in which case they typically appear directly before or after the verb or verb chain (6d1) and never in the circumstantial position (as tested by comma-separation, 6d3, or que-focusing, 6d2).

6a1) A @>N tarefa @SUBJ> era @FMV **nada @>A** fácil.
6a2) Tinha @FMV esta @>N idéia @<ACC **um=pouco @>A** iconoclasta @N<.

6b) A @>N menina @SUBJ> nadava @FMV **tremendamente @>A** bem @<ADVL.

6c1) Achava @FMV a @>N proposta @<ACC **muito @>A** sem @N< graça @P<.
6c2) *Foi uma proposta muito para agradá-lo.

6d1) **Pouco @ADVL>** importa @FMV a @>N sua @>N religião @<SUBJ.
6d2) *É pouco que importa a sua religião.
6d3) *Pouco, importa a sua religião.

6d4) *(algo, meio, nada, um=tanto <det>) importa a sua religião.

4.5.4.4 Complementiser adverbs: Interrogatives, relatives and comparatives

semantico-syntactic distinctions:
INTERROGATIVE FUNCTION <**interr**> como, onde, quando, quanto
RELATIVE POSTNOMINAL FUNCTION <**rel**> como, onde, quando
COMPARATIVE FUNCTION
EQUALITATIVE <**komp**><**igual**> como, quanto, quão, qual
REFERENTIAL[185] <**ref**> conforme, consoante, segundo

semantic distinctions:
QUANTIFIER FUNCTION (INTENSIFIERS) <**quant**> quanto
SPACE: onde, TIME: quando, MODE: como

syntactic distinctions:
CONJUNCTIONAL FUNCTION <**ks**> POSTNOMINAL RELATIVES and EQUALITATIVE COMPARATIVES: como, onde, quando, quanto, qual, quão, REFERENTIAL COMPARATIVES: conforme, consoante, segundo, COMPLEX ADVERBIAL "ABSOLUTE" RELATIVES: a=proporção=que, ainda=quando, ao=passo=que, ao=tempo=que, apenas, assim=como, assim=que, bem=como, cada=vez=que, da=mesma=maneira=que, enquanto, logo=que, na=medida=em=que, sempre=que, senão=quando, tal=como, toda=a=vez=que, todas=as=vezes=que, tão=como, tão=logo, à=maneira=que, à=medida=que, à=proporção=que
PREPOSITIONAL FUNCTION <**prp**> como, conforme, consoante, qual, segundo

Complementiser adverbs are those adverbs that can subordinate a subclause. The prototypical, non-complex members of the class are adverbial pronouns (*como, onde, quando, quanto,* qual). Within the class, functional and semantic distinctions can be made. Most common is the semantico-syntactic distinction between interrogatives (primarily semantic) and relatives (primarily syntactic), which for non-polylexicals[186] is functional rather than lexical, since the list of interrogatives is a virtual subset of the set of the list of relatives. Another semantico-syntactic distinction is that between two types of comparatives, which semantically compare either degree or quantity (*como, quanto*) or assertions (*conforme*), and syntactically link to comparative hooks (*tanto ... quanto, tão ... como*) or refer to out-statement source (*segundo, conforme*). Both interrogative and comparative relatives can be semantically subdivided in space, time, manner and quantitiy adverbs.

[185] Referential comparatives like *segundo* and *conforme* are relative to a whole statement, somewhat like the sentence apposition *o que* in '*apareceram - o que muito me surpreendeu.* The category can be syntactically defined by the fact that subclauses headed by this type of complementiser adverb do not allow direct objects: '*... segundo denunciava no congresso *a sua opinião.*' Note that in '*... segundo o que denunciava*' the relative pronoun *o que* functions as direct object of a lower level subclause, itself governed in its entirety by *segundo.*

[186] The parser recognized som interrogative preposition+adverb compound *(aonde, donde)* and polylexicals: por=que, por=quê, há=quanto=tempo, a=que=propósito.

An important distinction with regard to syntactic function is about *what* kind of constituent a complementiser adverb introduces, i.e. which syntactic *form* it is part of. In traditional grammar this distinction would determine what word class a complementiser adverb is assigned, - that of conjunction (where it heads a finite subclause: *venha quando quiser*), or that of preposition (where it links to a noun phrase: *grande como um urso*). This way, only the interrogative members of the class need to be recognised as adverbs[187], since they can appear at main clause constituents or as group level modifiers. However, since a "conjunctional" adverb like *'como'* in *'não sei como funciona'* is morphologically indistinguishable from prepositional comparative *'como'* or the "pure" adverbial variant in an interrogative sentence like *'como se chama?'*, I retain the morphological umbrella class of adverb in my system, using secondary tags, <ks> (conjunctional use) and <prp> (prepositional use) to make the syntactic distinction.

The various semantico-syntactic distinctions discussed above are all registered as pontentialities in the lexicon, their disambiguation presently being carried out at four different levels of the parser:

Word class disambiguation level	relative <rel> vs. interrogative <interr>
Mapping level	adverbial argument @ADV comparative function @KOMP<, @COM
Syntactic disambiguation level	adjunct @ADVL function vs. modifier @>A function complementiser @# function
Valency instantiationn level	prepositional <prp> vs. conjunctional <ks> function

Syntactically the difference between what I call relative and interrogative use of adverbs is that relative use is restricted to the complementiser position of typically non-nominal finite subclauses or averbal subclauses, while interrogative use does occur at main clause (@ADV, @ADVL) and group level (@>A), as well as in the complementiser position of typically nominal subclause (e.g. @#FS-<ACC, @#FS-P<), both finite and non-finite but not averbal:

Table: syntactic distribution of relatives and interrogatives

	syntactic distribution of relative adverb	syntactic distribution of interrogative adverb
main clause level constituent	-	yes (@ADV, @ADVL)
group level constituent	-	yes (@>A)

[187] The most easily recognised relative pronominal adverb in traditional grammars is ‘onde’, since usage as a direct postnominal relative is not too rare: *a casa onde morava, o lugar onde encontramos.*

complementiser in nominal FS or ICL subclause	-	yes (@#FS-ACC, @#FS-P<)
complementiser in adverbial or attributive finite subclause	yes , <ks> (@#FS-ADVL, @#FS-ADV, @#FS-N<, @#FS-KOMP<)	-
complementiser in AS	yes, <prp> (@#AS-A<, @#AS-KOMP<)	-

The syntactic distinction between relatives and interrogatives is important both for morphological disambiguation (especially the future subjunctive of verbs) and for semantic interpretation (e.g. MT), as the following examples are meant to show:

discuta comigo ...

(i) quando <rel> @#FS-<ADVL investir 3S FUT SUBJ @FMV na Ásia amanhã
(let's discuss it when [når, wenn] you invest in Asia tomorrow)

(ii) quando <interr> @#FS-<ACC investir INF @IMV na Ásia de novo
(let's discuss when [hvornår, wann] to invest in Asia again)

(iii) quando <rel> @#FS-<ADVL investir INF @#ICL-SUBJ> na Ásia fizer sentido de novo
(let's discuss it when [når, wenn] investing in Asia makes sense again)

(iii) quando <interr> @#FS-<ACC investir INF @#ICL-SUBJ> na Ásia faz sentido de novo
(let's discuss when [hvornår, wann] investing in Asia makes sense again)

Subclass membership of *Quando* and the inflexion morphology of *investir* in the examples are disambiguationally interdependent - an interrogative reading of *quando* prohibits a future subjunctive reading of *investir* and vice versa, while a relative reading for *quando* only allows an infinitive reading for *investir*, if another finite verb form to the right (*fizer*) suggests an embedded infinitive clause. Semantically, there is a relation between the cognitive/speech verb status of *discutir* in the matrix clause, and the possibility of an interrogative reading for *quando*. Once disambiguated with regard to subclass, the different readings of *quando* have semantic consequences for translation, as can be seen from the Danish and German equivalents (*hvornår/wann* for the interrogative, *når/wenn* for the relative).

Some examples of typical syntactic uses of **relative adverbs** are:

1. relative complementiser in attributive (postnominal) finite subclause (FS)

veio	[vir] <va+DIR> <ink> V PS 3S IND VFIN @FMV ‘[he] came’
para	[para] <+top> <move+> PRP @<ADV ‘to’
a	[a] <art> DET F S @>N ‘the’
cidade	[cidade] <by> N F S @P< ‘town’
onde	[onde] <rel> <ks> <aloc> ADV **@ADVL> @#FS-N<** ‘where’

nascera	[nascer] <ve> V MQP 1/3S IND VFIN @FMV ‘[he] was born’

2. absolute relative complementiser in adverbial finite subclause (FS)

mora	[morar] <ink> V PR 3S IND VFIN @FMV ‘[he] lives’
onde	[onde] <rel> <ks> <aloc> ADV **@ADVL> @#FS-<ADV** ‘where’
o	[o] <art> DET M S @>N ‘the’
vento	[vento] <vind> N M S @SUBJ> 'wind'
reina	[reinar] <vi> <vH> V PR 3S IND VFIN @FMV ‘reigns’

3 (comparative) relative complementiser in adverbial small clause (AS)

o	[o] PERS M 3S ACC @ACC> ‘him’
amava	[amar] <vt> <vH> <ink> V IMPF 1/3S IND VFIN @FMV ‘[he] loved’
tanto	[tanto] <quant> <KOMP> <igual> ADV @<ADVL ‘as much’
quanto	[quanto]<rel><prp><komp><igual><quant>ADV **@COM @#AS-KOMP<** ‘as’
ela	[ele] PERS F 3S NOM/PIV @AS< ‘her’

4a. absolute relative complementiser in adverbial small clause (AS)

quando	[quando] <rel> <ks> ADV **@ADVL @#AS-ADVL>** ‘when’
em	[em] <+top> PRP @AS< 'in'
Brasil	[Brasil] <top> PROP M S @P< ‘Brazil’
$,	
faça	[fazer] <xdr> <ink> V PR 1/3S SUBJ VFIN @FMV ‘do’
como	[como] <rel> <prp> ADV @COM @#AS-<ADVL ‘as’
os	[o] <art> DET M P @>N ‘the’
brasileiros	[brasileiro] <N> N M P @AS< ‘Brazilians [do]’

4b. comparative absolute relative complementiser in postadjectival, postnominal or adjunct-adverbial small clause (AS)

tem	[ter] <vt> <ink> <rH> V PR 3S IND VFIN @FMV ‘[he] has’
um	[um] <arti> DET M S @>N ‘a’
amigo	[amigo] N M S @<ACC ‘friend’
forte	[forte] ADJ M/F S @N< ‘strong’
como	[como] <rel> <prp> ADV **@COM @#AS-A<** ‘like’
um	[um] <quant2> <arti> DET M S @>N ‘a’
urso	[urso] <D> N M S @AS< ‘bear’

The distinction made here between “hooked” relative and absolute relative constructions is based on the presence of a syntactico-semantic "hook" in the non-absolute relatives, like 'cidade' in (1) and 'tanto' in (3). Comparative constructions with 'como' can appear in all 4 sentence types, with both finite and verbless clauses; for hooked AS-relatives (3) they even are the only ones. Another example for a comparison relative is given in (4b)[188].

A strong argument in favour of the existence of relative adverbs in Portuguese is the otherwise almost exclusive[189] use of the future subjunctive tense in relative subclauses. Departing from postnominal (attributive) or absolute nominal relative subclauses ('Seja quem for', 'Podem comprar os livros que acharem interessantes') as the prototypical and uncontroversial case, it seems logical to take future subjunctive inflexion as morphological evidence of a relative reading also in the case of *adverbial* subclauses (like the temporal "relative" in 'me avisem quando ele vier!'), - rather than creating *two* ad-hoc rules with no raison d'être but each other, i.e. (1) calling adverbs for conjunctions if and only if the subclause they head allows future subjunctive tense, and (2) allowing future subjunctive tense outside relative subclauses if and only if these are headed buy conjunctions that possess adverb homonyms[190].

Interrogative adverbs can function as adverbials, adverbial objects, adverbial (intensifier) adjects and complementisers:

1a. interrogative adverbial

Quando	[quando] <interr> ADV @ADVL> ‘when’
partiu	[partir] <ve> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] left’

1b. interrogative adverbial object

Onde	[onde] <interr> <aloc> ADV **@ADV>** ‘where’
mora	[morar] <ink> V PR 3S IND VFIN @FMV ‘[does he] live’
$?	

2. interrogative intensifier adject

Que	[que] <quant> <interr> <det> ADV **@>A** ‘how’
caro	[caro] <jh> <jn> ADJ M S @SC> ‘expensive’
foi	[ser] <vK> <ink> V PS 3S IND VFIN @FMV ‘was [it]’
$?	

[188] For a detailed discussion of comparison structures, including role predication ('trabalhar *como* professor'), see chapter 4.5.2).

[189] The only exception are conditional subclauses with *se* ('if'), where the future subjunctive tense is used to express a future condition: *Se tivermos dinheiro, compraremos* .. 'if we have the money, we'll buy ...

[190] Conjunctions without adverb homonyms, and polylexical conjunctions consisting of a preposition + "que" (e.g. *até que* 'until', *antes que* 'before') ask for *present* subjunctive inflexion in adverbial clauses with a *future* semantic interpretation, suggesting that not only is future subjunctive tense in adverbial subclauses restricted to adverbial complementizers, but also, that its interpretation is one of function (relative) than one of time or, rather, semantically motivated tense (future).

3a. interrogative complementiser in direct object FS

Quis	[querer] <x> <vH> <ink> V PS 1/3S IND VFIN @FAUX ‘[he/I] wanted to’
saber	[saber] <vt> <v-cog> <a+INF> V INF 0/1/3S @IMV @#ICL-AUX< ‘know’
como	[como] <interr> ADV **@ADVL> @#FS-<ACC** ‘how’
venceu	[vencer] <vi> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] won’

3b. interrogative complementiser in argument of preposition FS

a	[a] <art> DET F S @>N ‘the’
discussco	[discussco] <snak> <+sobre> N F S @AS< ‘discussion’
sobre	[sobre] PRP @N< ‘about’
como	[como] <interr> ADV **@ADVL> @#FS-P<** ‘how’
venceu	[vencer] <vi> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] won’

4.5.4.5 Adverb disambiguation and operator adverbs

Operator adverbs (ADV <setop>)

SET OPERATOR **<aset>** apenas, até, não, nem, senão [kun], sequer, somente, só, sobretudo, também, tampouco, mais +NUM
<post-adv> mais, menos, demais *(um bolo mais)*
<+num> mais *(comeu mais dois bolos)*
TIME OPERATOR **<atemp>** ainda, de=novo, em=breve, enfim, já, já=não, mais *(não mais)*, mal
META OPERATOR **<ameta>** simplesmente, obviamente, sobretudo ... + SET OPERATOR

Not all adverbs can appear in all adverbial slots of the Portuguese sentence, and lexical knowledge about which adverbs are allowed where, can be of great use to the CG-rules at the disambiguation level. In fact, when introducing functional subclasses for adverbs, the primarily intended trade-off for the CG grammar was the disambiguation of *other* - non-adverb - categories by providing syntagmatically useful landmarks in the sentence. However, with a fine-grained subclassification, many adverbs are themselves ambiguous as a lexeme, and their worth for the CG-disambiguation of non-adverb categories became interdependent on their own contextual disambiguation, and I have therefore tried wherever possible to functionally define subclasses that can be interpreted meaningfully outside the CG rule formalism, too, - not least in a semantic (or, more restrictedly, MT-oriented) way.

In the preceding chapters, a number of candidate classes for such categorical co-extension of syntactically and semantically defined adverbials has been discussed: Referentially heavy time-, place- and manner-adverbs or -adverbials (circumstantial adverbials, cf. also 4.5.4.2) prefer clause-initial or -final positions and trigger parentheses or commas when they intrude into a valency pattern. Among manner-adverbs, only predicative adverbs are allowed between subject and predicate. Quantifying adverbs (like *mais, menos, muito, imensamente,* cf. 4.5.4.3) function as (usually pre-) modifiers for attributive and adverbial adjects or as adjuncts for the main verb, and they always appear immediately before or after their head. Relative and interrogative adverbs (cf. 4.5.4.4) appear in the clause-initial *complementiser*-position (of either finite subclauses or averbal subclauses).

This chapter treats a syntactically especially intriguing (closed) class of adverbs comprised of what I will call operator adverbs. Some, like logical operators, work on absolute set membership and truth, but many also operate on the relative time or perspective conditions for such predications. I distinguish three classes:

A) Set operator adverbs (e.g. *apenas, não, só, também*)

The adverbs in this group are the only ones allowed to premodify nouns (@>N). They can also premodify non-verbal (B) or verbal (C) predicates. They are disallowed or awkward in adverbial adject or clause-final circumstantial position.

B) Time operator adverbs (e.g. *ainda, de=novo, mal, ?frequentemente*)

These adverbs can premodify predicates, both verbal (like C) and non-verbal. In the latter case they can assume the role of a "pseudo-complementiser", mimicking the role of the subordinating[191] particle in an averbal subclause (@#AS)[192]. In this view, the full clause replaced would be a copula clause, for instance *Ainda em Roma, ... ('Still in Rome, ..')* replacing *Quando ainda estava em Roma, .. ('When he was still in Rome, ...').* This would be structurally analogous to the absolute relative subclause reading for *Quando em Roma, .. ('When in Rome, ...').* Here, however, I prefer the adverbial-premodifier analysis, both (i) because this is in better harmony with the other uses of the time operator class and (ii) because an additional - relative - temporal adverbial ('when') is needed when unfolding the supposed AS.

C) Meta[193] operator adverbs (e.g. *simplesmente*, *obviamente, provavelmente*)

This last group of operator adverbs premodifies whole (verbal) predicates, or even - separated by a comma - whole clauses, but not noun phrases[194] (cf. A) and only rarely non-verbal predicates (cf. B). At the same time it still shares the reluctance of the other operator categories to appear in adverbial adject or clause-final circumstantial positions.

Table: Adverb class and word order

[191] Or co-ordinating, in the case of equalitative comparators *(tão depressa **como** possível),* if one accepts the notion of co-ordinating complementizers.

[192] Another example of both verb- and complementizer-less "clausal" predications is the pattern 'com/sem' + NP + ADVP-loc, like in *'com a mão na bolsa'* (with his hand in his pocket), where the preposition 'com' (or 'sem') functions as "pseudo-complementizer", and 'na bolsa' is predicated about 'a mão', - also in this case without a copula. In contrast with the time operator adverbial case, the preposition 'com' is not only supplemented by a comlementizer in the unfolded clause, but replaced by it: 'while his hand was in his pocket' or ' while he held his hand in his pocket'.

[193] The word is my coinage, another, semantically motivated, term is *attitudinal adverbs.*

[194] Unless they are of the attributive subclass (<attr>), which is occasionaly restricted by meta operators. These cases are, however, functionally non-verbal predicates, and will sometimes even be lexicalized as adjectives, too: *um manifesto obviamente @>A comunista @N<*. One might even argue that *obviamente* here is not the speakers meta-view, but *the way in which* the manifesto is communist, cp. *um manifesto agressivamente comunista*. And of course, also the view-point can be expressed internally rather than externally: *um manifesta religiosamente comunista*. The most "pure" case of a meta-operator with @>A function would probably be: *um manifesto provávelmente comunista.*

position: **adverb class:**	before noun or NP	before non-verbal predicate	before verbal predicate[195]	allows fronting	comma-isolated ad-clause[196]	adverbial adject	clause-final circumstantial position
A) set operator *até*	+	+	2				
B) time operator *já*		+	4	(+)			
C) meta operator *obviamente*		(+)	5	+	+		
intensifier *muito*			(1)[197]			+	+
MANNER *devagar* *de avião*				+			+
TIME/PLACE *hoje*			5	+			+
PREDICATIVE *timidamente*			3	+	+		+

(4a) [Ainda/já/também] no Rio, provavelmente [nem/até/só/também/*ainda/ *de=novo] o Paulo pode chegar muito [cedo/de=manhã/*até/*nem/ *simplesmente].
[Still - already - too] in Rio, probably [not even - even - only - too - still - again] Paul can arrive very [early, in the morning, even, not even, just]

Ainda	[ainda] <setop> ADV @>P ‘still’
em	[em] <sam-> <+top> PRP @ADVL>[198] 'in'
o	[o] <-sam> <art> DET M S @>N ‘-’
Rio	[Rio] <top> PROP M S @P< 'Rio'
,	
provávelmente	[provável] <lex> <setop> ADV @ADVL> ‘probably’
nem	[nem] <ka> <setop> ADV @>N ‘not even’
o	[o] <art> DET M S @>N ‘-’
Paulo	[Paulo] PROP M S @SUBJ> 'Paul'
pode	[poder] <x> V PR 3S IND VFIN @FAUX ‘could’

[195] The numbers show which type of adverb ”wins” in the closeness-contest for pre-predicator position. Meta operators, for example, appear after time/place adverbs, but respect other opertors’ right to be closer to the predicate. A typical sequence would be: ’Ele hoje provavelmente ainda não comeu.’

[196] i.e. related to the clause as a whole, rather than its VP kernel.

[197] Pre-predicate intensifiers have a poetical flair and are rare in colloquial Portuguese, but where they appear, even set-operators ”respect” them: ’Ela provavelmente também muito ama livros.’

[198] Another possible syntactic reading would be @SC>, as subject complement. For now, the parser chooses in such cases the @SC reading for PPs - if there is a copula. If there is none, the adverbial reading is chosen for TIME/PLACE-PPs, and an @SC/@ADVL ambiguity retained for other PPs: *Ela está* ***com @<SC*** *o amigo @P<* - ***No @ADVL>*** *Rio @P<, nunca chove* - ***Sem @SC> @ADVL>*** *o amigo @P<, ela não vem.*

chegar	[chegar] <ve> <vi> <sH> V INF 0/1/3S @IMV @#ICL-AUX< ‘arrive’
muito	[muito] <quant> ADV @>A ‘very’
cedo	[cedo] <atemp> <adj> ADV @<ADVL ‘early’
.	

(4b) Ele [já=não/só/também/simplesmente/*cedo/*de=manhã] quer trabalhar [cedo/de=manhã].
He [not any more - too - just - early - in the morning] wants to work [early - in the morning].

Sentence (4a) boasts all three types of operator adverbials, A ('nem'), B ('ainda') and C ('provavelmente'), as well as representatives of the adverbial quantifier class ('muito') and the circumstantial class ('cedo'). The switch board alternatives in []-brackets sketch the distributional potential of the operator subclasses. (4b) shows the restrictions for the position immediately to the left of verbal predicates, where all operator adverbials are allowed, but circumstantial adverbs - if not prohibited - often are felt as awkward.

CG-rules can exploit the above distributional tendencies in two ways, for "altruistic" disambiguation, i.e. disambiguation of *other* words, or "egoistic" disambiguation, i.e. of the word itself:

(i) on the one hand positions can be defined by their occupants, set operator adverbs, for instance, mark a left group boundary, time operator adverbs a left predicate boundary.

(ii) on the other hand, subclass disambiguation of (often polyambiguous) particles, though not necessarily improving primary (i.e. word class) tagging (*all* subclasses of a given word might be adverbial), can be exploited to make syntactic or even semantic distinctions accessible for parsing at higher levels. Thus, identifying 'senão' as set operator <setop> in pre-nominal but not sentence-initial position, suggests the translation 'only', whereas sentence-initial position favours the conjunctional adverb <ks> reading and should be translated as 'otherwise'.

Two special cases of set operator adverbs are pre- and post-nominal instances of *mais'* ('more') in connection with numerically modified NP-heads, as shown in (5).

(5a)

comprou	[comprar] <vt> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] bought’
um	[um] <card> NUM M S @>N ‘a’
livro	[livro] N M S @<ACC ‘book’
mais	[mais] <setop> <post-adv> ADV **@N<** ‘another [one]’

(5b)

comprou	[comprar] <vt> <vH> <ink> V PS 3S IND VFIN @FMV ‘[he] bought’
mais	[mais] <setop> <+num> ADV **@>N** ‘yet another’
dois	[dois] <card> NUM M/F P @>N 'two'
livros	[livro] N M P @<ACC ‘books’

(5b) fits fairly well into the (A) group, apart from the fact that there is a kind of semantically motivated valency relation with *another* premodifier, a numeral, whereas the other set operators more clearly modify the NP as a whole, and therefore, in dependency notation, its head. However, the adverbial adject link is not as strong as that of 'ao=menos' in *'ao=menos @>A dois @>N livros @NPHR'*, and I therefore use the same syntactic tag as in the other set operator adverbs, i.e. @>N.

(5a) is somewhat different, with the set operator candidate in the unusual post-nominal position. Only very few other adverbs can be classified as "post-adverbs", among them the numeral independent 'demais' (too much, too many), or the deictic 'aí' (*este rapaz aí* - this young lad here). In spite of the positional anomaly, I prefer to explicitly tag for the NP-dependency: @N<.

By the way, as in (5b), without the numeral and with a preceding 'não', a *time* operator reading would be required for 'mais' instead: *'não @ADVL> compra @FMV livros @<ACC mais @<ADVL.'* (he doesn't buy books anymore), and with an adjective to the right 'mais' might be a quantifier adverbial adject: *'nunca @ADVL> compra @FMV livros @<ACC mais @>A caros @N<.'*

4.5.4.6 Adverbial valency or prepositional retagging ?

Some adverbs seem to display a valency structure of their own. A well-known example are comparative structures like *‘mais/menos ... (do) que’* or *‘ tão ... como’*. A few others, like *‘inclusive’* (‘including’), can have nominal arguments, and it is a matter of choice whether to retain the ADV tag (1a') and provide for an NP argument slot, or to recategorise the head in question as a preposition (1a) because it governs an NP:

(1) adverb with nominal argument (valency marked <+NP>):
inclusive, por=exemplo

(1a)

Vieram	[vir] <vt> V PS/MQP 3P IND VFIN @FMV ‘[there] came’
trinta	[trinta] NUM M/F P S @<SUBJ '30'
,	
inclusive	[inclusive] **PRP @<ADVL** ‘including’
os	[o] <art> DET M P @>N ‘the’
estudantes	[estudante] <prof> N M/F P **@P<** 'students'

(1a')

Vieram	[vir] <vt> V PS/MQP 3P IND VFIN @FMV ‘[there] came’
trinta	[trinta] NUM M/F P S @<SUBJ '30'
,	
inclusive	[inclusive] **<+NP> ADV @<ADVL** ‘including’
os	[o] <art> DET M P @>N ‘the’
estudantes	[estudante] <prof> N M/F P **@A<** 'students'

(1b)

emprestou-	[emprestar] <hyfen> <vdt> V PS 3S IND VFIN @FMV ‘[he] gave’
lhes	[lhe] PERS M/F 3P DAT @<DAT ‘them’
inclusive	[inclusive] **<setop>** ADV **@>N** ‘even’
o	[o] <art> DET M S @>N ‘the’
poder	[poder] <am> <ac> <topabs> N M S @<ACC ‘power’
de	[de] PRP @N< ‘of’
curar	[curar] <vi> V INF 0/1/3S @IMV @#ICL-P< ‘healing’

(1c)

a	[a] <art> DET F S @>N ‘the’
lei	[lei] <rr> N F S @SUBJ> ‘law’
vale	[valer] <vi> <sN> V PR 3S IND VFIN @FMV ‘is valid’
para	[para] PRP @<ADVL ‘for’
todos	[todo] <quant1> <enum> DET M P @P< ‘everybody’
$,	
inclusive	[inclusive] **<setop>** ADV **@>P** ‘even’
para	[para] <+hum> PRP @<ADVL 'for'
o	[o] <art> DET M S @>N ‘the’
presidente	[presidente] <prof> N M/F S @P< 'president'

In (1a), *inclusive* heads a PP, and in (1a') an adverbial with an NP-complement denoting what is "to be included". Both have to be distinguished from (1b) and (1c), however, which are semantically different. Here, *inclusive* functions as an operator adverbial, operating on the following constituent, which makes it a prenominal in (1b) and a pre-prepositional in (1c). With the latter meanings *inclusive* belongs in a group with the (potentially prenominal) operator adverbials *até* and *mesmo*.

Similar considerations hold for the semantically antonymous group of exclusives *(salvo, exceto, menos, fora*) as well as *malgrado*. Though traditionally defined as prepositions, an ADV+NP analysis would be defendable, especially since also this group contains members, that homonymously appear as adverbs, cf. *menos* and especially the polylexicals *nao=obstante*, *por=exemplo* and *por=fora*, which differ from all other complex prepositions (*em=vez=de*, *encima=de* etc.) by not even having a preposition as the last element, which would otherwise "guarantee" a @P< reading for the dependent NP also in an analytical (non-polylexical) parse.

For some complex prepositions (*‘antes da festa’*, *'depois de comprar ..'* (2)), the polylexical solution runs into technical problems, since assignment of "word status" is a preprocessor task, and cannot easily be remedied later - failing in cases, where also an analytic reading is feasible, as in ‘falam @FMV antes @<ADVL de @<PIV tempos @P< passados @N<’, where ‘antes’ is a one-word adverb [rather], and ‘de tempos passados’ is a prepositional object. For similar reasons a complex conjunction reading for ‘antes que’ (3) is difficult to manage:

(2) adverb with prepositional arguments (<+PRP>):

antes/depois+de

Depois	[depois] **<+de+INF>** ADV @ADVL> 'after'
de	[de] PRP @A< '-'
(2': Depois=de	*[depois=de] PRP @ADVL> 'after')*
comprar	[comprar] <vt> <vH> V INF 0/1/3S @IMV **@#ICL-P<** '[he] bought'
um	[um] <arti> DET M S @>N 'a'
carro	[carro] <V> N M S @<ACC 'car'
$,	
não	[não] <dei> <setop> ADV @ADVL> 'not'
anda	[andar] <vi> <vH> <ink> V PR 3S IND VFIN @FMV '[he] goes'
mais	[mais] <setop> ADV @<ADVL 'any more'
em	[em] <+top> PRP @<ADVL 'by'
bicicleta	[bicicleta] <V> N F S @P< 'bicycle'

(3) adverb with a finite subclause argument (<+que>): antes/depois+que

Se	[se] <refl> PERS M/F 3S/P ACC/DAT @ACC> '-'
combinaram	[combinar] <vt> <ink> V PS/MQP 3P IND VFIN @FMV '[they] agreed'
antes	[antes] **<+que>** PRP @<ADVL 'before'
que	[que] KS @SUB **@#FS-P<** '-'
(3': antes=que	*[antes=que] KS @SUB @<ADVL 'before')*
fossem	[ser] <x+PCP> V IMPF 3P SUBJ VFIN @FAUX '[they] were'
separados	[separar] <vt> V PCP M P @IMV @#ICL-AUX< 'separated'

In all, one can conclude that candidates for valency bound adverb complements are few, heterogeneous and lexically idiosyncratic. Distributionally, apart from the semantically and syntactically unique comparatives, such adverbs either *function* like/as prepositions (heading NPs or completive que-clauses, *salvo, antes que*) or form part of lexically fixed expressions *together with* a preposition (*antes de*)[199], which is why I tend to conclude that a preposition reading should be favoured for all of the above cases, if the class be treated in a homogeneous way, - possibly by filtering the output of the parser *after* disambiguation.

[199] The adverbial phrase 'relativamente ao filme' as cited by Perini (1989, p.181) is a possible counter-example, being the adverbial analogon of the adjective phrase 'relativo ao filme' where 'ao filme' *is* an argument adject. If treated in the same way as 'antes de' and 'depois de', the term 'relativamente a' should, however, be regarded as a complex preposition ('em relação a'?) - which doesn't seem entirely satisfactory on the background of the related and argument-containing adjective phrase.

4.5.5 Violating the uniqueness principle: Reflexive impersonality structures

The word class ambiguity of the Portuguese word form "se", between subordinating conjunction on the one hand, and personal pronoun on the other, is one of the hardest tasks the morphological disambiguation grammar has to confront, involving many rules specifically written for this word form only. To make things worse, the ambiguity is interdependent with a characteristic ambiguity in its corresponding verb - that between the infinitive and finite (future subjunctive) readings[200], of which the first matches the pronoun reading, and the second, the conjunction reading of "se". And for the former, the woes of ambiguity even continue on the syntactico-functional level.

Traditionally, the pronoun "se" is regarded as *reflexive*, surface-syntactically implying a direct object reading. In terms of valency, reflexivity can - for one - simply match the syntactic pattern of the corresponding monotransitive usage, as in English "He hates *himself*" ('Se detesta'), or - with a plural verb form - as in the reciprocal "They love *each other*" ('Se amam'), where Portuguese can use the reflexive instead of the literal *um ao outro*. In these cases, both subject and object receive "real" thematic roles: The subject of such sentences functions as agent, and the direct object as patient. Like German and Danish, but unlike English, Portuguese can, however, "integrate" the reflexive pronoun into the verbal lexeme to such a degree that no clear thematic function can be assigned to it (so-called pronominal verbs). This is readily apparent where the subject of the pronominal verb form lacks the agent-feature of the monotransitive form, and instead displays the thematic role of patient itself, not leaving to the reflexive object any meaningful lower function along the hierarchy of thematic roles. In the cline below, (2a-b) are examples of analytic reflexives with "real" objects (testable by the addition *a si mesmo* ['himself'][201]), and (2c-e) are process or event reflexives where the object is void of thematic function. In the zero-subject construction (2f), finally, neither subject nor direct object receive thematic roles, and the second, prepositional, object functions as patient (theme).

		@SUBJ	**@ACC/se**	**@PIV**
(2a)	mata-se ('he kills himself)	AG	PAT	-
(2b)	lava-se ('he washes [himself])	AG	PAT	-
(2c)	habitua-se a ('he grows accustomed to')	PAT	?	-
(2d)	torna-se ('he/it becomes')	PAT	?	-
(2e)	passa-se ('s.th. happens')	PAT	?	-
(2f)	trata-se de ('it is about')	-	?	PAT

[200] The same verbal ambiguity is seen in conjunction with relatives/interrogatives, cp. 4.5.4.4.

[201] Yielding: *mata-se a si mesmo* and *lava-se a si mesmo*.

While (2c-e) can be described as process or event verbs[202] without explicitable agent, and the agent in (2a-b) is in the subject, it is unexpressed but explicitably non-subject in (3a-c):

(3a) Collor pode eleger-se deputado (pelos amigos). ('Collor may be elected MP.')
(3b) Derrubam-se as estátuas. ('Statues are [being] overthrown.)
(3c) Cobram-se mensalidades .... ('Monthly fees are charged.')
(3d) jamais se soube como ... ('it was never known how ..')

Here, *Collor* is subject, but <u>not</u> agent, and the subjects in (3b) - (3d), being inanimate, are not even semantically potential agents. In (3d) the problem is slightly different - if *se* is to be a (reflexive) direct object, then the other argument of the cognitive verb must be +HUM, which the interrogative clausal FS *como* ... clearly is not.

Rather, in all of (3), the patient role is represented by both subject and object. I would like to argue that "se" in (3) is not reflexive at all, neither analytically nor lexically. One solution for saving the uniqueness principle with regard to thematic roles in (3a) is to opt for a condensed causative matrix reading ('Collor can let/make himself be elected MP'), which is not, however, especially self-evident in the surface-structure of the *Portuguese* sentence. *Estátuas, mensalidades* and *como ...* , in (3b-d), are not even semantically capable of being agents, so a causative reading would not be an option anyway. Rather, all four examples can be described as (reflexive) passives[203], where the "reflexive" pronoun functions somewhat like the '-s' morpheme in the Scandinavian "reflexive" *synthetic passive* ('Månedlige gebyrer opkræve**<u>s</u>**.'), though "se" in the Portuguese construction is not a bound morpheme, and not even obligatorily enclitic[204]. In my parser I have therefore chosen a hybrid tag, @ACC-PASS, that at the same time satisfies surface syntactic necessities (i.e. having a [surface-] direct object for obligatorily transitive verbs) and the pseudo-morphological function of passivisation.

Thus, due to the @ACC function tag, both true reflexives (2) and passive constructions (3) are more or less in harmony with the morphological accusative case tag of *“se”*:

[202] I am here applying a simple two-feature typology, that is also used in my lexicon designate the semantic class, of deverbal nouns (cp. 6.3.2):

	imperfective	*perfective*
+ Agent-subject, + Control	activity <CI>	action <CP>
- Agent-subject, - Control	process <cI>	event <cP>

[203] In the context of this chapter, the term (reflexive) passive will be applied to instances where pronominal *se* roles as PATIENT, with no AGENT subject present or implied by the main verb's subject selection restrictions, and with number agreement permitting a non-subject reading for *se.*
[204] Diachronically, enclitic pronouns are on the retreat in Brazilian Portuguese.

for (2): *se "se" PERS M/F 3S/P **ACC**/DAT **@ACC***
for (3): *se "se" PERS M/F 3S/P **ACC**/DAT **@ACC-PASS***

There are, however, quite a few cases where either the verb's valency or the uniqueness principle make an @ACC reading rather unwanted. Consider the following sentences and their disambiguation possibilities concerning the pronoun "se":

(4a) não **se** @SUBJ> está familiarizado @<SC ('One isn't used to it.)
(4b) está-**se** @<SUBJ diante @<SC de um monstro ('we are facing a monster')
(4c) chega-**se** @<SUBJ ao @<ADV Primeiro Mundo ('you reach the First World')
(4d) tem-**se** @<SUBJ a impressão @<ACC que ...('one has the impression that ..')
(4e) o @ACC> empresta-**se** @<SUBJ a esta pergunta @<PIV a maior veemência @<ACC ('they give this question highest priority.')

In (4a-b) the intransitive verb *estar* has no valency slot for a direct object, and the presence of subject complements (*familiarizado* and *diante de um monstro*) make a subject reading a tempting alternative. In (4c) the ergative verb *chegar* does not invite an @ACC reading for *se* for similar reasons.

Assigning *se* subject function in these cases, creates what from a semantic point of view could be called an impersonal (or indeterminate) personal pronoun (Danish *'man'*, English *'they'*, *'one'*), since such a *se* does not provide anaphoric clues like other personal pronouns. The fact that the addition of an explicit (non-*"se"*) subject candidate to the examples in (4) yields in all cases ungrammatical sentences (like 4b: **ele está-se diante ...*, 4c: **ele chega-se ao Primeiro Mundo*, 4d: **ele tem-se a impressão que ..*, 4g: **ele costuma-se oferecer outros benefícios*), and that it is hard to find in the corpus examples of explicit (non-se) +HUM subject candidates in type (4) sentences (i.e. sentences with pronominal *se* and no free @ACC or @DAT valency slot for it to fit in), suggests that the subject slot is, in fact, already occupied - by *se*, that is, as other subject candidates in the examples are ruled out by morphosyntactic form, selection restrictions and the like. The lack of explicit "non-*se*"-subjects also makes yet another alternative - an indirect object[205] (@DAT) reading for *se* - a rather artificial and "patchy" solution, since a (non-impersonal!) dative reflexive pronoun would naturally want to refer to some anaphoric hook *in the same sentence* in at least some of the cases.

[205] Incidentally, it was such @DAT readings my early parser (prior to the introduction of the @ACC-PASS and @SUBJ choices for the pronoun *se*) used to suggest "all by itself" where the uniqueness principle and valency rules removed from the tag-line of *se* its (then) only other morphologically mappable syntactic reading, @ACC, - the reason being that a Constraint Grammar always leaves one reading, even if this means that a tag is chosen *via negationis.* Cf. the examples in (9) later in this chapter for a further discussion of *se @DAT*.

In (4d) *ter* ('have') does have monotransitive valency, but belongs to a verb class that cannot be passivised[206] or reflexivised to yield a passive or reflexive meaning *(*muito dinheiro é tido [por ele], *muito dinheiro se tem, *muito dinheiro se tem a si mesmo)*. Also, and disambiguationally more important, there is one @ACC candidate already, exerting its prohibitive influence by means of the uniqueness principle. In (4e), which is the kind of corpus jewel that makes the uniqueness principle (and many generative grammars) sweat blood and tears, there are even three direct object candidates, the object pronouns *o* and *se*, as well as the NP *a maior veemência*. Even if one concedes the NP the status of post-positioned subject, *o* is so strong a direct object, for morphological reasons, that a direct object *se* is in trouble with the uniqueness principle. With a subject-*se*, however, *o* can be interpreted as a kind of place-holder for the bigger NP-object later in the sentence, a technique not entirely uncommon in Portuguese.

Another context where neither a reflexive nor a passive reading make sense, are verb chains where a matrix verb governs a non-finite clause and "se" is linked, either by fronting (4g) or by hyphenation (4f), to a matrix verb demanding a +HUM subject (a condition clausal subjects can't comply with).

(4f)

Costuma-	[costumar] <hyfen> <x> <vH> V PR 3S IND VFIN @FAUX ‘usually’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@<SUBJ** ‘they’
oferecer	[oferecer] <vr> <vH> V INF 0/1/3S @IMV @#ICL-AUX< ‘offer’
outros	[outro] <diff> <KOMP> DET M P @>N ‘other’
benefícios	[benefício] <CP> N M P @<ACC ‘benefits’
$.	

(4g)

Talvez	[talvez] <+SUBJ> <ameta> ADV @ADVL> ‘maybe’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@SUBJ>** ‘one’
precisa	[precisar] <x> <vH> V PR 3S IND VFIN @FMV ‘needs’
acumular	[acumular] <vr> <sH> <sN> V INF 0/1/3S @IMV @#ICL-<ACC ‘to gather’
talento	[talento] <fhc> N M S @<ACC ‘talents’
$.	

The *obligatory* +HUM feature that unavoidably crops up with most of the *se*-cases in (4), both in subject selection restrictions and in translation, is by itself a strong argument in favour of the existence (or at least, ongoing evolution) of an independent

[206] 'Ter' is polysemic and can also occur with ditransitive (transobjective) valency, in *ter alguém como* ('regard someone as ...'), where passivisation is possible. The *como*- argument is, however, obligatory in this case, and I would therefore argue, that the case is irrelevant for the treatment of ordinary *ter*. The same holds for the few reflexive uses: *ter-se com* ('confront sb.') and *ter-se em* ('stay in').

Portuguese pronoun equivalent to French *'on'*, Danish *'man'* etc., which exhibit the same features: +HUM and third person singular (for a number discussion, cp. 5e and 5f).[207]

One might ask, why the impersonal pronoun analysis can't be applied to the "passive" constructions in (3) as well. While this might appear a tempting simplification when looking at singular 2-argument sentences like *vende-se um carro* ('a car is for sale'), two tests prove that this is not a viable alternative in the specific cases: First, making *se* the subject and *carro* the direct object - as one would in an impersonal analysis -, does not work in the plural (*vende**m**-se carro**s*** and 3b-c), as the number agreement relation of the verb shows - it clearly prefers *carros* as subject[208]. Second, since the object complement *deputado* in (3a and 3a') obligatorily needs a direct object, *se* cannot be subject, because it is the only object candidate in

(3a') pode eleger-se deputado

The fact that *Collor* can be omitted in (3a') is by itself evidence for that it is subject and not object, since Portuguese subjects are optional, but objects in the presence of object complements are not.

In general, surface-existing pre-positioned subject candidates (@SUBJ>) tend to prohibit the impersonal @SUBJ reading for "se" (French *'on'*, Danish *'man'*), since this would make the first NP a pre-positioned direct object (@ACC>) and imply either OSV word order (with pre-positioned "se") or (with enclitic "se") OVS word order, both of which are very rare in noun-caseless[209] Portuguese, and thus unlikely even without a number agreement problem:

(5a) o delta @SUBJ> **se** @ACC-PASS> divide em ..
('the delta is divided into ...')
(5b) nada @SUBJ> **se** @ACC-PASS> compare a ...
('nothing can be compared to ...')
(5c) o fim @SUBJ> da ditadura celebrou-**se** @<ACC-PASS
('the end of dictatorship was celebrated')

[207] Like 'on' in French and 'man' in German, impersonal *se* is the only third person pronoun in Portuguese that is obligatorily +HUM (not counting *você* , which is an etymological NP: *Vossa Mercê*). Though *chegar* ('to arrive') as such allows both +HUM and -HUM subjects, the sentence *chegou-se* ('they arrived') tests negative as a rewording of *chegaram todos os livros* ('All books arrived'), and positive for *chegaram todos os amigos* ('All the friends arrived).

[208] Though Portuguese grammars sometimes do cite the alternative singular form *vende-se carros*, I have not been able to verify the existence of such a construction in my *Brazilian* corpus for post-positioned "se". Cp. (5e) for pre-positioned "se".

[209] The only words capable of bearing an accusative feature in Portuguese are object pronouns like se itself, which would make the function of a fronted direct object clear, but seldom occur together with *se* - (4f) is a rare example -, maybe because the morphological case marking makes the uniqueness principle problem more "palpable".

When the potential subject is post-positioned, however, the impersonal reading is a powerful alternative, because it embodies the MORE "natural" VSO (5d') as opposed to a VOS word order in the passive reading (5d):

(5d) celebrou-**se** @<ACC-PASS o fim @<SUBJ da ditadura
('the end of dictatorship was celebrated')
(5d') celebrou-**se** @<SUBJ o fim @<ACC da ditadura
('one/they celebrated the end of dictatorship')

As one might expect from word order considerations (i.e. the normalcy of SVO as opposed to OVS), there is some indication in my corpus that the tendency towards reading "se" as subject is strongest if it is pre-positioned (non-hyphenated), as in (5e) where there is number agreement evidence for the subject reading. Number agreement, or rather, the lack of it, can, of course, force the subject reading in (hyphenated) *post*-position (5f), too, though such instances[210] seem to be few and special. Thus, in (5f), due to the interfering PP, there is no direct "clash" between the singular V+se entity and the plural NP candidate for either @ACC or @SUBJ role. It may be that such a clash would provoke agreement in conservative speakers, i.e. a plural verb form and a consequently facilitated subject reading for the clause final NP.

(5e)

Quando	[quando] <rel> <ks> ADV @ADVL> @#FS-ADVL> ‘when’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@SUBJ>** ‘one’
compara	[comparar] <com^vtp> <vH> V PR 3S IND VFIN @FMV ‘compares’
os	[o] <art> DET M P @>N ‘the’
evangelistas	[evangelista] <attr> N M/F P **@<ACC** 'Evangelists'
com	[com] PRP @<PIV ‘to’
outras	[outro] <diff> <KOMP> DET F P @>N ‘other’
fontes	[fonte] <topabs> <sfP> N F P @P< ‘sources’

(5f)

Brande-	[brandir] <hyfen> <vt> <sH> V PR 3S IND VFIN @FMV ‘braces’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@<SUBJ** ‘one’
contra	[contra] PRP @<ADVL ‘against’
os	[o] <art> DET M P @>N ‘the’
joguinhos	[joguinho] N M P @P< ‘toys’
os	[o] <art> DET M P @>N ‘the’
espectros	[espectro] <HM> <topabs> N M P **@<ACC** ‘specter’
de	[de] <sam-> PRP @N< 'of'

[210] An entirely speculative explanation for this might be that both the fronting of object pronouns and the use of *se* as impersonal pronoun are, in a historical perspective, grammatical innovations (with at least the first being far more common in Brazil than in Portugal), so conservative speakers might have a tendency to both use more hyphenated (i.e. postpositioned) *se*'s and obey the number agreement rules for reflexive constructions that prevent *se* from achieving real (singular) impersonal pronoun status.

```
a           [a] <-sam> <art> DET F S @>N ‘-’
violência   [violência] <am> N F S @P< ‘violence’
$.
```

A special case of "weak" pre-positioned subject candidates are relative pronouns which often do function as @ACC>. Here, both analyses seem to be acceptable:

(5g) **o que** @ACC> **se** @SUBJ> deve fazer ('what one has to do')
(5g') **o que** @SUBJ> **se** @ACC-PASS> deve fazer ('what has to be done')
(5h) hoje, **o que** @ACC> **se** @SUBJ> busca, é um emprego ..., ('today, what one is looking for, is a job ...,')
(5h') hoje, **o que** @SUBJ> **se** @ACC-PASS> busca, é um emprego ..., ('today, what is being looked for, is a job ....,')

In subclauses headed by (comparative) adverbial relatives, the case is more simple - and biased in favour of an impersonal subject reading. Here, the direct object is always omitted (i.e. "implied" by this type of adverbial complementiser) even with obligatorily transitive verbs (6a-b). In the English translation, a *what*-object has to be added:

(6a) ..., conforme dizem, .. ('according to what they say')
(6a') ..., conforme eles dizem, .. ('according to what they say')
(6b) ..., segundo denunciou na época, .. ('according to what he said at the time')

Applying the uniqueness principle, we are then left with a natural subject reading for "se" in (6c). An @ACC-PASS reading is ruled out by the agrammaticality of surface objects in these constructions (6c'), and the @SUBJ reading for "se" is further backed by the fact that no other explicit subject can be added (6c") - in contrast with *se*-less (6a') -, suggesting that the subject slot is in fact filled by *se* (and, implicitely, that a non-subject *se* would in this context be understood as direct, not *indirect,* object for *estima*, producing agrammaticality for the same reasons as 6c').

(6c) conforme/segundo se @SUBJ> estima ('according to what one estimates')
(6c') *conforme/segundo o @ACC> estima ('*according to he estimates it')
(6c") *conforme/segundo ele @SUBJ> se estima ('*according to what he estimates himself')

In table (1), I have inspected 288 running instances of pronominal "se" from a newspaper corpus, applying the different categories defined earlier in this chapter. The distribution shows that the "prototypical" reflexive usage is still the most common (about two thirds of all cases), while passive readings are rare, and the impersonal pronoun use is as frequent as 1 in 5.

(1) **Table: functions of pronominal “se”**

usage	reflexive		passive		impersonal (indeterminate)		reciprocal		all	
position	pre	post	pre	post	pre	post	pre	post	pre	post
VFIN	73	58	20	8	35	25	2	1	130	92
INF	27	25	1	5	1	-	2	-	31	30
GER	1	3	-	-	-	-	1	-	2	3
all	101	86	21	13	36	25	5	1	163 **(57%)**	125
	187 **(65%)**		34 **(12%)**		61 **(21%)**		6 **(2%)**		288 **(100%)**	

The text in question being Brazilian, it comes as no surprise that pre-posed pronouns ("pre") are more frequent than enclitic ones ("post"), accounting for 57%. Interestingly, non-finite forms are more conservative in this respect than finite ones, with fronting percentages of 50% and 59%, respectively.

And non-finite forms are different in yet another aspect: They (almost) never occur with impersonal "se", and in passive "se" constructions they avoid pronoun fronting (almost) altogether. A rare but typical context, where indeterminate "se" does occur before non-finite verb forms, are clausal arguments of prepositions:

(7a) a capacidade **de** @N< se @SUBJ> tirar @#ICL-P< proveito dela
('the capacity to take advantage of her')
(7b) **além=de** @ADVL> se @SUBJ> misturar @#ICL-P< folhos, ...
('apart from mixing leaves')
(7c) o plano **de** @N< se @SUBJ> montar @#ICL-P< uma rede de lavenderias
('the plan to mount a network of washing saloons')
(7d) **A** @ADVL> se @SUBJ> julgar @#ICL-P< pela reação que despertou, ...
('to judge by the reaction he provoked')

Statistical findings like the above can be useful on the heuristic level of disambiguation, after all other lexical and contextual information has been exhausted. So far the following disambiguation principles have been discussed for the pronoun "se":

- lexically marked reflexive valency favours the prototypical <refl> @ACC reading
- lack of reflexive valency suggests @SUBJ or @ACC-PASS readings, the latter only for transitive valency
- an explicit subject-NP candidate to the left disallows se-@SUBJ
- a plural verb form prohibits se-@SUBJ

- lack of a slot for direct object (<vt>) or reflexive object (<vr>) in the verb's valency inventory suggests se-@SUBJ

Apart from valency, agreement and word order, another - fourth - type of information can be useful:

Since the impersonal se-@SUBJ *per defitionem* is assigned the semantic feature +HUM (cp. the discussion of (4)), it can be ruled out for verbs without the <vH> subject selection restriction, which is marked in the lexicon. For transitive <vt> verbs obligatorily marked <vH>, but without a <vr> valency reading, in principle both a se-@SUBJ and a @ACC-PASS reading are feasible. The latter will win even in fair battle (i.e. with non-specific CG rules), if:

- there is no NP-candidate to fill the verb's @ACC valency slot
- the verb is in the plural

and it would enjoy a strong word order bias where

- an NP @ACC-candidate is pre-positioned.

My intuition after some corpus-inspection is that the border line necessary to delimit the remaining cases of se-@SUBJ in transitive verbs *with* direct object candidates is too soft to make disambiguation practicable. I therefore prefer to make my parser tag, as a default, *all* instances of "se" with transitive non-reflexive verbs as @ACC-PASS, and only choose @SUBJ if number agreement, the uniqueness principle or a special clausal structure (6c) enforce this solution. Undecided cases of pronominal *se*, like those discussed in (5g-h) or (5d) will thus be resolved by the default rule as reflexive passives. Since passivisation (both analytic and synthetic) is a fairly universal metaphor for indeterminacy in Indo-European languages, this approach seems to be tenable from a semantic and MT-point of view, too:

(8a)

```
Corta-     [cortar]  <hyfen> <vr> V PR 3S IND VFIN @FMV 'cut is'
se         [se] <refl> PERS M/F 3S/P ACC/DAT @<ACC-PASS '-'
a          [a] <art> DET F S @>N 'the'
cabeça     [cabeça] <anmov>N F S @<SUBJ 'head'
de         [de] <sam-> <+hum> PRP @N< 'of'
o          [o] <-sam> <art> DET M S @>N 'the'
rei        [rei] <title> N M S @P< 'king'
$,
fuzila-    [fuzilar] <hyfen> <vt> <vi> <vH> V PR 3S IND VFIN @FMV 'shot is'
se         [se] <refl> PERS M/F 3S/P ACC/DAT @<ACC-PASS '-'
o          [o] <art> DET M S @>N 'the'
```

ditador	[ditador] <prof> N M S **@<SUBJ** ‘dictator’
$,	
derrubam-	[derrubar] <hyfen> <vt> <vH> V PR 3P IND VFIN @FMV ‘brought down are’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@<ACC-PASS** ‘-’
estátuas	[estátua] <cc> <HM> N F P **@<SUBJ** ‘statues’
$.	

(8b)

em	[em] <sam-> PRP @<ADVL 'in'
um	[um] <-sam> <arti> DET M S @>N 'an'
exemplo	[exemplo] <ac> <+de+interr> N M S @P< 'example'
de	[de] PRP @N< 'of'
como	[como] <interr> ADV @ADVL> @#FS-P< ‘how’
se	[se] <refl> PERS M/F 3S/P ACC/DAT **@ACC>-PASS** ‘-’
constrói	[construir] <vt> <vH> V PR 3S IND VFIN @FMV ‘is built’
a	[a] <art> DET F S @>N ‘-’
impunidade	[impunidade] <am> N F S **@<SUBJ** ‘impunity’

Note that *corta-se* in (8a) is marked for reflexive valency, too, suggesting the default @ACC tag for *se* in this case. However, if used in the active voice, *cortar* subcategorises for a +HUM subject. Since *cabeça* is -HUM, only a passive reading makes sense. Compare the "active voice" sentence *O rei @SUBJ> cortou-se @<ACC com uma faca* ('the king cut himself with a knife). Another problem is that *fuzila-se* in (8a) could, in another context, well be read as ambiguous, - it might be the dictator shooting himself (se-@ACC). Without the context of analogous constructions in the same sentence, as in (8a), only word order could be used to make the distinction, as an agent-subject statistically is more likely to precede the corresponding verbal constituent.

Besides its reflexive, passive and impersonal (indeterminate) uses there is one last functional interpretation applicable to the pronoun *se* - that of dative object, or, in terms of thematic role, "benefactive", a function which can be tested by substitution with the morphologically unambiguous dative object pronoun *lhe*. Dative/benefactive *se* is fairly rare, and from a statistical point of view it might well be defendable to ignore it altogether. Still, in my parser I do map also the @DAT function onto "se", imposing a heavy extra burden on the disambiguation rules. Consider the following examples:

(9a) **Se** @DAT> arrancou uns cabelos @<ACC [a si mesmo].
('He pulled out one of his hairs.')

(9b) Ele @SUBJ> **se** @DAT> comprou um carro @<ACC.
('He bought himself a car')

(9c) Os pais @SUBJ> preocupados devem **se** @DAT> lembrar que @#FS-<ACC só é um jogo.
('Worrying parents have to remind themselves that it is but a game.')

A typical case is *arrancar* ('to pull out') in (9a), where the @DAT function simply fills the semantically pre-existing valency slot of "affected" (here a negative variant of "benefactive": 'to pull out from') - which can be tested for its "dativeness" by substitution with a predicative object headed by the preposition *a (arrancar-lhe ac. - arrancar ac. a alg.)*, in analogy with other dative governing ditransitives (e.g. *dar* 'give'). Such cases can be marked <vdt> (or <a^vtp>) in the lexicon, and the only special feature of the *se*-constructions concerned, is that subject and dative object coincide, rather than subject and direct object, as usual in reflexive constructions.

Lembrar (9c), too, has a valency slot for "affected" or "benefactive" ('to remind s.o. of s.th.') and allows alternation between dative *lhe* and a PP headed by the preposition *a* (*lembrar-lhe ac. - lembrar ac. a alg.*), - though it can also govern a *"de"*-prepositional object on top of reflexivisation (<de^vrp>), as in *devem se @ACC> lembrar do @<PIV jogo* or *devem se @ACC> lembrar de @<PIV que só é um jogo.* The direct (i.e. "prepositionless" as in 9c) attachment of a *que*-clause, however, creates problems with the uniqueness principle, since both reflexive pronouns and object clauses are tagged @ACC. The dative tag, in this case, is also a way of saving the uniqueness principle (which is of central importance to the disambiguation rules) without adding elliptic zero elements (here, the preposition *de* would be needed to match a normal valency pattern.

(9b), finally, is an example of what one might call *optional benefactive dative*. Here, the dative object is not valency bound in a the "normal" syntactic way: Though "se" can be replaced by morphologically unambiguous dative forms of other personal pronouns *(ele **lhe** comprou um carro)*, its @PIV replacement is headed by the preposition *para*, not *a* (which is the default for valency bound dative objects), and the VP-isolation test is positive:

(9b') Comprou um carro **para** Pedro (para si mesmo).
O que fez **para** Pedro (para si mesmo)? - Comprou um carro.
(9a') Arrancou um cabelo **a** Pedro (a si mesmo).
O que fez **a** Pedro (a si mesmo)? - *Arrancou um cabelo.
(9c') Devem lembrar **a** Pedro (a si mesmo) que só é um jogo.
O que devem fazer **a** Pedro (a si mesmo)? - *Lembrar que só é um jogo.

Note that the question-answer pairs for the tighter bound cases (9a') and (9c') are ungrammatical for *neutral* que ('to do') and *half-VP*-answers - understood as *o que fez contra Pedro* ('what [malice] did he do to Peter'), *full-VP*-answers are permissible: *arrancou-lhe um cabelo, lembrar-lhe que só é um jogo.*

In other contexts, *para* is not usually valency bound either, at least not as such[211]. Like the benefactive in (9b), it appears to be *semantically* governed when it fills the direction slot of adverbial complements of verbs like *ir* ('go'), *viajar* ('travel'), allowing for alternation with other directional prepositions like *a* and *até*. In the case of (9b), the benefactive seems to match an inherent thematic role slot of a whole group of "aquirance" verbs. While the semantic anatomy of ordinary <vdt> ditransitive verbs ('to give', 'to send') can be described as CAUSE HAVE, "aquirance" verbs ('to take', 'to buy') seem to display a BECOME HAVE structure, the difference being reminiscent of that between the causative and inchoative/ergative aspects (to kill - to die). In the first, change is complemented of the object, in the second, of the subject. This explains why "se" is so loosely bound in (9b) - *if* the subject thematic role - as normal for ergatives - were that of patient, no pronoun would be needed (as in other verbs of the BECOME HAVE group: 'to get', 'to receive'). However, "aquirance" verbs subcategorise for *agent* subjects, which makes MAKE (BECOME HAVE) a better semantic dissection, and explains why there is a certain potential for the surface manifestation of a patient role - as when *se*-@DAT is added in (9b). In Danish, the distinction can be made in much the same way: 'få en øl' (= 'to get a beer') can be made "active" by adding the reflexive pronoun *sig*: 'få **sig** en øl' (= 'to go and get a beer').

Technically, in terms CG-rule dynamics, the distinction of 4 different "functions" of pronominal *se* has proven quite effective, preventing uniqueness principle and valency pattern based rules from making wrong argument choices with regard to *other* constituents, especially subject and direct object NPs in clauses containing pronominal *se*. Thus, while somewhat controversial in its linguistic interpretation and somewhat dynamic in its distinctions, and in spite of introducing new ambiguities (and resulting, new, *“se-internal”* errors), the more fine-grained approach has helped improve - all other things equal - *overall* parser performance (as compared to the original, purely morphologically based, case mapping approach with only @ACC and @DAT readings)[212] With regard to the functional tagging of pronominal *se* itself, however, the use of subcategories like impersonal (indeterminate) subject and reflexive passive may well create benchmark and correctness measurement problems, which are being circumvented by the optional use of a post-parser filter program fusing the above distinctions into a functionally underspecified Portmanteau tag, @REFL, - in much the same way pronominal *se* is left underspecified with respect to (morphological) case (ACC/DAT) on the morphological tagging level.

[211] It may be valency bound, as a preposition that can head a - valency bound - directive adverbial.

[212] Of course, as a computational linguist, I am inclined to think that the disambiguational usefulness of a category may well by itself be regarded as an indication that the category in question is not entirely without a structural base, but reflects - to a certain degree - corpus reality and system "uncontradictoriness".

4.6 The transformational potential of CG

4.6.1 Exploiting word based tag notation with string manipulation tools

Technically, a Constraint Grammar rule can be regarded as a context dependent string manipulation command to be executed by a computer on an ambiguously tagged sentence. A whole sentence with all its words and tags could be written as one long line of text, and a given CG rule could then be rewritten as an awk or perl language substitution rule (s/.../..../g), using so called regular expressions to express optional or dummy string segments, or to bracket-mark a string of characters for conditioned deletion, repetition or movement. Thus, information from different levels of analysis (morphology, syntax, semantics etc.), both form and function, can be represented in the same formalism, and interact in transparent, string-based disambiguation process.

It seems only logical, after disambiguation, to go a step further and exploit the text-tool friendliness of the tagging notation for other purposes, like corpus searches, information extraction, IT-based grammar teaching and the like. All of the application examples mentioned are about identifying, extracting and standardising string chunks from a text context. Common UNIX tools like grep in conjunction with substitution commands will do the job, and are, in fact, used at the applicational front end of my parser. In theory, however, the CG-formalism itself could be applied to the same end. The *mapping* operator, for example, could be used to mark corpus occurrences of certain linguistic patterns (1a), which could then be extracted by a chained grep-command. A *replace* operator, as suggested in (Tapanainen, 1996), though non-existent in the cg1-compiler and inflexible in cg2, could complement the *mapping* operator and be used for notational standardisation (1b).

(1a) MAP (@EXAMPLE) TARGET (@#ICL-SUBJ>) (*-1 @SUBJ> BARRIER NON-ADV LINK 0 N) ; # Find an example of: a non-finite subclause functioning as subject that itself has a subject noun, preceding the main verb with nothing but adverbs in between

(1b) REPLACE (KS) TARGET ("quando" <rel> ADV) IF (0 @#FS) ; # Replace the relative adverbial reading of "quando" by a subordination conjunction reading if the word heads a finite subclause

4.6.2 Theory dependent tag filters

One of the most recalcitrant problems of grammatical analysis, in both corpus annotation and grammar teaching, is - from a practical point of view - the simple fact of life that the work of different grammatical schools working on the same language, is notationally more or less incompatible. While initiatives like the European EAGLE project do suggest minimum standards of distinctional complexity (for example with regard to word classes), a large degree of theory-dependent "idiosyncrasy" remains. Such - notational - idiosyncrasy is not, of course, entirely arbitrary, but, for each different school of thought, a logical result of the application of that particular school's general principles on individual problems, and the quality of a system must always be measured against what goals it is to serve and which practical uses (e.g. explanatory, pedagogical, informational, translational) it is going to be put to. Thus notational details are largely dictated by (internal) consistency and (external) adequacy. From the technical point of view of automatic analysis, an extremely interesting question is: For grammatical systems tackling roughly the same overall topic (say, syntax), - how much information is lost, and/or how much information must be added when "translating" one notation into another? Consider the following questions and their theory dependent, notationally different, answers:

(1a) Is "proper noun" a PoS word class or just a semantic distinction?
(1b) Is "do que" a syntactic entity or a word?
(1c) Is the distinction "interrogative" - "relative" for the adverb *quando* semantic or syntactic, and is it still an adverb when used "conjunctionally" to head a finite subclause?
(1d) Is *to Peter* in "He has sent a letter to Peter" an adjunct adverbial, valency bound adverbial object, prepositional object or postnominal modifier?
(1e) Is *a letter* in the same sentence a VP constituent or a clause constituent? And if it is a VP constituent, is *to Peter* one, too? And what is a VP anyway, a verb chain ("continental"), the predicate ("Anglo-Saxon") or verb+subject+objects ("Portuguese") ?
(1f) Is *has* in the verb chain "has sent" head (as in dependency grammar) or dependent (as in some constituent analyses) of *sent*, or neither (but rather a clause constituent, as in Chomskyan generative grammar) ?

Apart from distinctional complexity (1d) - which is simply about information loss or gain -, systems also disagree on which level of analysis a certain categorical distinction belongs (1a-c). More difficult to treat than distinctional and level-incompatibility is *structural* notational incompatibility (1e-f), since here, one can not simply relabel a given grammatical unit, but may have to use smaller, larger or structurally different units.

In principle, all systems are distinctionally "downward" compatible, i.e. distinctions can be dropped in favour of informationally "poorer" Portmanteau terms. Both such "Portmanteau-fusing" and the relocation of a given distinction from one level of analysis to another, however, is much facilitated in an automatic system, if *all* information is (a) of the same notational type and (b) string coded rather than structure coded. Tag notation combines both advantages - information from all levels can be lumped together as "words" in a string. Constituent analysis, in contrast, involves either graphical computation tools or at least bracket matching algorithms, both of which are more complex than the ordinary search-and-replace tools needed for manipulating tag strings.

As mentioned, structural incompatibility (1e-f) is harder to handle than mere category distinctions, for a variety of reasons. One of the reasons is, that a change from one structural notation to another may force a theory/notation-inherent distinction, that is left under-specified in the other system. In these cases, *no* equivalent notation exists, all transformation is mandatorily "upward", like hierarchical PP-attachment when moving from traditional CG grammar to PSG tree-structures, or head-dependent distinction (1f) when moving the other way.

In tag notation, function cannot be expressed other than *explicitly*, and dependency relations are a minimal way of doing this. Higher level function tags, like subject, object and the like, make things even worse, since they are often only *implicitly* marked in constituent analyses, as when defining an (English) subject as a clause-minus-VP-constituent. Moving from function-tag to constituent-tree is, by comparison, relatively "easy" (cp. 4.6.3) - though still considerably more demanding than Portmanteau-fusing or level-movement (1a-c). This asymmetry in transformability is yet another argument for making (word based) tagging the primary notation.

The 1:1 transformation of pronoun subclasses in table (2) can serve as an example for a theory dependent tag filter. In my system, there are three pronoun classes, all morphologically defined: Personal pronouns (PERS), determiner pronouns (DET) and specifier pronouns (SPEC). Functionally, all can substitute for whole NPs in the role of subject, object etc., but only DET pronouns can appear as adnominal modifiers (@>N). The three classes correspond in traditional Portuguese grammars to 6 "pseudosemantic" pronoun classes[213] and the "functional" article class:

(2) **Table: pronoun subclass filtering**

Traditional pronoun class	**CG tags**
personal pronoun	PERS
possessive pronoun	DET <poss>
demonstrative pronoun	DET/SPEC <dem>
interrogative pronoun	DET/SPEC <interr>
relative pronoun	DET/SPEC/ADV <rel> @#FS/AS-N<
indefinite pronoun	DET <quant1/2/3>

[213] Sometimes the category of *reflexive pronoun* is added, which would have to be filtered as a syntactic subclass of personal pronouns: <refl> PERS.

	SPEC <quant0> DET/SPEC <rel> ¬ @#FS/AS-N<
article	DET <art>/<arti>

As can be seen, both morphological word class tags, secondary lexical tags and syntactic tags are made use of in the filter, underlining the rather hybrid and idiosyncratic character of traditional pronoun classes.

Traditionally, the Portuguese word forms *a, as, o, os* would be described either as definite articles or - when modified by PPs or relative clauses - as demonstrative pronouns. In my system, they belong to the same word class (DET), but have two lexically bound secondary tags, <art>[214] and <det> that are then functionally disambiguated by valency-level CG-rules, which can - from lower level disambiguation - draw on both syntactic function (3a) and word class context (3b):

(3a) REMOVE (@%art) (NOT 0 @>N) ; #remove definite article reading if not prenominally used
(3b) REMOVE (@%dem) (0 <artd>) (NOT 1 DET-REL OR SPEC-REL OR PRP) ; # remove demonstrative reading of potential definite articles if the following word is not a relative pronoun or a preposition

The article tag for the indefinite articles *um, uma, uns, umas* can be directly filtered from my DET class, since the necessary ambiguity resolution does not concern syntactic function, and the secondary <arti> (indefinite article) need not be disambiguated once the word class is determined. This, however, is still a considerable disambiguation task, as can be seen from the rules in (4), that perform differentiation from the NUM word class (numerals) necessary for the two singular forms, *um* and *uma*..

(4)

<u>ordinary rules:</u>

SELECT ("um" <quant2> DET) (1C SO-ADJ) ; # **um** só
SELECT ("um" <quant2> DET) (-1C <vK>) (NOT 1 VEZ OR PRP-DE/PRP_DET) ; # foi **um** ano extraordinário, not: foi **um** dos ...
SELECT ("um" <quant2> DET) (1 PRP-DE) (NOT 2 N/V-P) ; # O vice dá **uma** de galante titular
SELECT ("um" <quant2> DET) (-1 PRP-DE) (*-2 KOMP-CORR BARRIER NON-NOMINAL/PRP) (1 N) (2 <rel>) ; # a marca mais agradável d**um** ano cuja retrospectiva ...

SELECT ("um" NUM) (-1 MAIS) (0 S) ; # mais **uma** vez
SELECT ("um" NUM) (1C <dur>) (0 S) (NOT 2 <rel>) ; # durava **uma** semana; *not:* a marca mais agradável d**um** ano cuja retrospectiva ...
SELECT ("um" NUM S) (*2 OUTRO BARRIER CLB LINK NOT 1 N) ; # **um** em=frente ao outro
SELECT ("um" NUM S) (1 PRP-DE/PRP_DET LINK 0 P) ; # **um** desses cidadões
SELECT ("um" NUM S) (1C VFIN) ; # **um** vive, outro morre

[214] For these words, <art> implies <artd> (definite article), which is the icon used as a set definition and in output filtering.

SELECT (NUM) (-1C PRP-DE) (-2 MAIS) ; # mais de **um**
SELECT (NUM P) (1C <num+>) ; # **cinco** por=cento
SELECT (NUM) (0 NUM-POST-E) (-1 KC) (-2 NUM) ; # trinta e **cinco**, quatro ou **cinco**

<u>heuristic level 1:</u>

SELECT ("um" <quant2> DET S) (1 S/<hyfen>) (NOT 1 V-NONAD) (NOT -1 <+num>) (NOT 1 <num+>) ; # o que é que come **uma** baleia? ; *choose the quantifier (indefinite article) reading before a singular noun phrase constituent that cannot at the same time be a verb, if the immediate context isn't characteristic of numeral context (<num+> or <+num>)*
REMOVE ("um" <quant2> DET S) (NOT 1 S/<hyfen>) ; # vinte e **um** crianças; *remove the quantifier (indefinite article) reading if the immediate right hand context isn't singular (and nominal)*
REMOVE ("um" <quant2> DET S) (1C V-NONAD) ; # **um** basta

SELECT (NUM) (1 <num+>) ; # o prazo de **uma** semana
SELECT (NUM) (-1 <+num>) ; # número **um**

Sometimes, notational transformation can be done by simple lumping of classes: If a grammarian using my parser wants to drop the morphological distinction between the noun (N) and proper noun (PROP) word classes for, say, pedagogical reasons, PROP can be filtered into <prop> N, which would move the proper noun distinction from the word class level to a semantic (secondary) level.

4.6.3 Tree structures for constituent analysis

Given the popularity of constituent grammars in present day linguistics, it is an interesting question whether the flat structure of traditional CG can in some way be transformed into a constituent tree structure, and, if so, - will information be lost, or ambiguity added ?

Both, I suggest. First, in replacing functional tags by mere PSG constituent order, information will be lost, especially on the sentence and clause levels, where the tag system is richest, less in the ad-N and ad-A structures, where argument and modifier function normally is expressed only in the head's valency tags, not in the dependent element's syntactic tag. This problem can be remedied by enriching the mere constituent tree by tagging its nodes with the original CG function tags.

On the other hand, certain attachment underspecifications will be brought to the foreground when building an explicit tree structure from a CG-notation, as in the following examples:

i) @N< after a postnominal which features a nominal head itself:

... o gigante Venceslau comedor de gente ***famoso*** *...*

(‘the giant Venceslau, eater of men well-known’)

ii) co-ordination problems, like in the single/double attachment of the @N< (postnominal) in:

... cinco homens e quatro mulheres ***do Rio*** *...*

(‘five men and four women from Rio’)

iii) @<PRED after interfering nominal arguments (like @<ACC):

... viu o amigo, ***exausto*** *... - não viu nenhuma solução, exausta*

(‘he saw his friend, exhausted’ - ‘exhausted, he saw no solution’)

Possible solution strategies might involve agreement [in (i), for Portuguese, at least], minimal attachment [or minimal co-ordination], and semantic matching between head and modifier [in (iii)]. Of course, it is hard to see, how *any* primarily syntactic description should be able to totally resolve these ambiguities, - so elegant underspecification might even qualify as the best solution yet ...

The following is an outline of an algorithm for constructing constituent trees from flat dependency structures as used in my parser:

- 1. Attach all adnominal adjects (@>N, @N<) and adverbial adjects (@>A, @A<) to their heads, choosing as head the closest word of eligible word class in the direction indicated by the attachment markers (>, <). The syntagms that are established in step (1) are later moved and co-ordinated as wholes in steps (4) and (2).

- 2. Co-ordinators are regarded as intra-phrasal, where a step-(1)-element has to cross them in order to find an eligible head. Otherwise co-ordinators are applied to the highest level pair of heads that otherwise would contradict the uniqueness principle.

- 3. Clause boundaries are introduced where two uncoordinated arguments clash due to the uniqueness principle, and between same-level attachment markers that point away from each other.

- 4. Clause level arguments (including clausal @#-arguments) are attached to the closest main verb (@MV) in the indicated direction, unless there is an interfering clause boundary, in which case the first @MV after the second clause boundary is chosen, and so on.

The gradual introduction of layered parentheses (or tree branching sections) might look like this:

a) unprocessed flat description:
O pai de o menino que venceu comprou dez cervejas.
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>-@#FS-N<-@SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC
(The father of the boy who won bought ten beers)

b) attacchment of prenominals:
(O pai) de (o menino) que venceu comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>-@#FS-N<-@SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC

c) attachment of postnominal relative and finished PP:
(O pai) (de ((o menino) que venceu)) comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>-@#FS-N<-@SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC

d) finished first NP:
((O pai) (de ((o menino) que venceu))) comprou (dez cervejas).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>-@#FS-N<-@SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC

e) clause boundaries (marked by '-') due to the uniqueness principle between unco-ordinated 2x @SUBJ> and 2x @FMV, plus subject/object argument attachment:
(((O pai) (de ((o menino) - (que venceu)-))) comprou (dez cervejas)).
DET-@>N N-@SUBJ> PRP-@N< DET-@>N N-@P< <rel>-@#FS-N<-@SUBJ> V-@FMV V-@FMV NUM-@>N N-@<ACC

In order to show the implementability of this concept I have written a computer program (called *brackets*) that identifies group and clause boundaries in a flat CG-style description, thus delimiting constituents, then marks constituent boundary brackets as complex *form* (np, pp, icl etc.), and finally assigns to every complex constituent a function tag derived from the syntactic CG-tag of its head. Below, the transformation is shown for the following sentence: *A crise apura o paladar do consumidor e valoriza o* *dono de restaurante que pilota a própria cozinha* (‘The crisis sharpens the palate of the consumer and values the restaurant owner who pilots his own kitchen’):

(8a) Analysed text, in flat, word based CG-notation:

word form	*base form*	*valency & semantics*	*word class & inflexion*	*syntax*
*a	[a]	<art>	DET F S	@>N
crise	[crise]	<sit>	N F S	@SUBJ>
apura	[apurar]	<vt> <sN>	V PR 3S IND VFIN	@FMV
o	[o]	<art>	DET M S	@>N
paladar	[paladar]	<anost> <fh>	N M S	@<ACC
de	[de]	<sam->	PRP	@N<
o	[o]	<-sam> <art>	DET M S	@>N
consumidor	[consumir]	<DERS -or>	N M S	@P<
e	[e]		KC	@CO
valoriza	[valorizar]	<vt> <sN>	V PR 3S IND VFIN	@FMV
o	[o]	<art>	DET M S	@>N
dono	[dono]	<H>	N M S	@<ACC
de	[de]		PRP	@N<
restaurante	[restaurante]	<inst>	N M S	@P<
que	[que]	<rel>	SPEC M/F S/P	@SUBJ> @#FS-N<
pilota	[pilotar]	<vt> <vH>	V PR 3S IND VFIN	@FMV
a	[a]	<art>	DET F S	@>N
própria	[próprio]	<jn>	ADJ F S	@>N
cozinha	[cozinha]	<ejo>	N F S	@<ACC

(8b) after tree structure transformation, with added group/clause tags and hierarchical tabs:

```
@SUBJ>:np
 |-@>N:DET F S                *a            [a] <art>
 |-@H:N F S                   crise         [crise] <sit>
@FMV:V PR 3S IND VFIN         apura         [apurar] <vt> <sN>
@<ACC:np
 |-@>N:DET M S                o             [o] <art>
 |-@H:N M S                   paladar       [paladar] <anost> <fh>
 |-@N<:pp
    |-@H:PRP                  de            [de] <sam->
    |-@P<:np
       |-@>N:DET M S          o             [o] <-sam> <art>
       |-@H:N M S             consumidor    [consumir] <DERS -or>
@CO:KC                        e             [e]
@FMV:V PR 3S IND VFIN         valoriza      [valorizar] <vt> <sN>
@<ACC:np
 |-@>N:DET M S                o             [o] <art>
 |-@H:N M S                   dono          [dono] <H>
```

```
|-@N<:pp
|  |-@H:PRP                  de            [de]
|  |-@P<:N M S               restaurante   [restaurante] <inst>
|-@N<:fcl
   |-@SUBJ>:SPEC M/F S/P     que           [que] <rel>
   |-@FMV:V PR 3S IND VFIN   pilota        [pilotar] <vt> <vH>
   |-@<ACC:np
      |-@>N:DET F S          a             [a] <art>
      |-@>N:ADJ F S          própria       [próprio] <jn>
      |-@H:N F S             cozinha       [cozinha] <ejo>
```

[**word classes:** DET=determiner, N=noun, V=verb, PRP=preposition, KC=co-ordinating conjunction, SPEC=specifier-pronoun, ADJ=adjektiv; **bøjning:** S=singular, P=plurar, M=male, F=female, PR=present, 3S=third person singular; **derivation:** <DERS -or>=suffiksderivation på '-or'; **syntaks:** @>N=prenominal, @SUBJ>=subject, @FMV=finite main verb, @<ACC=accusative object, @N<=postnominal, @P<=argument of preposition, @CO=co-ordinator, @#FS-N<=finite subclause functioning as postnominal; **valens:** <art>=article, <rel>=relative, <vt>=monotransitive verb; **semantics:** <H>=human, <sit>=situation, <ejo>=functional place, <inst>=institution, <anost>=anatomical bone structure; **selektionsregler:** <fh>=human feature, <sN>=has non-human subject, <vH>=has always human subject, <jn> has non-human head; **ortografi:** <sam->&<-sam>=first and second part of fused expression]

(8c) Same clause, automatically transformed into horizontal tree structure notation (for more graphical trees, with the notational conventions used in the VISL teaching system, cp. chapter 7.2.5):

```
0 |
0 @STA:cu
0 |______________________________________________________________________________
1 |                                                               |       |
1 @CJT:fcl                                                        @CO:KC  @CJT:fcl
1 |__________________________                                     |       |________
2 |               |       |                                       |       |       |
2 @SUBJ>:np               @FMV:V  @<ACC:np                                        |   @FMV:V
          @<ACC:np
2 |_________      |       |__________________                     |       |       |_______
3 |       |       |       |       |       |                       |       |       |
3 @>N:DET @H:N    |       @>N:DET @H:N    @N<:pp                  |       |       @>N:DET
3 |       |       |       |       |       |________               |       |       |
4 |       |       |       |       |       |       |               |       |       |
4 |       |       |       |       |       @H:PRP  @P<:np          |       |       |
4 |       |       |       |       |       |       |________       |       |       |
5 |       |       |       |       |       |       |       |       |       |       |
5 |       |       |       |       |       |       @>N:DET @H:N    |       |       |
5 |       |       |       |       |       |       |       |       |       |       |
  A       crise   apura   o       paladar de      o  consumidor   e     valoriza  o


0
0
0
1
1
1
2
2
2_______________________________
3 |       |                       |
3 @H:N    @N<:pp                  @N<:fcl
3 |       |_______        |_______________
4 |       |       |               |       |       |
4 |       @H:PRP  @P<:N      @SUBJ>:SPEC  @FMV:V  @<ACC:np
4 |       |       |               |       |       |______________
5 |       |       |               |       |       |       |       |
5 |       |       |               |       |       @>N:DET @>N:ADJ @H:N
5 |       |       |               |       |       |       |       |
  dono    de      restaurante     que     pilota  a   prórpria    cozinha
```

[@H =head, np =noun phrase, pp =prepositional phrase, fcl =finite clause, ':' =separator for function and form]

4.7 Elegant underspecification

An important difference between the flat CG-notation and the constituent tree notation is that the latter *must* make explicit (i.e. "overspecify") certain ambiguities left underspecified by flat syntax, e.g. in connection with postnominal attachment (especially prepositional phrases, 1a-c), co-ordination (2a) and adnominal adjects.

(1a) He saw **((**the man @<ACC with @N< the bicycle @P<**)** from @N< China @P<**)**.
(1b) He saw **(**the man @<ACC with @N< **(**the bicycle @P< from @N< China @P<**))**.
(1c) Viu o homem @<ACC com @N< a bicicleta @P< de @N< a China @P<
(1d) Foi buscar o homem @<ACC com @N< @<ADVL a bicicleta @P< de @N< a China @P<

(2a) **((**married @>N women @NPHR**)** and @CO men @NPHR**)**
(2b) **(**married @>N **(**women @NPHR and @CO men @NPHR**))**

But specifying as *grammatical* such ambiguity as in real discourse is resolved by *communicational* and *cognitive* context, is not necessarily a (notational) advantage. On the contrary, it is (structural) underspecification that in many (especially applicational) contexts makes the "flat" dependency of CG-syntax a purer, if not truer, model of *grammatical* structure than constituent based tree models. Take, for instance, machine translation, where much of the work is done not by *understanding* a text, but simply through lexical and *grammatical* transformation, where the possibility of underspecification becomes an important asset: - First, many instances of ambiguity represent "true *syntactic* ambiguity", that can only be understood by the fully contextualised - human - listener/reader, and are therefore best passed on to the communicational level in "raw" form. - Second, a number of these structural ambiguities (especially co-ordination [2a,b] and "short" [1b] vs. "long" [1a] attachment of postnominal PPs) are relatively universal, i.e. language independent within a certain language family, so that they can be preserved in translation, which can then be based directly on the "flat" description (1c), which “contains” both (1a) and (1b). Even more economical is (1d), where the additional possibility of a clause level adverbial PP reading with instrumental ‘com’ is written into the same word-tag string (‘He fetched the man with a bicycle from China’). Even here, translation into English is unaffected by the structural choices and preserves the same ambiguity.

To make an ambiguity of this type explicit (for a language pair that otherwise treats it alike), will only burden the translation module with irrelevant ballast. Adjectival modifiers, on the other hand, either postnominal or as free predicative adjuncts, are more problematic, since there may - in Portuguese, but not in English - be agreement relations (2c) between head and modifier, which could be exploited by a tree transformation module:

(2c) homens @NPHR e @CO **mulheres** @NPHR casad**as** @N<
‘men and women she-married’

Even where ambiguity is not underspecified by structural "flatness", i.e. where it concerns the *function* of individual - well established - constituents, Constraint Grammar notation, being word based, offers an elegant means of expressing multiple readings: Several (competing) function or dependency tags can be added to the same word, so that ambiguity can be expressed in *one* analysis rather than in two or more different trees. Especially with long sentences this may be more transparent - and pedagogically superior - than a multi-page list of complete alternative (tree) analyses.

5

The uses of valency: A bridge from syntax to semantics

5.1 Valency as lexical "boot strapping": A research paradox

The role of valency in my parser is rather dynamic: On the lower, morphosyntactic, levels of analysis, valency is treated as “God given” lexical information inherently bound to a given lexeme, which is expressed by secondary “untouchable” tags. On higher levels of analysis, valency tags are themselves disambiguated (ch. 5.4 and 5.5), turning into primary tags and even allowing semantic interpretation. Such dynamics is natural and wanted in the concept of progressive level parsing, and fairly easy to implement in the CG formalism. From a research point of view it is more problematic that valency is treated as a “God given” feature on a *lexicographic* level in the first place. Valency patterns, especially *verbal* valency patterns, are not stable in the same way inflexional patterns are. Consider the following cline of monotransitivity, from least to most transitive with respect to direct objects:

verb	**direct object**		
cair <ve> (<va>)	?	(no chão)	‘fall (on the ground)’
dormir <vi>	um sono feliz		‘sleep [a happy sleep]’
viver <vi><vt>	uma guerra		‘live/experience [a war]’
comer <vt><vi>	peixe		‘eat [fish]’
lançar <vt>	uma ofensiva		‘launch [an offensive]’

Cair is an ergative (inaccusative) verb (<ve>), with a patient subject which prevents it from taking a direct (patient) object, ever (?). The verb does take *adverbial* arguments, though (‘cair no chão’). *Dormir* is usually known as an intransitive (inergative[215]) verb (<vi>), but a direct object can be forced by replicating the verb’s meaning in the object noun (‘sleep’ V - ‘sleep’ N). *Viver* (‘live’) is usually used as an intransitive, though mostly with a place or time complement. However, with a change in meaning (‘experience’), direct objects of the <occ> (occasions) or <per> (periods) semantic prototypes are allowed. *Comer* is usually used as a transitive (<vt>), but not

[215] If used in the *perfeito simples* tense (‘dormiu’), a more ergative reading and perfective aspect can be forced (‘he fell asleep’), and the adjectival use of the participle ‘dormido’ (unthinkable in a “pure” inergative intransitive verb) *is* possible, making ‘dormido’ mean ‘adormecido’ (‘asleep’, “fallen asleep”).

obligatorily. In the intransitive, there is a meaning change stressing the *process* of eating, but this seems to be analogous in English and Danish, and cannot be detected in translation. *Lançar*, finally, is next to obligatorily transitive.

In general, due to ellipsis and anaphoric usage, it is more common for a verb to move "down" the ladder of transitivity than "up". In Portuguese, yes/no-questions are answered by repeating the "naked" finite verb in the first person, without any complements:

Come peixe? - Como.
Posso telefonar? - Pode.

Thus, even auxiliaries and otherwise obligatorily transitive verbs can appear in the intransitive, at least from a parsing point of view where the window of analysis is the sentence.

On these grounds, short of tagging all verbs for both <vi> and <vt>, it is difficult to find a safe lexicographic strategy for marking valency potential. In the parser's current lexicon, the strategy has been to list the *maximal* valency potential, i.e. preferring <vt> or <vi><vt> over <vi>, since it is more dangerous for the parser's syntactic performance to have a rule discard an @ACC reading due to a missing <vt> tag than allowing an @ACC reading for longer than necessary, due to a superfluous <vt> tag. The reason for this is that my CG rule set is very "cautious" - with much more REMOVE rules than SELECT rules - and that, once discarded, tags cannot be recovered, while wrongly undiscarded readings can always be discarded by another rule later. Also, lexically unprovided-for intransitivity comes as a natural by-product in the robust CG system in those cases where there isn't even a candidate constituent for the role of @ACC, while unprovided-for transitivity is always a problem, precisely because there *is* a constituent (the @ACC candidate) waiting to be tagged, and risking to be tagged wrongly.

Therefore, only very rare valencies are omitted. With regard to <vi> and <vt>, the *preferred* valency is listed first, a fact which can be exploited in the CG by declaring order sensitive sets:

LIST <vt-vi> = (<vt> <vi>) ; *preferably transitive, but potentially intransitive*
LIST <vi-vt> = (<vi> <vt>) ; *preferably intransitive, but potentially transitive*

Typically, these sets will be used in order to make rules more cautious, as in the following context example, where monotransitivity (<vt>) is demanded as the preferred (but not necessarily only) valency:

(... <vt> LINK NOT 0 <vi-vt>)

Returning to a research view concerned with lexicography, the softness of valency tags, while relevant for the parsing grammar, is not the only problem. Epistemological methodology is another one. With lexicographic research into valency being one of its potential applications, how can a syntactic parser defend using valency information in its own (tool) grammar? If direct object tags (@ACC) were assigned *if and only if* the closest main verb candidate is tagged <vt>, then it would indeed not make sense to use the parser's output for extracting information on verbs' monotransitive valency.

However, CG rules are not absolute rules like the rules in standard generative systems (for instance DCG). In a DCG, if a verb terminal is listed only as <vi>, and there are no rewriting rules for a clause with a <vi> main verb and an additional np (the direct object), then a sentence with the combination <vi> and direct object will not be parsed. It just isn't part of the language defined ... implying that all lexicographic research has to be done a priori. Similarly, a standard DCG parse will fail with a sentence featuring a <vt>-only main verb but no direct object candidate np.

Technically, a CG parses the function of words, not sentences, and it does so by *removing* information, opting for the most "resilient" tag. Therefore, the only verb candidate in a sentence will always become be assigned main verb function, - even if it is listed in the lexicon as only <vt>, and there is no direct object. More laboriously, but likewise, the CG rule set may well arrive at providing the (correct) direct object tag (@ACC) even in the absence of a <vt> tag on the main verb, simply because all *other* readings have been discarded by rules stronger than the one that would have removed the @ACC reading. Also, in a CG parser, safe - unambiguous - contexts can force readings that contradict the original valency. An example are pronouns in the accusative ('o/a/os/as' - as enclitics or directly to the left of a verb), which will force a direct object (@ACC) reading already at the mapping stage. Likewise, que-clauses will be tagged as direct objects (@#FS-<ACC) directly after a transitive verb, even if it is tagged only for general (np) monotransitivity (<vt>) and not for the special <vq> (que-clause) transitivity.

For all these reasons, if a parsed corpus is used for the lexicographic extraction of valency information, it will offer *more* information than was originally fed into it in the shape of a valency entries in the parser's lexicon. One could say that the parser is "seeded" with valency information, and then only catalyses the decoding of structural information inherent in the corpus itself.

If the lexicographic net gain in the shape of additional valency pattern information is entered into the parsing lexicon, a new round of information gain is possible on the next corpus: the valency module of the parser "bootstraps" itself.

Of course, valency pattern information that is in conflict with the original lexical information will be subject to a higher error rate. This is of little importance, however, if *qualitative* information is what is wanted, for instances where a substantial number of automatically extracted but manually inspected corpus examples is used to add or modify a lexicon entry. *Quantitative* evaluation will be affected "locally", i.e. for the valency pattern of individual lexemes (e.g. the frequency of 'dormir' with a direct object), not for categorical conclusions (e.g. the frequency of @ACC or <vt> as such). If valency statistics for individual lexemes is mandatory, two solutions offer themselves:

(a) A qualitative valency analysis of the corpus is done first, i.e. the automatically extracted lexical valency patterns are automatically compared to those found in the lexicon, and all differences inspected manually and added to the lexicon, where they are deemed to contain "new" valency information rather than plain erroneous tagging. After this, a new, quantitative, run is performed.

(b) False *positive* valency readings (e.g. superfluous @ACC due to a wrong <vt> mark on the main verb) - especially *unambiguous* false positive readings - are extremely rare due to the "negative" way a cautious CG works - it relies mostly on REMOVE rules and the correct syntactic tag will survive (alongside or instead of, for instance, a false @ACC), since there would be nothing in the syntactic context calling for its removal. False *negative* valency readings (e.g. no @ACC because of a lacking <vt> mark) are more common, but will often lead to the survival of a "dummy" tag (@NPHR for nominal, @ADVL for adverbial material), which is used as function labels for the top node of utterances without a verbal top node. Now, sentences containing extra dummy labels can be filtered automatically, inspected and used to correct the individual lexeme valency statistics.

5.2 Valency instantiation: From secondary to primary tags

On the morphological and syntactic levels, valency information is used by the parser in the shape of angle-bracketed so-called secondary tags. The syntactic uses are obvious - a marker for monotransitivity (<vt>), for example, can be used for deciding whether a nominal group to the right qualifies for direct object (@<ACC), or whether a N-V ambiguous word form with the word form 'o' to the left, is more likely to be tagged as an article-noun pair, or as a sequence of personal (accusative) pronoun and verb. Valency information on preposition binding - for both verbs, nouns and adjectives - is essential for the functional pp-disambiguation between @PIV, @ADV, @N< and @A< on the one hand, and @ADVL, on the other. After lexicon look-up, in the morphological and syntactic modules, the valency potential of a word appears as an unordered sequence of (one or more) valency tags attached to word forms. At this stage, valency tags do not express any kind of grammatical analysis, but merely a lexical potential.

Valency tags can, however, become themselves bearers of syntactic information:

- 1. For a given word, valency tags can be used for making explicit syntactically motivated **word class subcategories** not part of the primary word class inventory. For example, a syntactic subclass of "title nouns" can be defined by means of valency (<+n>). As a primary tag, <+n> will be removed if the word in question does not bind a proper noun. Thus, one can distinguish between *'senhor'* ('master', 'lord') and *'senhor'* ('mister'), where the latter is a title noun, and the former is not. Also, if one wishes to regard the category of auxiliary as a word class rather than as a function, then the fact whether or not a class ambiguous verb governs an infinitive (or other auxiliary complement) is crucial for making the distinction, and valency tags like <x> (governing infinitive) or <de^xp> (governing 'de'-mediated infinitive) can be disambiguated in order to provide the necessary tags[216].

- 2. With regard to a word's dependent, valency tags allow more fine-grained **syntactic subdistinctions**. A simple example is the distinction between modifiers and arguments in np- and ap-groups, where the parser's *syntactic* tags, @N< and @A< only express dependency and underspecify valency. Compare:

  A confiança <+em> (no governo) @N< (argument postnominal)
  Um espião (no governo) @N< (modifier postnominal)

[216] In my system, 'auxiliary' *is* a function (@AUX), and the valency tag <x>expresses both the binding of infinitive auxiliary complements (in the case of @AUX) and the binding of direct object infinitive clauses (in the case of @MV).

Here, the <+em> valency tag allows the identification of a ‘em’-headed pp as postnominal argument rather than modifier.

On the clause level, the @ADV tag (adverbial argument), for pp’s expressing place or direction, underspecifies whether the *predicative* subfunction of the adverbial relates to the subject or to the direct object:

Morava ***<va+LOC>*** *(em Londres) @<ADV*
Colocou ***<vta+LOC>*** *a nova sede da empresa (em Londres) @<ADV*
Vai ***<va+DIR>*** *(para Londres) @<ADV*
Manda-o ***<vta+DIR>*** *(para Londres) @<ADV*

Here, the valency tag <va> indicates that the @ADV relates to the subject, while <vta> means that the @ADV relates to a direct object. A simple filtering program could capitalise on the valency information and change (enrich) the *syntactic* information on the pp, creating, for instance, the tags @ADV-SC and @ADV-OC, respectively.

- 3. Valency markers can be regarded as **dependency hooks** for making explicit ambiguous or disjunct dependency relations.

(a) *O pintor <+n> (de retratos) @N< (****Manoel Braga****) @N<*
(b) *A criação, (neste ano) @N<, (****duma fundação nacional de arte****) @N<*
(c) *A confiança <+em> (do povo) @N< (****no governo****) @N<*
(d) *a terra era mais <komp> rica em óleo (****do que imaginava****) @KOMP<*

In the examples above, the dependents in bold face are ambiguous with regard to their potential heads. Only in (b) is there a graphical indication (commas) of where to attach the pp in question. Simply using the principle of close attachment, would result in errors, attaching ‘Manoel Braga’ to ‘retratos’ in (a) and ‘no governo’ to ‘povo’ in (c), treating pp’s hierarchically rather than as sister postnominals. By using valency tags as hooks, however, correct attachment can be achieved. ‘Pintor’ in (a) is a title noun (<+n>) and binds the proper noun pp ‘Manoel Braga’, and ‘confiança’ is marked <+em> allowing the argument-attachment of pp’s headed by ‘em’, as is the case for ‘no governo’ in (c). In (d), finally, a valency tag, <komp> (comparative) helps to link head (‘mais’) and dependent (‘do que imaginava’) in a disjunct constituent functioning as adverbial adject (@>A) of ‘rica’. The comparandum dependent (‘do que imaginava’) could itself have been tagged as adverbial adject (@A<), but that would leave attachment ambiguous (left to either to ‘rica’ or ‘mais’). Therefore, a more explicit function tag (@KOMP<) is chosen which functions as the “far” end of the valency link between comparandum head (‘mais’) and body (‘do que imaginava’).

Note that valency is treated as a feature of the *head* of a dependency relation, which explains why a different type of marking is used for the two ends of a valency link - (functional) syntactic tag on the dependent and valency tag on the head.

5.3 Disambiguating valency tags

Naturally, in order to supplement word class, syntactic or dependency tags, valency tags will have to be treated as primary tags, and - if necessary - disambiguated accordingly. On a yet higher level of analysis, such disambiguation can also be useful with regard to polysemy resolution. As we will see in chapter 6, valency instantiation is one of the semantic tools that lends itself to the CG approach, the reason for this being that a word's meaning (or, hence, translation) often depends on the presence and type of valency bound arguments.

Incidentally, the fact that valency instantiation - within a CG framework - allows fairly seamless and effortless progression from syntax to semantics, seems (in a technical way) to support Halliday's view on semantics as "ever more delicate syntax".

From a technical, CG point of view, valency tagging differs from morphosyntactic tagging in principled ways. Morphology and word class is first derived from the lexicon and then disambiguated with the use of other types of primary tags. Syntactic function, with few exceptions, cannot be derived from the lexicon, but is mapped and then disambiguated *with* the use of previously disambiguated *morphological* primary tags. Valency tags, finally, are primarily lexicon-based, like morphological and word class tags, and "come into existence" by mere - root-based - lexical look-up, expressing a kind of (lexico-syntactic) potential. On the other hand, disambiguationally, valency tags are even more dependent on previously disambiguated *other types* of primary tags (here: syntactic) than is the case in syntactic tagging. Simply, *that* valency tag is chosen (from the valency tag list of a given word) which matches an already established argument category. In most cases, then, the *real* tagging work has been done on the syntactic level already, and no new information is gained. Thus, the fact that a direct object function has been established, implies transitive valency in the head verb, i.e. monotransitive valency (<vt>) with no other objects present, ditransitive valency (<vdt>, <vtp>) *with* a dative object or prepositional object (@PIV) present, and transobjective valency (<vtK>) in connection with an object complement (@OC). Likewise, in the case of pp-attachment to nouns or adjectives with matching valency ('confiança no governo'), one could say that the grammar - at an earlier level - already must have "seen" the relevant valency tag at the noun/adjective when deciding for attachment and against an @ADVL function tag. Therefore, CG-rules at the valency disambiguation level are fairly simple, and only in the few cases of ambiguous valency attachment is there any "real" work left. This is why I prefer to call the process *valency instantiation* rather than *valency disambiguation.*

The following rules are taken from the rule section that instantiates monotransitive valency (<vt>):

```
REMOVE (@%vt) (0 @%vK) (*1C @<SC BARRIER CLB/VFIN) ;
```

Monotransitivity is abandoned in favour of a copula reading (<vK>) if there is an unambiguous subject complement to the right in the same clause.

```
REMOVE (@%vt) (NOT 0 @MV OR PCP) ;
```

Monotransitivity is out of the question if the verb isn’t functioning as main verb, unless the verb is inflected as a past participle. As a participle, monotransitive verbs loose their direct object, but they still have to be instantiated as monotransitive if the valency tag is to be used for polysemy resolution.

```
REMOVE (@%vt) (0 @%vr) (1 @%refl) ;
REMOVE (@%vt) (0 @%vr) (*-1 @%refl BARRIER CLB) ;
```

Monotransitivity is abandoned in favour of a reflexive reading if their is a reflexive 3. person pronoun (<refl>) directly to the right, or to the left without an interfering clause boundary.

```
REMOVE (@%vt) (0 @%vq) (*1 QUE LINK 0 @#FS-<ACC LINK NOT *-1 CLB) ;
```

The general monotransitivity reading <vt> is abandoned in favour of the more precise “cognitive verb” valency for finite que-clauses (<vq>), if there is a conjunctional ‘que’ to the right without interfering clause boundary and heading a direct object clause (@#FS-<ACC).

```
REMOVE (@%vt) (0 @%xt) (*1 @#ICL-<ACC BARRIER @<ACC) ;
```

The general monotransitivity reading <vt> is abandoned in favour of the more precise matrix verb valency for ACI and causative constructions (<xt>), if there is a non-finite subclause to the right functioning as direct object and without another - leftward pointing - direct object (@<ACC) in between.

```
REMOVE (@%vt) (NOT *1 @<ACC&) (*-1 SB/VFIN BARRIER @ACC>) (NOT 0 @#ICL-AUX< OR @ATTR) ;
REMOVE (@%vt) (NOT *1 @<ACC&) (*-1 SB/VFIN BARRIER @ACC> LINK NOT -1 @ACC>) (NOT 0 @ATTR) ;
```

These last two rules are very general, stating that a <vt> reading is to be ruled out if there is no left pointing direct object anywhere to the right, and no right pointing direct object to the left before a finite verb or sentence boundary is encountered. The NOT 0 @ATTR exception takes care of participles (that are to retain their <vt> valency in spite of not having a (surface) direct object.

```
REMOVE (@%vt) (0 @%vp-all) (*-1 @PIV> BARRIER @MV OR @#FS) ;
REMOVE (@%vt) (0 @%vp-all OR @%xtp) (*1 @<PIV BARRIER @MV) (NOT *-1 @MV) ;
REMOVE (@%vt) (0 @%vp-all OR @%xtp) (*1 @<PIV BARRIER @MV) (*-1 @MV BARRIER @#FS) ;
```

This group of rules abandons the <vt> reading, if the possibility of af <vp>, <vtp> or <xtp> reading is made likely by the presence of a prepositional object (@PIV) to the right or left - pointing the right way and without another interfering main verb.

6

The semantic perspective: Incremental Semantic Parsing (ISP)

6.1 Semantic tagging

Language analysis in a parsing perspective (both manual and automatic) is traditionally subdivided into different levels, where different applications take an interest in different levels. Thus a morphological or word class (PoS) analysis may be satisfactory for corpus based lexical frequency analyses, while internet based grammar teaching as in the VISL project (Bick 1998 and http://visl.hum.sdu.dk) needs syntactic, and machine translation needs semantic analysis.

One might assume that each of these levels needs its own parsing tools, - for instance, probabilistic tools for PoS-tagging, generative grammars for syntactic parsing and logical meta-languages for the semantic level. My research, however, indicates that it is possible to extend at least one rule based tool - Constraint Grammar - to ever “higher” levels of analysis - provided a lexical data base containing the necessary lexicographic information is developed in parallel. One could say that in the case of Constraint Grammar, the quality and granularity of the analysis is not inherent in the technique, but rather goal driven, and that it can be improved in an incremental way.

Are semantic tags practical?

After morphological and syntactic tagging the next logical step in Progressive Level Parsing appears to be the semantic tagging of word items. But is semantic tagging possible, and is tagging a sensible approach at all to the semantic analysis of free text? Clearly, in terms of referent resolution or information content, any type of semantic analysis of NL texts is a task not to be fully resolved in the near future, - whatever the notational conventions used. Since human language is intertwined with human intelligence and human knowledge, full semantic analysis will not work without a certain degree of artificial intelligence and a huge bank of "knowledge about the world", both unavailable at the present time.

Nevertheless, in a more limited sense, semantic tags can be useful for present day tasks:

For one, semantic tags can be used, as secondary tags, for the disambiguation of *syntactic* ambiguity. <+HUM>, for instance, in a noun connected to the verb "build", would rule out the object reading, and favour a subject reading, in a language where word order can't be counted on to make the distinction. To a lesser degree, secondary semantic tags can help resolve morphological word stem ambiguities, too - e.g. where two words share all graphical features (in some or all of their inflexional forms), and only differ in pronunciation, or not at all. The Portuguese verbs ‘ser’ and ‘ir’, for example, have a large overlap in inflexional forms. Thus, ignoring auxiliary functions, ‘foi’ can mean both ‘I/he was’ and ‘I/he went’. However, given the semantic subject restrictions of a MOVE-verb like ‘ir’, ‘foi’ can be disambiguated as a copula (‘ser’) if the subject is not <+MOVE>.

Finally, a list of alternative (mutually exclusive) semantic tags attached to *one* word can be exploited for sense distinction in polysemic lexical items, as desirable in, for instance, machine translation or information extraction. To this end, the semantic tags themselves need to be subjected to disambiguation. This can be achieved by unifying semantic tags from syntactic arguments with semantic slots from their valency head. Thus, speech verbs would have a semantic subject slot, that says +HUM, action and activities would have a slot saying +ANIM. Thus, a singing star (+HUM) is not the same kind of star as a falling star (-HUM).

Even before the introduction of semantic tags proper, semantic information has been used in the CG-rules of my parser - by declaring semantically motivated *sets of base forms* (e.g. V-SPEECH, V-MOVE), and sets of *syntactic* tags, from which a certain semantic feature can be inferred (<vq> for cognitive verbs, or N-HUM for {<+n>, <attr>}. In principle, semantic tags *could* be mapped on the basis of such set membership, blurring the line between (CG) lexicon and (CG) grammar. Instead, for reasons of consistency, I have treated semantic tags in the same way I have treated valency and word class tags, by introducing additional information at dictionary level, integrated into the lexicon entries concerned.

6.2 Semantic prototype tags for nouns: Minimal distinction criteria

What should semantic tags denote?

It is hard to see how semantic tags could be determined formally (like morphological tags), or structurally (like syntactic tags), without themselves becoming morphological or syntactic tags. But then, of course, refining syntactic distinctions *is* one way of approaching semantics. For example, I have used valency grouping for sense distinction in verbs. I think, however, that tagging and the choice of tags, to a certain degree, has to be goal oriented. So what are semantic tags good for? My own ultimate goal is machine translation, and so sense distinction in polysemic words is of primary importance.

Therefore, there is a case for introducing "real" (i.e. not primarily syntactic) semantic classes on top of “mappable” (i.e. syntax or morphology based), semantic distinctions.

But *which* and *how many* semantic tag categories should be chosen? And how should they be organised with regard to each other? When answering this question, it is important to make the teleological distinction between *defining meaning* and *distinguishing meaning*. In a system that does not claim to *understand* text, but only to *structure* or *translate* it, the final semantics lies (only) in the eye of the beholder.

Therefore, in order to make a sense distinction, it isn't necessary to *define* a given sense, or *identify* its referent, - it is enough to draw a line across the semantic map which separates two different senses of a word. If a parser can determine - in a given text - just which side of the line a given usage has to be placed, then a disambiguated tag will allow the system to pick the exact word sense from the *lexicon* – the sense need *not* be in the *tag itself*. Constraint Grammar is an ideal tool to make such choices in a context dependent and robust way. Like on the morphological and syntactic level, CG rules will not *define* but *disambiguate* from a given set of distinctions. Semantic tags, then, have to reflect the structure of the semantic landscape rather than the meaning of individual words. Consequently, for a CG to work well, the system’s set of semantic tags should be inspired by some ordering or classification principle for word senses.

Word meanings – to mention some basic systems - can be ordered in autonomous groups (word fields, picture dictionaries), semantic hierarchies (monolingual thesauri, biological classification, Princetown WordNet) or word nets (multilingual relational systems like EuroWordNet). Finally, word meanings can be ordered by semantic decomposition:

vivo
-> *morto = ‘NOT alive’*
viver = ‘BE alive’
-> *nascer = BECOME alive’*
-> *morrer = ‘BECOME NOT alive’*
-> *matar = ‘MAKE BECOME NOT alive*

I believe that the most cost efficient[217] way to draw distinction lines across the semantic landscape is by *prototype similarity*. Prototypes can be conceived as both class hyperonyms (‘animal’: *dog*, *pig*, *lion*, ‘plant’: *oak*, *sunflower*) and common well known set members of classes otherwise too “abstract” (‘knife’ for ‘cutting tools’: *knife*, *sword*, *saber*, ’book’ for ‘readables’: *book*, *paper*, *magazine*). Whatever the abstraction level, prototype similarity testing is done by asking *‘is it more like A or more like B?’* rather

[217] On the one hand, in terms of CG rule efficiency, on the other, in terms of the work load needed to enter the relevant information into the lexion.

than by asking *'is it an A?'* or *'is it a B?'*. Thus, senses are distinguished by which semantic prototype is closest, not by set membership proper - and since the distinguishing line for one prototype is halfway to the next, neighbouring prototype, this leaves a large margin of potential *definitional* uncertainty (labour saved for the parser). Disambiguational prototype similarity is much easier to resolve than referential or compositional identity or even set membership (as in hierarchical systems): Let's assume that sense A and sense B of a given word could be *defined* (within the system) by 10 features each, of which they share three. Instead of grammatically treating all of the 17 defining features, it is *disambiguationally* enough that 1 of the 14 *distinguishing* features tests positive. For example, if the Danish word *marsvin* (meaning both 'dolphin' and 'guinea pig' could be prototyped as *fish* in one sense, and as *rabbit* in the other, and if the grammar knows that fish swim in water, and rabbits don't, then all 'swimming' predicators and all 'water' place adverbials in the same clause make the dolphin-reading very likely, though dolphins by no means can be defined as biologically belonging to the set of 'fish' – they are just more close to the 'fish' prototype than to that of rabbit, in terms of their lexical collocational potential.

Disambiguationally, it is important to distinguish between "lexical" polysemy where the different senses are what could be called "thesauric heteronyms" belonging to different areas of a thesauric system, and "thesauric" polysemy where the different senses are "thesauric hyponyms" of the same hyperonym. In the second case, disambiguation is *only* needed where the distinctions are forced by the need for translation into different terms, in the first case, disambiguation will be useful in monolingual analysis, too, since it may help disambiguate the form and function of the word in question, or of other words in the sentence.

In lexical polysemy, one doesn't need many prototypes in order to distinguish between senses that are accidental homographs (converging polysemy) or stem from metaphorical or metonymic transfer (diverging polysemy). In the first case, different etymology will usually ensure a fair degree of "semantic distance", like in *'fato'* ('suit', 'flock') that has absorbed the meanings of *'facto'* ('fact') in Brazilian Portuguese because of a phonetic and graphical disappearance of 'c' before 't' in many Brazilian words. In the second - metaphorical – case, "semantic distance" with regard to one or more features is what characterises a metaphor in the first place. Thus, metaphoric transfer often moves from abstract to concrete or vice versa, or from animal to human, as when Danish *'sild'* in its Portuguese translation has to be disambiguated into its 'fish' and 'girl' meanings. Metonymic transfer is often used "live" as a purely rhetorical tool, where no disambiguation is necessary before translation, since the effect can be assumed to be the same in the target language ('an angry letter' – 'uma carta furiosa', 'o Itamarati hesitou' – 'the Itamarati [palace of government] hesitated'), but some synecdochic relations (pars pro toto, totum pro parte) do need to be treated lexically, as for Danish *'træ'* that becomes *'árvore'* ('tree') in Portuguese as a plant, but *'madeira'* ('wood') as a material. In all the above cases, whether disambiguation is wanted in order to choose the right translation equivalent, or in order to resolve a metaphor, polysemy crosses over thesauric branches for every new word sense, allowing the prototype similarity technique to work:

	prototypes	**distinguishing features**
marsvin	fish or rabbit?	±SWIM, ±WATER-PLACE
fato	clothing, group or abstract?	±CONCRETE, ±ANIM
sild	fish or human?	±HUM
carta furiosa	book or human?	±HUM, ±READABLE
Itamarati	house or human group?	±HUM, ±PLACE
træ	plant or material?	±COUNT, ±MASS

Senses that are simply *thesauric* hyponyms of a standard general meaning of the term, however, are more problematic - at least where a bilingual view point is taken, and where the target language doesn't share the hyperonym-hyponym-structure of the source language, - in particular, where the target language *lacks* the hyperonym concerned. An example is the Portuguese word 'dedo' which can mean both 'toe' and 'finger' in English. The meanings of the English translations are, of course, thesauric hyponyms of 'dedo', and are therefore difficult to distinguish by prototype similarity.

The choices of prototypes on the one hand, and minimal distinction criteria on the other, appear to be interdependent, and I want to argue that the best list of prototypes is not the one that gives the best descriptions, but the shortest one that can *handle* the sense distinctions given an operationally feasible list of distinction criteria. Likewise, for my purpose, the best list of minimal distinction criteria is not the one that allows perfect compositional semantic analysis of all word senses, but the shortest one that can resolve prototype similarity for all semantic prototypes. After all, it is easier to single out an Italian by the native language he speaks, than by defining a prototypical dark-haired gesticulating wine-consuming pasta-eater, since nowadays, you would find a lot of those among other nationalities, too.

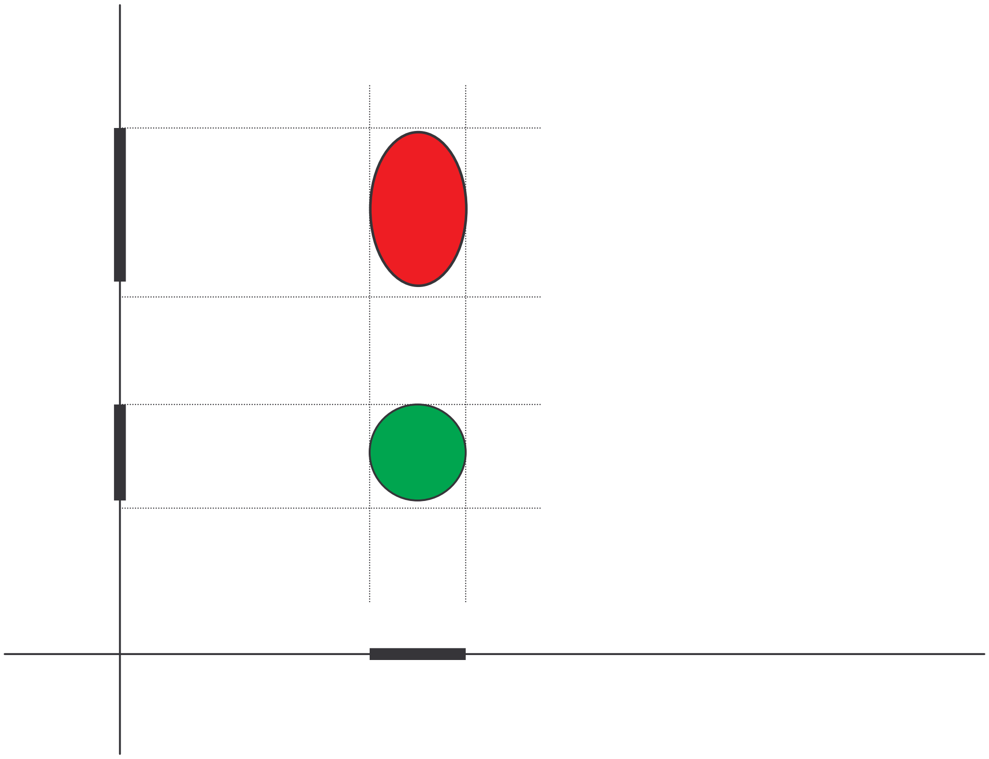
6.3 A semantic landscape

One can imagine semantic prototypes as multidimensional bubbles in a multidimensional semantic landscape, where the dimensions (or co-ordinates) are expressed as *semantic features,* and the balloons are *feature bundles.* Binary semantic features (±HUM, ±MOVE) are about "hemisphere" membership, east-west, north-south etc. I will call such features ***atomic*** *semantic features.* For other features, like size, temperature, colour, prototypes will allow a certain (fuzzy) range along a dimension. Sometimes, these ranges will be subdivisions of what could be expressed as one hemisphere membership[218], - ‘SWIM’, ‘WALK’, ‘FLY’, for instance, cover different meaning ranges of ±MOVE. Different animal prototypes, ‘fish’, ‘mammal’, ‘bird’, ‘duck’, all belong in the same quadrant of a two-dimensional universe with the co-ordinates of ±LIVE and ±MOVE, but their “meaning bubbles” still have different positions and shapes. For example, there is a partial overlap between ‘bird’ and ‘duck’ since both fly and walk, but a prototypical bird doesn’t swim.

The multidimensional prototype bubbles as well as any word sense bubble can be projected onto semantic landscapes with fewer dimensions, even one-dimensional and two-dimensional, like the shadow of a balloon hovering over a sun-lit plane, or in front of a vertical cliff. The concept of minimal distinction criteria is about comparing balloons of different shape and position in as flat a universe as possible (i.e. with the fewest possible dimensions). The trick is simply about which projection to choose: a round and a vertically cigar-shaped balloon will yield the same circular shadow on the plane, but different shadows on the cliff side. Likewise, quadrant-size balloons flying on top of each other (like the ones for ‘SWIM’, ‘WALK’ and ‘FLY’ on the ‘MOVE’ axis) will blend their square shadows on the plane (e.g., expressing ‘+LIVE’ and ‘+CONCRETE’), but be distinguishable by form and/or position against the cliff side (‘MOVE’).

[218] *Any* feature could, of course, be expressed as its own dimension, if one so chooses, i.e ±swim etc., but this does not appear to be a “cost-effective” method, if these features do not have independent combination patterns with *other* dimensions.

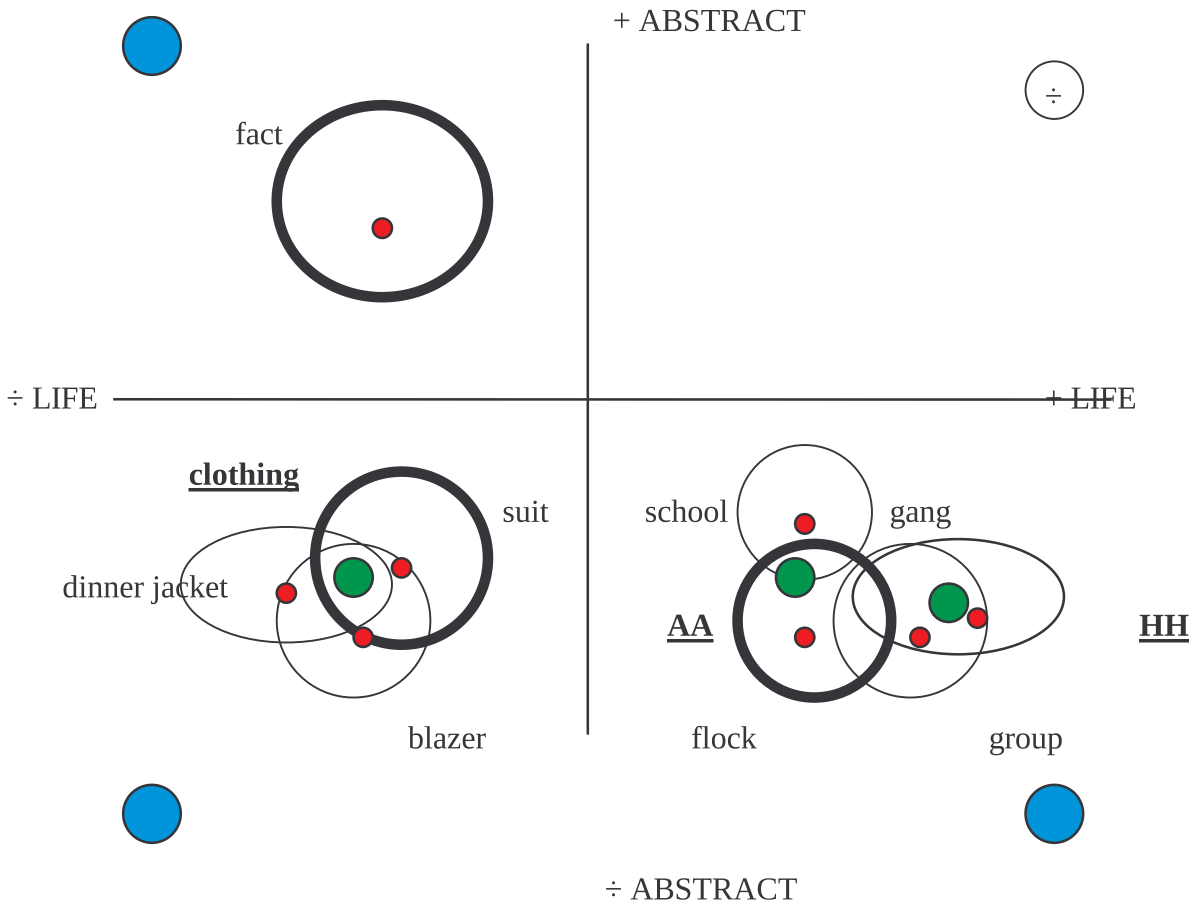
Illustration: Disambiguation of semantic prototype bubbles by dimensional downscaling (lower-dimension projections)



Let’s take a specific example. As suggested above, it may be near impossible in a principled way to *define* the Brazilian Portuguese word *fato* in however vast a data base, but in a bilingual (i.e. practically oriented MT-) perspective it would be quite possible to distinguish between the three *Danish* translations 'kendsgerning' (fact), 'habit' (suit), and 'flok' (flock), by means of only two *atomic semantic features*, ±abstract and ±HUM, in different combinations, like *abstract not living* ('fact'), *not abstract not living* ('suit') and *not abstract living* ('flock'). These features are furthermore enough to delimit and discriminate (not define!) larger prototype families in relation to each other, like *"clothing"* (<tøj>) and *"group of animals"* or *"group of people"* (in the illustration AA and HH, respectively). In a Constraint Grammar context a hierarchy of lexicon and context driven grammatical rules can "discard" or "select" these features in a given sentence either individually or group-wise in the form of prototypical "feature families" (like 'clothing') [219]. Thus, if the parser knows from the lexicon that ‘fato’ includes in its prototype range <AA> (‘animal multiplicity’) and <tøj> (‘clothing’), rules will

[219] At present there are altogether ca. 200 different tags for semantic prototypes in my parser. The semantic features of nouns can be reduced to 16 hierarchically ordered "atomic" features. Verbs are tagged for ±HUM subject selection, and adjectives for ±HUM nominal selection.

disambiguate the senses ‘suit’ and ‘flock’ as a by-product of a more general prototype disambiguation from which other semantically ambiguous words will profit, too.

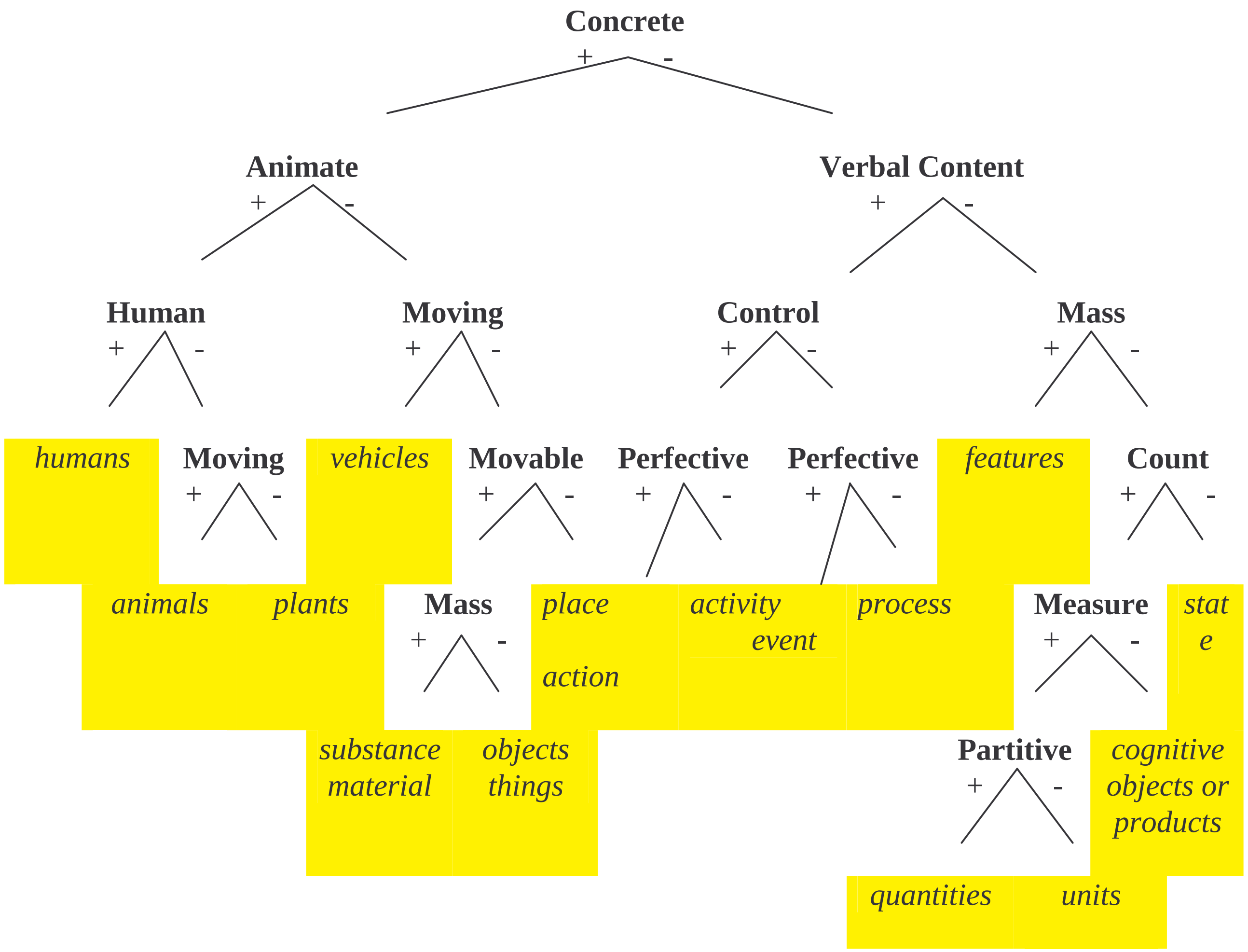


The diagram places two lists of related words in a semantic field, relating them to each other and to prototypical notions (medium size green circles) or feature combinations (big blue circles). The nuclear meanings of the word are marked by small red dots, and their semantic reach by ordinary circles of varying size. The graphical representation illustrates how 'suit', 'blazer' and 'dinner jacket' are difficult to distinguish, since they all belong to the same prototype, 'clothing'. However, a single atomic feature, ±LIFE is enough to distance all three words from others like 'flock' or 'gang'. In order to make a distinction between words within one LIFE/ABSTRACT quadrant, more features are necessary, for instance, ±ANIMAL for the distinction between the AA-word 'school' (of fish) and the HH-word 'gang'. 'Flock' and 'gang' have a semantic overlap - *the priest lectured his flock* - , which is best described in a metaphorical way: 'lecture' projects its +HUM-object selection restriction features at the valency bound 'flock'. The feature combination +ABSTRACT/+LIFE is empty, since ±LIFE is a binary subdivision of ÷ABSTRACT.

A. Atomic semantic features

In all, the parser uses 16 atomic semantic features for nouns. These features are strictly binary in the sense that every feature has only two states, plus and minus (±). Also, most of the 16 features are organised in a hierarchic binary tree with every higher level ± feature choice leading to two new (lower level) ± feature choice. In the feature tree there are 15 binary branching nodes, representing 12 atomic features (±MOVING, ±MASS and ±PERFECTIVE occur in two branching nodes each) and yielding 16 terminal categories as feature combination paths. For example, the terminal category of ‘place’ can be defined (in terms of atomic features) as +CONCRETE, -ANIMATE, -MOVING and –MOVABLE.

Illustration: atomic semantic features - binary decision tree



Apart from the semantic features shown in the decision tree, the parser uses 4 additional features:

± HUMAN EXPRESSION[220] (i.e. qualifiable by adjectives etc. usually associated with humans), e.g. ***política** agresiva* ('aggressive politics')
± ADJECTIVAL (FEATURE), e.g. *tamanho* ('size')
± LOCATION (as an independent feature, because -MOVE -MOVABLE alone doesn't cover abstracta), e.g. *concerto* ('concert')
± TEMPORAL

Note that the feature tree is organised as a decision tree. Therefore, it is possible to infer only those features from terminals, that lie in the decision path leading to that terminal. Animals, for example, are moving, non-human, animate and concrete, and the parser can disambiguate the concept 'animal' as true for a given sense by instantiating all these features, or as wrong, by disallowing at least one of the features. However, due to the way in which features are lumped into prototypes by physics or language/cognition, other atomic features may be inferred or used for disambiguation as well. Thus, all moving things are also movable, i.e. something that isn't movable, can't be an animal. Some prototypes, on the other hand, violate the branching rules of the feature tree because of metaphorical usage. The prototype <inst> (institutions), for instance, comprising words like *igreja* ('church') or *justiça* ('justice'), routinely promotes buildings or abstract nouns to +HUM status by allowing them to act, think, decide and order.

Therefore, the feature tree must be understood as a theoretical point of departure, while the real parsing system is built upon feature set intersection and disjunction for individual prototype bundles, thus relying heavily on the inclusion or exclusion of certain key features with a maximum of distinctive power. Exactly which positive or negative features can be inferred from which other positive or negative features, is shown in the table below, where all 16 atomic semantic features are plotted against the parser's prototype classes for nouns, which are bundled according to the atomic feature bundles they match (one table row per prototype/feature bundle). The prototypes of institution (<inst>), town (<by>) and country (<land>), for instance, form a bundle, and share the same set of atomic features, positive for CONCRETE/ENTITY (they can be touched), HUM (they can wish, decide and act), LOCATION (something can be in them), and negative for all other features.

A feature X can be inferred in a given bundle, if there is a feature Y in the same bundle such that – with respect to the whole table - the set of prototype bundles with feature X is a subset of the set of prototype bundles with feature Y. In the table below, inferability is marked in the following way:

[220] The difference between ±HUM and ±HUMAN EXPRESSION is to a certain degree isomorphic to the distinction between argument selection and modifier selection: ±HUM words can fill the subject slot in, for instance, speech verbs, while ±HUMAN EXPRESSION (only) can fill the head slot for human modifiers: *triste* ('depressed'), *otimístico* ('optimistic').

‘+	uninferable positive feature
+	inferable positive feature
.	uninferrable negative feature
(empty)	inferable negative feature

Within the particular prototype/feature bundle (<inst> [institution], row 3) and (<by> [town], <land> [country], row 4), +CONCRETE/ENTITY and +HUMAN EXPRESSION can be inferred from +HUM, while +HUM itself and the other remaining positive feature (+LOCATION) are uninferrable in the sense, that there is no feature in the bundle representing a superset of bundles for the +HUM or +LOCATION bundle sets. Among the bundle’s negative features, -ANIMATE (living) and -MOVABLE, for example, cannot be inferred, while –MOVE can be inferred once –MOVABLE is given. What distinguishes the prototype bundles in row 3 and 4, is the ±COUNTABLE feature (N), since human settlement words like *cidade* (‘town’), *aldeia* (‘village’) and *bantustão* (‘banana republic’) are countable[221], while institutions are not: *a polícia* (‘the police’ - which in Portuguese can’t mean ‘the officers’). Actually, institution words used in the plural (i.e. as +COUNTABLE) often switch to the prototype category of <hus> (‘building’, row 9), loosing the +HUM and +X features in the proces.

[221] In the prototype scheme, uman settlement nouns behave somewhat like nouns denoting individual humans (with the additional feature of fixed location). Thus, human settlement nouns can have names, and as such, become members of another morphological PoS class, making singular form a lexeme category.

Atomic semantic features for differentiating semantic word classes

E = entities (±CONCRETE) V = ±VERBAL
C = ±CONTROL P = ±PERFECTIVE
I = ±MOVING S = ±MEASURING
J = ±MOVABLE D = ±PARTITIVE
A = ±ANIMATE (living) X = ±HUMAN-EXPRESSION (allowing human modifier-ADJ)
H = ±HUMAN ENTITY F = feature (±ADJECTIVAL)
M = ±MASS L = ±LOCATION
N = number (±COUNTABLE) T = ±TEMPORAL

E	C	I	J	A	H	M	N	V	P	S	D	X	F	L	T	
+		'+	+	'+	'+	.	'+			.		+	.	.	.	H, HM, N, prof, fam, title, attr
+		'+	+	'+	'+	.	.			.		+	.	.	.	HH, parti
+			.	.	'+	.	.			.		+	.	'+	.	inst
+			.	.	'+	.	'+			.		+	.	'+	.	by, land
+		'+	+	'+	.	.	'+			.		.	.	.	.	A, AM, AB, zo, D, orn, ent, ich
+		'+	+	'+	.	.	.			.		.	.	.	.	AA,DD
+			.	'+	.	.	'+			.		.	.	.	.	bo,B
+			.	'+	.	.	.			.		.	.	'+	.	BB
'+			.	.	.	.	'+			.		.	.	'+	.	top, agua, sky, vej, area, ejo, hus, ta, bar, an, anorg, anost, anfeatc, sygc, anzo, anorn, anich, anbo, star
.			.	.	.	.	'+			.		.	.	'+	.	topabs, spids, hul
+		.	'+	.	.	.	'+			.		.	.	'+	.	tm, fælde, kovr, ujo, rør, bild, r
.			.	.	.	.	.			.		.	.	'+	.	stil, sit, anfeat
.			.	.	.	'+	.			.		.	.	'+	.	vejr, vind, regn, ling
+		.	'+	.	.	.	'+			.		.	.	'+	.	anmov
'+			.	.	.	.	.			.		.	.	'+	.	surf
+		'+	+	.	.	.	'+			.		.	.	.	.	V, skib, fly, or
+		'+	+	.	.	.	.			.		.	.	.	.	VV
+		.	'+	.	.	'+	.			.		.	.	.	.	cm, liqu, mat, stof, mad, kul, drik, rem
+		.	'+	.	.	.	'+			.		.	.	.	.	cc, part, er, sten, stok, ild, vejrc, madc, kulc, il, kniv, fio, klud, sejl, paf, lys, ten, mu, tøj, sko, hat, smyk, tøjzo
+		.	'+	.	.	.	'+			+	'+	.	.	.	.	CC/ar
.						'+	.			.		.	.	.	.	am, amh
.						.	'+			.		.	.	.	.	ac, featc, p, l, w, s, f, tegn
.						.	'+			.		.	'+	.	.	ret, akc, meta
.						.	'+			.		'+	.	.	.	reg, pp, ll, ww, ss, sd, rr
.						.	'+			.		'+	'+	.	.	right
.						.	'+			.		.	.	'+	.	geom
.						.	.			.		.	'+	.	.	ax, state, sh, feat, fh, ak, syg, col, o, ling
.						.	.			'+	.	.	'+	.	.	featq, fq
.						.	.			.		'+	'+	.	.	ism, akss
.	'+					.	.	+	.	.		+	.	.	+	CI, lud, sp, fag, terapi, tæsk, dans
.	'+					.	'+	+	'+	.		+	.	.	+	CP, CPP, d, kneb (<vt>)
.	'+					.	.	+	'+	.		+	.	.	+	CPS (-ação)
.						.	.	'+	.	.		.	.	.	+	cI (<vi><ve>)
.						.	'+	+	'+	.		.	.	.	+	cP, cPP (<vi>), snak, strid
.						.	'+	+	'+	.		'+	.	'+	+	occ (= human place-event)
.						.	.	+	'+	.		.	.	.	+	cPS (<ve>)
.						.	.			'+	.	.	.	.	.	num+ & unit
.						.	.			+	'+	.	.	.	.	num+ & qu, qus
.						'+	.			'+	.	.	.	.	.	mon ?
.						.	.	.		.		.	.	.	'+	temp (= non-deverbal event)
.						.	'+	.		'+	.	.	.	.	'+	dur (= time-unit)
.						.	'+	.		.		.	.	'+	'+	per (= time-place)

'+ = underivable pos. feature, + = derivable pos. feature, . = underivable neg. feature

6.5 A semantic Constraint Grammar

In the semantic, bilingual part of the Portuguese CG lexicon, meanings are expressed as translation alternatives, and kept apart by listing - for each alternative - a set of *discriminators,* - CG tags that are to be present in the final analysis for the alternative in question to be chosen:

Portuguese word	**Danish translation**	**to be chosen if tagged as**
palavra	*ord1*	tag1 tag3
	ord2	tag1 tag5
	ord3	tag2
	ord4	tag4 tag5 tag6

In principle, all kinds of tags can be used as discriminators, e.g. semantic prototype class and inflexion for nouns, valency, and subject class for verbs, semantic argument class for prepositions etc. Comparing discriminator lists with actual tag strings, the system goes for maximal discriminator instantiation. If *palavra* in the example has received the tag string 'tag1 tag2 tag5', then *ord2* will be chosen as the correct meaning (translation).

As explained in chapter 6.3 and 6.4, the semantic module of my parser aims at resolving bilingually motivated noun polysemy by disambiguating semantic prototype membership. Rules can either target prototypes directly, or indirectly via atomic semantic features. Chapter 6.5.1 discusses the techniques used to disambiguate atomic semantic features, while chapter 6.5.2 is concerned with direct prototype disambiguation. Chapter 6.5.3 explains how polysemy resolution can be achieved by exploiting information that is not primarily semantic, such as morphosyntactic information disambiguated at a lower level, and by the "instantiation" of valency patterns.

6.5.1 Using feature inheritance reasoning

In the parser's lexicon, every noun entry features a '±' list for all 16 atomic semantic features used in the system. This list is computed by a special program from the prototype spectrum of the noun in question. Positive features are marked with capital letters, negative features with small letters. Since the feature structures of a noun's semantic prototypes are compiled on top of each other, many atomic features of polysemous nouns will appear as both positive (capital letter) and negative (small letter). Consider the following examples of polysemous institution nouns, with their Danish translation equivalents given according to semantic prototype inventory:

word	Ee = entities (±CONCRETE)	Cc = ±CONTROL	Ii = ±MOVING	Jj = ±MOVABLE	Aa = ±ANIMATE (living)	Hh = ±HUMAN ENTITY	Mm = ±MASS	Nn = number (±COUNTABLE)	Vv = ±VERBAL	Pp = ±PERFECTIVE	Ss = ±MEASURE	Dd = ±PARTITIVE	Xx = ±HUMAN-EXPRESSION	Ff = feature (±ADJECTIVAL)	Ll = ±LOCATION	T = ±TEMPORAL	polysemy spectrum
faculdade	Ee	c	i	j	a	Hh	m	N	v	p	s	d	Xx	f	Ll	t	
	E					H							X		L		<inst> fakultet
	e					h							x		l		<featc> evne
fundo	Ee	c	i	j	a	Hh	m	Nn	v	p	s	d	Xx	f	Ll	t	
	e					h		N					x		L		<topabs> bund
	E					H		n					X		L		<inst> fond
	e					h		N					x		L		<hul> nåleøje
	e					h		N					x		l		<ac><smP> midler
indústria	Ee	c	i	j	a	Hh	M	n	v	p	s	d	Xx	f	Ll	t	
	E					H	m						X		L		<inst> industri
	e					h	m						x		l		<am> flid, snilde
							M										
justiça	Ee	c	i	j	a	Hh	M	n	v	p	s	d	Xx	f	Ll	t	
	e					h	m						x		l		<am> ret(færdighed)
	E					H	M						X		L		<inst> justits
							m										
rede	E	c	i	Jj	a	Hh	m	Nn	v	p	s	d	Xx	f	L	t	
				J		h		N					x				<ujo> net
				j		H		n					X				<inst><+n> netværk
				J		h		N					x				<tm> hængekøje

Ambiguous feature potential - here involving E, J, H, M, N, X and L - is shaded. All words in the table are ambiguous with regard to the features H and X (±HUM), and since it is the <inst> prototype that contributes the +HUM (H/X) feature potential, and since all other readings have -HUM (h/x), the <inst> prototype can be singled out by (positively) disambiguating features H or X, i.e. by discarding features h or x.

Thus, the grammar need not address the <inst> prototype directly, but can achieve a considerable effect by merely disambiguating *one* atomic feature, ±HUM. For every prototype bundle a list of features can be built such that each member of the list is capable - by “negative instantiation”[222] - of discarding all prototypes in the bundle. Prototype bundles are lumped together by CG set definitions, for instance:

```
LIST <inst-proto> = @%inst ;
LIST <anim-proto> = @%A @%AM @%AB @%zo @%D @%orn @%ent @%ich
```

As can be seen, the <inst> prototype forms a bundle of its own <inst-proto>, while the animal prototype bundle (<anim-proto>) consists of several sister prototypes.

For the <inst-proto> prototype bundle, the list of “killing features”, i.e. of features that - if discarded - disallows all prototypes in the bundle, is called <inst-slet> in the set-definition section of the parsers semantic CG module:

```
LIST <inst-slet> = (@j @a @H @m @n @s @X @f @L @t) ;
```
[223]

Apart from the above mentioned H- and X- features (+HUM), the set includes -MOVABLE (j), -ANIMATE (a), -MASS (m), +LOCATION (L) and others. The actual prototype tag killing is performed by a simple CG rule:

```
REMOVE <inst-proto> (NOT 0 <inst-slet>)
```

Note that the basic type of this rule has no context conditions, only tag or set membership conditions for the target itself. In practice, of course, a rule like the above could be split up into individual rules for every target prototype and atomic feature condition, and made more cautious or specific by adding real context conditions. Still, the basic idea of the rule is *not* to establish a context, but rather to implement a kind of *feature inheritance reasoning*, inferring one feature/tag from another: If a word form is not A, then it cannot be B either. One could also say that the rule helps to express (atomic) semantic feature information in terms of (complex) prototype bundles.

Due to the way atomic features are lumped into prototype bundles, certain positive or negative atomic features imply certain others, and this, too, can be handled by rules expressing feature inheritance reasoning. Atomic features, like prototypes, have their own “killing sets”. In the prototype killing set only the essential, underivable, atomic features are listed for the prototype bundle concerned (here <inst-slet>), but information from other “secondary” features, is “translated” into disambiguation of primary features by feature inheritance rules:

```
REMOVE (@=FEATURE) (0 @=feature) (NOT 0 <FEATURE-slet>) ;
REMOVE (@=feature) (0 @=FEATURE) (NOT 0 <feature-slet>) ;
```

[222] i.e. by not being true in a given context with a given set of CG rules.

[223] With the present CG-compilers, @-tags (i.e. tags to be disambiguated on the same tag line) cannot be AND-ed as is the case here. Rules involving contexts like <inst-slet> must therefore be unfolded - at the latest, at compile time - into as many individual rules as there are AND-ed features in the set concerned.

In the real rules, FEATURE is a positive atomic feature (capital letter), ‘feature’ a negative atomic feature (small letter). In the case of <inst-proto>, the primary features j, H, s, L and t can draw on secondary features from outside the primary set.

In the table below, all applicable “killing sets” for atomic semantic features are listed:

```
LIST <E-slet> = (@=c @=v @=p @=s @=t) ;
LIST <C-slet> = (@=e @=i @=j @=a @=h @=m @=V @=s @=d @=X @=f @=l @=T) ;
LIST <I-slet> = (@=c @=E @=J @=m @=v @=p @=s @=d @=l @=t) ;
LIST <J-slet> = (@=c @=E @=v @=p @=s @=f @=t) ;
LIST <A-slet> = (@=c @=E @=v @=p @=s @=d @=f @=t) ;
LIST <H-slet> = (@=c @=E @=m @=s @=d @=X) ;
LIST <M-slet> = (@=c @=h @=n @=v @=p @=d @=x @=f @=t) ;
LIST <N-slet> = (@=m) ;
LIST <V-slet> = (@=e @=i @=j @=a @=h @=m @=d @=f @=T) ;
LIST <P-slet> = (@=e @=i @=j @=a @=h @=m @=V @=d @=f @=T) ;
LIST <S-slet> = (@=c @=e @=i @=j @=a @=h @=v @=p @=x @=l) ;
LIST <D-slet> = (@=c @=i @=a @=h @=m @=v @=p @=x @=f @=l @=t @=S) ;
LIST <X-slet> = (@=e @=i @=j @=a @=h @=m @=s @=d) ;
LIST <F-slet> = (@=c @=e @=i @=j @=a @=h @=v @=p @=d @=l @=t) ;
LIST <L-slet> = (@=c @=i @=s @=d @=f) ;
LIST <T-slet> = (@=e @=i @=j @=a @=h @=m @=d @=f) ;
LIST <e-slet> = (@=i @=j @=a @=h) ;
LIST <v-slet> = (@=c @=p) ;
LIST <t-slet> = (@=c @=v @=p) ;
LIST <j-slet> = (@=i) ;
LIST <s-slet> = (@=d) ;
```

Feature E (entity), for instance, isn't listed in the killing set for <inst-proto>, but rules discarding E will still work, since feature H (human entity) *is* listed, and not-H can be inferred from not-E by:

```
REMOVE (@=H) (0 @=h) (NOT 0 <E-slet>) ;
```

In the same way, it can be inferred that, if something can't be moved (not J), then it cannot move either (not I). These relations are not symmetric, of course: books, for instance, cannot move (not I), but they can *be* moved (J).

```
REMOVE (@=e) (*-1 @%col LINK 0 @>N LINK NOT *1 @NON->N) (NOT 0 @%topabs) ;
REMOVE (@=e) (*1 @%col LINK 0 @N< LINK NOT *-1 @NON-N<) (NOT 0 @%topabs) ;
REMOVE (@=e) (*1 PRP-DE BARRIER NON-POST-N LINK 0 @N< LINK 1 @P< LINK 0 (@=M)
   AND (@=E) LINK NOT 0 (<@=e>)) (NOT 0 @N-META);

REMOVE (@=H) (0 @SUBJ>) (*1 @MV BARRIER CLB-ORD LINK 0 V-NON-HUM LINK NOT 0
   V-HUM&) ;
REMOVE (@=H) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
   @MV LINK 0 V-NON-HUM LINK NOT 0 V-HUM&) ;
REMOVE (@=H) (*-1 @%jn LINK 0 @>N LINK NOT 0 @%jh LINK NOT *1 @NON->N) ;
REMOVE (@=H) (*1 @%jn LINK 0 @N< LINK NOT 0 @%jh LINK NOT *-1 @NON-N<) ;

REMOVE (@=h) (0 @SUBJ>) (*1 @MV LINK 0 V-HUM LINK NOT *-1 CLB-ORD) ;
REMOVE (@=h) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
   @MV LINK 0 V-HUM) ;
REMOVE (@=h) (*-1 @%jh BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jn) ;
REMOVE (@=h) (*1 @%jh BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jn) ;
REMOVE (@=i) (0 @SUBJ> AND @=I) (*1 @MV BARRIER CLB-ORD LINK 0 V-MOVE);
REMOVE (@=i) (0 @SUBJ>) (*1 @MV LINK 0 V-MOVE LINK NOT *-1 CLB-ORD) ;
REMOVE (@=i) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-MOVE) ;

REMOVE (@=j) (0 @<ACC) (*-1 @MV LINK 0 V-MOVE-TR) ;

REMOVE (@=A) (0 @SUBJ>) (*1 @MV BARRIER CLB-ORD LINK 0 V-NON-HUM LINK NOT 0
    V-DYR&) ;
REMOVE (@=A) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-NON-HUM LINK NOT 0 V-DYR&) ;

REMOVE (@=a) (0 @SUBJ>) (*1 @MV LINK 0 V-DYR LINK NOT *-1 CLB-ORD) ; # hvis @=a
    fjernes uden @=A, fås metaforisk dyre-læsning ...
REMOVE (@=a) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-DYR) ;

REMOVE (@=M) (0 P) (NOT 0 S) ;

REMOVE (@=m) (0 S LINK 0 @=M) (-1 DET-MASS LINK 0 @>N) ;
REMOVE (@=m) (0 S LINK 0 @=M) (*-1 NON-PRE-N OR MMM-QUANT BARRIER
    DETA/B/C/D/E LINK NOT 0 PRP) (NOT 0 @%HH) (NOT -1 @>A) ; # ikke: tem $gente
    morrendo no Brasil, ikke: um=pouco iconoclasta

REMOVE (@=n) (0 @=N LINK 0 P) (NOT 0 S) ;

REMOVE (@=X) (*-1 @%jn BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jh) ;
REMOVE (@=X) (*1 @%jn BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jh) ;

REMOVE (@=x) (0 @=X) (*-1 @%jh BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jn);
REMOVE (@=x) (0 @=X) (*1 @%jh BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jn) ;
```

6.5.2 Using semantic context

In most cases, the syntactic relation between head constituent and dependent constituent is restricted not only in terms of the head's valency potential and the dependent's word class or syntactic form, but also in semantic ways. Head and dependent are subject to a kind of semantic "agreement" relation - in the case of modifiers (a) -, or semantic "valency" - in the case of arguments like subjects (b) or direct objects (c):

(a)	carinhoso <jh> *'tender'*	curto <jt> *'short'*	líquido <jn> *'liquid'*
um pai <H> *'a father'*	ok.	*	*

um verão <per> *'a summer'*	*	ok.	*
ferro <mat> *'iron'*	*	*	ok.

(b)	carne <mad> *'meat'*	cerveja <drik> *'beer'*	paulistas <H> *'paulistas'*
comemos <+ACC-mad> *'eat'*	ok.	?	*
bebemos <+ACC-drik> *'drink'*	*	ok.	*
convidamos <+ACC-hum> *'invite'*	*	*	ok.

(c)	falou <sH> *'spoke'*	latia <sA> *'barked'*	aconteceu <sN> *'happened'*
a criança <H> *'the child'*	ok.	*	*
o cachorro <A> *'the dog'*	*	ok.	*
a festa <occ> *'the party'*	*	*	ok.

In (a), a human noun as np-head (<H>) matches - and is matched by - a human adjective-modifier (<jh>). Neither head nor dependent can be exchanged with another class from the table: Periods of time (<per>) and materials (<mat>) can't be modified by adjectives asking for human heads, and time adjectives (<jt>) or thing adjectives (<jn>) can't modify human heads.

In (b) and (c), the valency of the three verbs for direct objects (b) or subjects (c) is semantically specified, and must be matched by the right semantic class in the object or subject noun. Like in the head-modifier case (a), conditions work both ways: one can eat meat, but not beer or paulistas, and meat can be eaten, but not drunk or invited. Children can talk, but don't bark or happen, and talking is done by children, not dogs or parties.

In the parser's present lexicon, the semantic classification of verbs and adjectives is less fine-grained than that of nouns. The basic set of distinctions is 'human', 'animal, 'plant' and 'non-living':

semantic class X	adjectives modifying X	verbs taking class X subjects	verbs taking *only* class X subjects	verbs taking class X objects
human	<jh>	<sH>	<vH>	<+ACC-hum> <+PIV-hum>
animal	<ja>	<sA>	<vA>	-
plant	<jb>	<sB>	<vB>	-
non-living	<jn>	<sN>	<vN>	-

Only the first three columns are fully implemented in the lexicon, while semantic object restrictions are listed in the lexicon only for a few verbs. However, these object restriction classes and other semantic verbal classes can be defined as ad-hoc sets and used, or even mapped, in the Constraint Grammar, drawing on syntactic classes and individual base forms. Of course, semantic sets are not tags, and while any set can be used to disambiguate other material (active disambiguation), only lexicon derived or mapped tags can themselves be disambiguated (passive disambiguation).

Some of the grammar’s more semantically motivated set definitions (SETS section) are listed below:

```
LIST V-EXIST = "existir" "faltar" "haver" "ter" ;
LIST V-HUM = @%vH @%+interr @%de^vtpK @%como^vtpK @%como^vta ;
LIST V-HUM& = @%vH @%sH ;
LIST V-TING = @%vN ;
LIST V-TING = @%vN @%sN ;
LIST V-NON-HUM = @%vN @%vA @%vB ;
LIST V-NON-HUM& = @%vN @%sN @%vA @%sA @%vB @%sB ;
LIST V-NON-TING = @%vH @%vA @%vB ;
LIST V-NON-TING& = @%vH @%sH @%vA @%sA @%vB @%sB ;
LIST V-ALL = V ;
SET V-HUM-SAFE = V-ALL - V-NON-HUM& ;
SET V-TING-SAFE = V-ALL - V-NON-TING& ;
SET V-NON-HUM-SAFE = V-ALL - V-HUM& ;
SET V-NON-TING-SAFE = V-ALL - V-TING& ;
LIST V-DYR = @%vA ;
LIST V-DYR& = @%vA @%sA ;
LIST V-KOMERC = "cobrar" "comprar" "pagar" "vender" ;
LIST V-NONCONTROL = @%vN @%va+TID> "achar" “estar" "gostar" "haver" "parecer" "ter" ;
LIST V-SPEAK = "admitir" "afirmar" "alfinetar" "analisar" "assegurar" "atirar" "bradar" "comemorar"
    "confessar" “contar” "definir" "determinar" "dizer" "falar" "garantir" "gritar" "insistir" "lembrar"
    "marcar" "observar" "planejar" "proclamar" “propor” "prosseguir" "rebater" "responder" "retrucar"
    "saber" "sugerir" "testemunhar" <+interr>;
LIST V-STUD = "ler" "estudar" "ter" ;
LIST V-MOVE = @%va+DIR "andar" "cair" "chegar" "correr" "entrar" "escorregar" "girar" "ir"
    "viajar" "vir" "voltar" ;
LIST V-MOVE-TR = @%vta+DIR "botar" "carregar" "colocar" "jogar" "levar" "pôr" "virar" ;
SET V-MOVE/TR = V-MOVE OR V-MOVE-TR ;
LIST V-NONCONTROL = <va+TID> <vN> "achar" "estar" "gostar" "haver" "parecer" "ter" ;
LIST V-EKFIN = "acabar" "chegar" "começar" "nascer" "romper" ;
LIST VT-NONPASS = "haver" "ter" ;
```

In the *rule* body (CONSTRAINTS section) of the semantic CG module, heads can be subjected to polysemy disambiguation by semantic projections from their dependents (modifiers or arguments), and the polysemy of dependents can be resolved by using semantic projections from their heads.

An example of **argument-head polysemy disambiguation**[224] are the monotransitive verbs *tirar* and *pôr*. A common translation into Danish is 'trække' *(tirar, 'pull')* and 'sætte' *(pôr, 'put')*, but with direct objects of the <tøj> class (clothes), the translation is 'tage af' [take off] and 'tage på' [put on], respectively. Which translation is chosen, depends on the discriminator <+ACC-tøj> and whether or not it has been removed by a CG-rule like the following:

```
REMOVE (@%+ACC-tøj) (*1 @<ACC& LINK NOT 0 @N-TOJ OR PERS) (NOT -1 @ACC>) ;
```

In terms of (polysemy) disambiguation gain by selection restrictions, **head-argument disambiguation**[225] is probably more important than argument-head disambiguation, since the latter - as opposed to the former - can also draw on the valency tag instantiation technique (cp. chapter 6.5.3) which is a very efficient tool. In the present grammar, (semantic) head-argument disambiguation is most advanced for subjects. For instance, the H/h atomic feature is disambiguated in subjects by checking for a matching V-HUM main verb.

```
REMOVE (@=H) (0 @SUBJ>) (*1 @MV BARRIER CLB-ORD LINK 0 V-NON-HUM-SAFE OR V-
    NON-HUM) ;
REMOVE (@=H) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-NON-HUM-SAFE OR V-NON-HUM) ;
REMOVE (@=h) (0 @SUBJ>) (*1 @MV LINK 0 V-HUM-SAFE LINK NOT *-1 CLB-ORD) ;
REMOVE (@=h) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-HUM-SAFE) ;
```

Another case of head-argument reasoning are rules stating that MOVE-verbs *(correr, viajar)* need +MOVING subjects (feature I), and transitive MOVE-verbs *(carregar, levar)* need +MOVABLE direct objects (feature J):

```
REMOVE (@=i) (0 @SUBJ> AND @=I) (*1 @MV BARRIER CLB-ORD LINK 0 V-MOVE);
REMOVE (@=i) (0 @SUBJ>) (*1 @MV LINK 0 V-MOVE LINK NOT *-1 CLB-ORD) ;
REMOVE (@=i) (0 @SUBJ>) (*1 CLB-ORD/VFIN LINK 0 CLB-ORD LINK *1 @MV LINK *1
    @MV LINK 0 V-MOVE) ;
REMOVE (@=j) (0 @<ACC) (*-1 @MV LINK 0 V-MOVE-TR) ;
```

In order to unravel the difficult and largely idiosyncratic semantics of prepositions, both argument-head and head-argument disambiguation can be exploited. The problem with prepositions is that it is very difficult to establish generally accepted semantic subclasses for an individual preposition. Dictionaries consistently disagree on the number of different meanings, and large dictionaries list dozens of different "meanings"

[224] i.e. arguments used for resolving polysemy in their head.
[225] i.e. heads used for resolving polysemy in their dependents.

for every preposition. Though some solutions have been suggested for monolingual polysemy mapping of prepositions (like the Zeugma-test in Togeby [1993]), I have decided to treat preposition polysemy from a *bilingual lexical* point of view, defining "meaning" as the [Danish] translation of a preposition.

The preposition *sobre* , for instance, is unlikely to mean 'om' in Danish if its head is not a cognitive verb, and its literal "physical" meaning 'på' is unlikely if its argument is not an entity, and this could be expressed by CG-rules like the following:

(a) *REMOVE (@om) (0 PRP-SOBRE LINK 0 @<PIV) (*-1 @MV LINK NOT 0 V-COG) ;
(b) *REMOVE (@på) (0 PRP-SOBRE LINK 0 @<ADVL) (*1 @P< LINK NOT 0 @=E) ;

In general, however, valency bound prepositions of the @N< and @PIV types[226] take their meaning (translation) from their head in individualized lexical ways (e.g. "phrasal verbs"), and general, semantic class based, rules like (a) are rather an exception than the rule, as the following list shows:

discussão-entrevista-livro sobre	diskussion-interview-bog **om**	discussion-interview-book **about**
controle sobre	kontrol **med**	control **of**
domínio sobre	magt **over**	power **over**
investigação sobre	undersøgelse **af**	investigation **into**
perícia sobre	kendskab **til**	knowledge **of**
vigilância sobre	omsorg **for**	care **of**
falar-meditar-votar sobre	tale-meditere-stemme **om**	talk-think **about**, vote **on**
especular sobre	spekulere **over**	speculate **on/about**
influir sobre	influere **på**	influence **-**
jurar sobre	sværge **ved**	swear **on/by**
reinar sobre	regere **over**	govern **-**
prevalecer sobre	slå igennem **overfor**	prevail **over**

Instead of crafting *rules* for individual words, I have chosen to exploit the prepositional valency tags (<sobre^vp> for verbs and <+sobre> for nominals) in *two* semantic ways, basing the translation of *both* the head *and* the preposition on valency tag instantiation at the head, and listing both in the head's lexicon entry:

jurar#....#<vi><por^vp><sobre^vp><vt># ..
__ <vi> bande
__ <vq> love
__ <por^vp><sobre^vp> **sværge /ved**

[226] With valency-bound adverbials (adverbial arguments or objects, @ADV), the meaning spectrum is more literal and systematic, covering 'på' (on, in contact with), 'over' (*over*, not in contact with) and 'hen over' (*across*, implying movement), depending on referential "physical" data.

The program module implementing the translational look-up (called *trad*), takes the translation of a valency governed preposition from its head, removing and replacing the original literal "dummy" translation at the preposition itself.

For prepositions that are not valency bound, the valency tag instantiation technique is not practical since no valency tags are used in the first place. Also, there is a higher degree of lexical variation in both head and dependent, which would lead to a proliferation of tags - if they *were* added -, since most verbs and nouns would have to be "dependency-potential" tagged for *all* prepositions (<+a><+com><+de> ...). But since "free" prepositions retain more of their literal meaning spectrum, the physical and other semantic features of the @P< argument can be exploited (argument-head polysemy disambiguation). To this end, the preposition in question is tagged with the relevant markers for what could be called "semantic valency". In the case of *sobre,* the following cases can be distinguished:

semantic valency tag	translation	type of argument	example
<+attr>	udover	ADJ, N <attr>	*sobre rico, é inteligente* (‘apart from rich, he is intelligent’) *sobre poeta, é um bom narrador* (‘besides being a poet, he is a good story teller’)
<+temp>	henimod	N <temp>	*sobre a noite, sobre o anoitecer* (‘towards nightfall’)
<+bar>	over	N <bar>	*sobre a mureta, sobre o gradil* (‘over the wall, over the fence’)

These "semantic valency" tags can then be disambiguated and instantiated in the same fashion as ordinary valency tags.

In some cases, the spatial semantics of the (nominal or verbal) head could be used, too (head-dependent disambiguation), as in the following two translation mapping rules. Typically, heads will be involved in terms of semantic category rather than lexically ("one lexeme at a time").

MAP (‘over’) TARGET ("sobre") IF (0 @N<) (*-1 N BARRIER @NON-N< LINK 0 <sky> OR <star>) ; Translate ‘sobre’ as ‘over’, if it functions as postnominal dependent of a noun belonging to the cloud- (<sky>) or star (<star>) prototypes.

MAP (‘over’) TARGET ("sobre") IF (0 @<ADVL) (*-1 @MV BARRIER CLB LINK 0 V-FLY) ; Translate ‘sobre’ as ‘over’, if it functions as a free adverbial dependent of a "fly"-verb (SET-defined as, for instance, *voar, pairar, passar, pender*)

At group level, a common example of **head-modifier disambiguation** are adnominal adjectives that are disambiguated with regard to ±HUM (<jh> vs. <jn> and <ja>), depending on the semantic class of the head in question, @N-HUM/X (human), *not* @N-HUM/X (not human), or @N-DYR (animal):

```
REMOVE (@%jh) (0 @>N) (*1 @NON->N LINK NOT 0 @N-HUM/X) ; # head to the right
REMOVE (@%jh) (0 @N<) (*-1 @NON-N< LINK NOT 0 @N-HUM/X) ; # head to the left
REMOVE (@%jn) (0 @>N) (*1 @NON->N LINK 0 @N-HUM/X) ; # head to the right
REMOVE (@%jn) (0 @N<) (*-1 @NON-N< LINK 0 @N-HUM/X) ; # head to the left
REMOVE (@%ja) (0 @>N) (*1 @NON->N LINK NOT 0 @N-DYR) ; # head to the right
REMOVE (@%ja) (0 @N<) (*-1 @NON-N< LINK NOT 0 @N-DYR) ; # head to the left
```

Similarly, adjectives functioning syntactically as different kinds of predicatives (@SC, @OC and @PRED) can be semantically disambiguated by checking the semantic class of the relevant nominal head[227], @SUBJ for @SC, @ACC for @OC, and in the case of free predicatives, @SUBJ for rightward pointing @PRED>, and the closest nominal head (set @NOM-HEAD) for leftward pointing @<PRED:

```
REMOVE (@%jh) (0 @<SC) (*-1 @MV BARRIER @SUBJ> LINK *-1 @SUBJ> BARRIER @MV
   LINK NOT 0 @N-HUM/X&) ;
REMOVE (@%jh) (0 @<OC) (*-1 @<ACC BARRIER CLB-ORD LINK NOT 0 @N-HUM/X&) (*1
   NON-PRE-N LINK NOT 0 @<ACC) ;
REMOVE (@%jh) (0 @<OC) (*1 @<ACC BARRIER @MV LINK NOT 0 @N-HUM/X&) ;
REMOVE (@%jh) (0 @<PRED) (*-1 @NOM-HEAD LINK NOT 0 @N-HUM/X&) ;
REMOVE (@%jh) (0 @PRED>) (*1 @SUBJ> LINK NOT 0 @N-HUM&) ;

REMOVE (@%jn) (0 @<SC) (*-1 @MV BARRIER @SUBJ> LINK *-1 @SUBJ> BARRIER @MV
   LINK 0 @N-HUM/X) ;
REMOVE (@%jn) (0 @<OC) (*-1 @<ACC LINK 0 @N-HUM/X LINK NOT *1 @MV) ;
REMOVE (@%jn) (0 @<OC) (*1 @<ACC LINK 0 @N-HUM/X LINK NOT *-1 @MV) ;
REMOVE (@%jn) (0 @<PRED) (*-1 @NOM-HEAD LINK 0 @N-HUM/X) ;
REMOVE (@%jn) (0 @PRED>) (*1 @SUBJ> LINK 0 @N-HUM/X) ;
```

With the same lexical and syntactic information as used in the head-dependent disambiguating rules, semantic **modifier-head disambiguation** can be achieved, too. In the rule examples below, the atomic semantic features H/h and X/x of nouns are used as target (rather than context), and checked against the semantic class of possible adnominal adjectives, with <jh> favouring H/X, and <jn> favouring h/x:

```
REMOVE (@=H) (*-1 @%jn LINK 0 @>N LINK NOT 0 @%jh LINK NOT *1 @NON->N) ;
REMOVE (@=H) (*1 @%jn LINK 0 @N< LINK NOT 0 @%jh LINK NOT *-1 @NON-N<) ;
REMOVE (@=h) (*-1 @%jh BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jn) ;
REMOVE (@=h) (*1 @%jh BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jn) ;
REMOVE (@=X) (*-1 @%jn BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jh) ;
```

[227] In the case of clause level predicatives, this is not, strictly speaking, a case of head-modifier disambiguation, since the syntactic head is the main verb, but both morphologically (number and gender agreement) and semantically (information structure), there is a link between clause level predicatives and the subject or object.

```
REMOVE (@=X) (*1 @%jn BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jh) ;
REMOVE (@=x) (0 @=X) (*-1 @%jh BARRIER @NON->N LINK 0 @>N LINK NOT 0 @%jn);
REMOVE (@=x) (0 @=X) (*1 @%jh BARRIER @NON-N< LINK 0 @N< LINK NOT 0 @%jn) ;
```

Since the CG rule set keeps reiterating as long as further disambiguation can be achieved, it is not necessary to formulate similar rules for all +HUM prototypes individually: The discarding of the H- or h-feature will propagate to other atomic semantic features and all relevant prototype bundles by means of semantic feature inheritance rules as described in chapter 6.5.1. Of course, such generalisation is not mandatory, and in order to reduce the error rate of the parser's semantic module, individual prototype rules and context conditions may later be added to the general rules.

6.5.3 Parsing level interaction in polysemy resolution

A particularly elegant and "incremental" solution for polysemy resolution of semantically ambiguous words is the semantic exploitation of "lower level parsing information" (morphological form and syntactic function), which the system already *has* disambiguated.

Lexicographically this approach can be implemented by means of what I will call (polysemy-) discriminators, a concept reminiscent of the discriminators (style, register, diachronicity, dialect etc.) used in ordinary paper dictionaries (cp. Bick, 1993). Drawing on subsequent levels of parsing, the following types of *pre*-semantic discriminators are used:

- **word class subcategory discriminators**

An example is the distinction <artd> DET vs. <dem> DET for the determiner 'o/a/os/as', where the translation is 'the' and 'this/those', respectively.

- **inflexional discriminators**

Sometimes, nouns change their basic meaning in the plural, so number inflexion helps to resolve polysemy, as in *costa N S* ('coast') vs. *costas NP* ('back'). Verbs, when inflected as past participles (PCP), sometimes acquire a somewhat different, "adjectival", meaning:

V VFIN		V PCP	
abastar	*forsyne (supply)*	***abastado***	*velforsynet, velsitueret (well off)*
aborrecer	*afsky (detest)*	***aborrecido***	*træls (irritating)*
brigar	*slås (fight)*	***brigado***	*vred (angry)*
calar	*tie (not speak)*	***calado***	*tavs (quiet)*

Another, rarer, example is the verb *saber*, which translates 'knew' in the PS tense, but 'got to know' in the IMPF tense.

- **syntactic function discriminators**

A number of Portuguese adjectives has two meanings depending on whether they occur prenominally (@>N) or postnominally (@N<), for instance:

@>N	@N<	
novo	*'new [another], nouveau'*	*'new [just produced], neuf'*
grande	*'big, famous, stor'*	*'big, high, høj'*
raro	*'rare, sjælden'*	*'strange, mærkelig'*
triste	*'lousy, ussel'*	*'depressed, sørgmodig'*

- **valency instantiation discriminators**

This is an abundant and very useful group of discriminators. Not least, many verbs allow polysemy discrimination by valency instantiation, as *tirar*, which is translated 'pull' when used monotransitively with a direct object (<vt>), but can mean 'shoot [at]' with a prepositional object with 'em' (<em^vp>). Another example is *viver*, where the Danish translation is 'leve' ('live') in the <vi> case, but 'opleve' ('experience') with <vt> valency.

Note that ordinary word class discrimination (e.g. V vs. ADV for 'como') need not be expressed by means of discriminators, since a difference in word class leads to two different lexicon entries, i.e. the word form in question is treated as covering two *lexemes* (each with its own semantics) – and disambiguated accordingly.

As an example for how different kinds of polysemy discriminators can work together in one word, let's look at the verb *saber* , meaning 'to know' when inflected in the imperfeito tense, but 'get to know' in the *perfeito* tense. Here the difference in aspect is expressed lexically (or, rather, phrasally) in English, but by means of a tense distinction in Portuguese. Thus, morphological information can be exploited for a semantic purpose. Word class function could be used, too: if *saber* appears as an auxiliary (@AUX), it means 'to be able to'. Finally, syntactic information from other constituents of the clause can be used in order to instantiate one of several lexically possible valency patterns of the verb *saber*: While both 'to know' and 'get to know' ask for direct objects, the translation alternative 'to taste' is to be chosen for adverbial complements (bem/mal - 'good'/'bad'), and 'to taste of' before a prepositional object introduced by the preposition *'a'*.

(11) **saber V**

@MV, IMPF, <vq><vt>	'know'
@MV, PERF, <vq><vt>	'get to know'

@AUX, <+INF>	'know how to, can'
@MV, <va>	'taste'
@MV, <a^vp>	'taste of'
@MV, <de^vp>	'know of'

[@ = syntactic function: MV =main verb, AUX=auxiliary; <> =valency: <vt> =transitive, <+INF> governs infinitive, <va> =with adverbial object, <vp> =with prepositional object, a^ =preposition "a", de^ =preposition "de"; morphology: IMPF =imperfeito tense, PERF =perfeito tense]

Finally, monosemous (or semantically already disambiguated) words can help disambiguate those that are (still) semantically ambiguous. Thus the Portuguese preposition *'de'* translates as 'fra' (‘from’) if its argument is a place (+LOC), but 'af' (‘made of’) if the argument of the preposition is a word denoting a substance (e.g. *de ouro* 'af guld' - 'made of gold'). The genitive is to be used if the complement is a person (+HUM: *o cachorro do homem* 'mandens hund' - 'the man's dog'). Again, the relevant discriminators have to be introduced into the lexicon, in the form of semantically enriched valency information (so-called selection restrictions).

(12) The following sentence illustrates some of the possibilities:

apesar=**de**	[apesar=de] <sam-> PRP @ADVL> 'på trods **af**' (in spite of)	*
a	[a] <-sam> <art> DET F S @>N 'den' (the)	
advertência	[advertência] <s> N F S @P< 'råd' (advice)	
de	[de] <sam-> <+hum> PRP @N< '**(genitiv)**' (of)	*
o	[o] <-sam> <art> DET M S @>N 'den' (-)	
meu	[meu] <poss 1S> DET M S @>N 'min' (my)	
pai	[pai] <fam> N M S @P< 'far' (father)	
,		
que	[que] <rel> SPEC M/F S/P @#FS-N< 'som' (who)	
não	[não] ADV @ADVL> 'ikke' (not)	
gosta	[gostar] **<de^vp>** <vH> <ink> V PR 3S IND VFIN @FMV 'kunne lide' (liked)	
de	[de] <sam-> PRP @<PIV '**(objekt)**' (-)[228]	*
a	[a] <-sam> <art> DET F S @>N 'den' (-)	
minha	[meu] <poss 1S> DET F S @>N 'min' (my)	
nova	[novo] <ante-attr> <jn> ADJ F S @>N 'ny' (new)	
vida	[vida] <feat> <per> N F S @P< 'liv' (life)	
,		
comprei	[comprar] <vt> <vH> <ink> V PS 1S IND VFIN @FMV 'købe' ([I] bought)	
uma	[um] <quant2> <arti> DET F S @>N 'en' (a)	
carroçada	[carroçada] <qus> N F S @<ACC 'læs' (load)	
de	[de] <quant+> PRP @N< '**(partitiv)**' (of)	*
coisas	[coisa] <cc> <ac> N F P @P< 'ting-1' (things)	
$,		

[228] In the case of a prepositional object, the actual Danish preposition used - or, in this case, the zero-preposition option - is read directly from the relevant translation equivalent listed with the valency bearing head verb.

por=exemplo	[por=exemplo] <adv> <+NP> PP @<ADVL 'fx' (e.g.)	
um	[um] <quant2> <arti> DET M S @>N 'en' (a)	
fato	[fato] <tøj> <AA> N M S @ACC< 'habit' (suit)	
de	[de] <+mat> PRP @N< '**af**' (of)	*
lã	[lã] <cm> <stof> N F S @P< 'uld' (wool)	
preta	[preto] <col> <jn> ADJ F S @N< 'sort' (black)	
que	[que] <rel> SPEC M/F S/P @SUBJ> @#FS-N< 'som' (which)	
veio	[vir] <va+DIR> <sN> V PS 3S IND VFIN @FMV 'komme' (came)	
de	[de] <sam-> <+top> PRP @<ADV '**fra**' (from)	*
a	[a] <-sam> <art> DET F S @>N 'den' (-)	
*argentina	[Argentina] <top> PROP F S @P< 'Argentina' (Argentine)	
de	[de] <+V> PRP @N< '**med**' (by)	*
avião	[avião] <fly> N M S @P< 'fly' (plane)	
em=menos=**de**	[em=menos=de] <c> PRP @<ADVL 'på mindre **end**' (in less than)	*
uma	[um] <card> NUM F S @>N 'een' (a)	
semana	[semana] <dur> <num+> N F S @P< 'uge' (week)	
.		

The relevant lexicon entry first lists a number of valency and semantic context options for the preposition *'de'* , and then indicates which translation is to be chosen in case one or the other of these polysemy-discriminators is instantiated (i.e., survives all disambiguation-constraints in the grammar). Information about syntactic function - from the "next lower" parsing level - (like @<ADVL [adverbial], @N< [postnominal modifier] or KOMP< [comparative complement]) can be used as discriminators, too:

de PRP <komp><corr><+hum><+mat><+top><+V><+feat><+il><+tøj><quant+>

1. @N<	af	*(default postnominal translation)*
2. <quant+> @N<	(partitive)	*(after quantitiva)*
3. <+mat> @N<	af	*(postnominally before substance word)*
4. <+hum> @N<	(genitive)	*(postnominally before proper nouns or person-words)*
5. <+feat><+tøj> @N<	med	*(before features or clothing)*
6. @<ADVL @ADVL> @ADVL	fra	*(default adverbial translation)*
7. <+V><+il> @<ADVL	med	*(before vehicles or tools)*
8. <+top> @<ADVL	fra	*(adverbially before toponyms or other place words)*
9. <+hum> @<ADVL	fra	*(adverbially before proper nouns or person-words)*
10. <komp> @KOMP<	af, blandt	*(as comparative complement: "the biggest of ...")*
11. <komp><corr> @KOMP<	end	*(as correlative comparative complement: "bigger than"*

Note that the MT engine will choose the *first* alternative if an equal number of matches is found for several of the above discriminator groups. Thus, if the CG rules have removed <corr> from the tag string of a comparative 'de', the second last translation, 'af/blandt' will be preferred over the last one, 'end', though both lines still match <komp> and @KOMP<. With Danish as the target language (but not for German, for instance), line (8) and (9) could be fused with each other and with (6), but that would create a problem with institution words, which are tagged both <hum> and <top>, while obviously being able to occur as postnominals [@N<] or comparative complements [@KOMP<], too. <+hum> and <+top> would therefore have to be added on these lines, too, in order to prevent these alternatives from having their discriminators outnumbered by case (6) even in cases where 'de' is not, *syntactically*, tagged as adverbial [@ADVL].

For the noun *'fato'* the lexicon offers the following polysemy discriminators, some of which are valency instantiations (<+que>, <+de+que>, <+de+INF>), some semantic prototypes (<ac><tøj><AA>) and one a feature listing of all *those* atomic semantic features, that the prototypes mentioned cover *jointly* (e.g. A = +ANIM, a = ÷ANIM).

fato N M <ac><tøj><AA><+que><+de+que><+de+INF><=EecIiJjAahmNnvpsdxflt=>

__ <ac><+de+que><+de+INF>	kendsgerning (fact)
__ <tøj>	habit, kostyme (suit)
__ <AA>	flok {fx geder} (flock {e.g. of goats})
fato=de=banho N M	badedragt (bathing suit)
fato=de=macaco N M	kedeldragt (dungarees, jump suit)

In the sentence *'Um fato de ovelhas corria no campo'* the parser uses 8 rules to disambiguate the polysemy of *'fato'*, - not included those rules for *other* words in the sentence that created the necessary unambiguous context conditions:

(12a)

*um	[um] <quant2> <arti> DET M S @>N 'en' (a)
fato	[fato] **<AA>** N M S @SUBJ> '**flok**' (flock)
de	[de] <quant+> PRP @N< '(partitiv)' (of)
ovelhas	[ovelha] **<zo>** N F **P** @P< 'får' (sheep)
corria	[correr] <vi> V IMPF 1/3S IND VFIN @FMV '**løbe**' (ran)
em	[em] <sam-> <+top> PRP @<ADVL 'i' (across)
o	[o] <-sam> <art> DET M S @>N 'den' (the)
campo	[campo] <BB> <top> <topabs> N M S @P< 'mark-2' (field)

A test run of the parser with rule-tracing shows that the first three rules to be used are valency instantiation rules:

*REMOVE (<+de+que>) (*1 CLB/SB LINK NOT 0 QUE-KS) ;*

... if the following clause boundary is not the conjunction 'que'
*REMOVE (<+que>) (*1 NON-ADV LINK NOT 0 QUE-KS) ;*
... if the first following non-adverbial word is not the conjunction 'que'
*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*
... if there is no preposition complementing infinitive before the next clause boundary or infinitive governing preposition

Next, positive (capital letters) or negative (small letters) semantic features are removed. The only "real" rule is the first one, stating that *'fato'* in this sentence has the capacity of moving; the other rules are "reflex" conclusions based on the feature +MOVE.

*REMOVE (<i>) (0 @SUBJ> AND <I>) (*1 @MV BARRIER CLB-ORD LINK 0 V-MOVE);*
if it is subject and there follows a MOVE main verb without interfering clause boundary, then the target word has the feature +MOVE
REMOVE (<j>) (NOT 0 <i>) ;
if it can move (active movement, +=I, ÷ =i), then one can move it (passive movement, + =J, ÷ =j).
REMOVE (<tøj>) (NOT 0 <i>) ;
it cannot belong to the prototype 'clothing' if it can move.
REMOVE (<e>) (NOT 0 <i>) ;
it cannot be abstract (i.e., a non-entity) if it can move.
REMOVE (<ac>) (NOT 0 <e>) ;
it cannot belong to the prototype 'abstract countable', if it isn't abstract.

[*atomic semantic features:* <i> = ÷move, <I> = +move, <j> = ÷passive movement, <J> = +passive movement, <e> = ÷abstract, <E> = +abstract; *semantic prototypes:* <tøj> = clothing, <ac> = abstract countable]

In the expression *'Um fato de lã preta'* 4 of the same rules are used, plus a rule recognising the postnominal substance context (made of black wool) to the right.

(12b)

*um	[um] <quant2> <arti> DET M S @>N 'en' (a)
fato	[fato] **<tøj>** <AA> N M S @NPHR '**habit**' (suit)
de	[de] <+mat> PRP @N< 'af' (of)
lã	[lã] <cm> **<stof>** N F **S** @P< '**uld**' (wool)
preta	[preto] <col> <jn> ADJ F S @N< 'sort' (black)

*REMOVE (<+de+que>) (*1 CLB/SB LINK NOT 0 QUE-KS) ;*
*REMOVE (<+que>) (*1 NON-ADV LINK NOT 0 QUE-KS) ;*
*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*
*REMOVE (<e>) (*1 PRP-DE BARRIER NON-POST-N LINK 0 @N< LINK 1 @P< LINK 0 <M> AND <E>);* it cannot be abstract, if the preposition 'de' follows without other interfering material than postnominals, and if this preposition has af directly adjacent (i.e. article-less) argument of the type +MASS and -ABSTRACT (i.e. cloth, substance).
REMOVE (<ac>) (NOT <=e>) ;

Even where none of the semantic rules applies, there can still be a valency instantiation that resolves the polysemy[229]. Here, it is the valency tag <+de+que> that survives all constraints.

(12c)

*o	[o] <art> DET M S @>N 'den' (the)
fato	[fato] **<ac>** <tøj> <AA> **<+de+que>** N M S @NPHR '**kendsgerning**' (fact)
de	[de] PRP @N< '(af)' (-)
que	[que] KS @SUB @#FS-P< '**at**' (that)
sua	[seu] <poss 3S/P> DET F S @>N 'hans' (his)
namorada	[namorada] <title> N F S @SUBJ> 'kæreste' (girl-friend)
tem	[ter] <vt> <sH> V PR 3S IND @FMV 'have' (has)
um	[um] <quant2> <arti> DET M S @>N 'en' (a)
emprego	[emprego] <stil> <ac> N M S @<ACC 'arbejde' (job)

*REMOVE (<+que>) (*1 NON-POST-N LINK NOT 0 QUE-KS) ;*

if the first word after possible postnominals is not the conjunction 'que'.

*REMOVE (<+de+INF>) (*1 CLB/SB OR <+PRP+INF> BARRIER @#ICL-P<) ;*

if there is, without interfering non-finite preposition complement, a clause or sentence boundary or a word that can govern a preposition followed by an infinitive.

Finally, to give a kaleidoscopic view of the different translation oriented semantic disambiguation techniques, let’s have a look at the polysemic word form ‘revista’, which can mean any one of the following:

(a) revista N F S	‘tidsskrift’ (magazine)
(b) revista N F S	‘inspektion’ (inspection)
(c) revista V PCP F S	‘gennemset’ (inspected)

Though morphosyntactic disambiguation is capable of separating (a) and (b) from (c), it cannot resolve the noun polysemy, both (a) and (b) have the same gender and inflexion pattern. This is why they are treated as one - polysemous - lexeme in the parser’s lexicon. Semantic disambiguation is, however, possible: The ‘magazine’ reading can be arrived at by valency instantiation (<+n>, i.e. proper noun argument) or by head-argument selection of the semantic prototype class <rr> (readable) in the face of a main verb of the READ-class:

N "magazine"

Leu	*a*	*revista*	*VEJA*
'ler' (V-READ)	→	<rr>	

[229] The parser chooses those translation alternatives that have the most un-discarded discriminators left. In case this criterion results in ambiguity, the first reading on the list (the semantic default) is chosen in a heuristic fashion.

<+n> ← **PROP**

([he] read the news magazine VEJA)

N "inspection"

uma *rápida* *revista* *da tropa/casa/plano* *pelo coronel*
<+de> ← **'de'**
<CP> (action):
+CONTR ← **+HUM**
@>N <temp> → +VERBAL ← **@N<PASS (SUBJ)**
→ +PERF ← **@N< (OBJ)**
+TEMP

(a quick inspection/review of the troops/house/plan by the major

The 'inspection' reading is favoured by the valency instantiation of <+de> (genitivus objectivus), and by modifier-head disambiguation of the atomic semantic features V (+VERBAL) and T (+TEMP) in the face of a temporal adjective modifier ('rápida' - 'quick'). The 'magazine'-reading is negative for both features (v and t). A human agent of passive argument ('pelo coronel') supports the semantic prototype <CP> (action), where @N<PASS matches the atomic feature V (+VERBAL), and +HUM matches C (+CONTROL).

A comparison of the positive atomic semantic features of the two readings shows that the prototypes <rr> (book) and <CP> (action) differ in four categories, C-V-P-T, of which three have been used in the example:

	C (control)	N (countable)	V (deverbal)	P (perfective)	X (human adj.)	T (temporal)
<rr> (book)		+			+	
CP (action)	+	+	+	+	+	+

The verbal form 'revista V PCP F S' covers two etymologically different base form lexemes, *rever ('re-ver' - 'see again' or 'inspect')* and *rever (Latin 'repere' - 'leak through')*. The only difference lies not in inflexional morphology, but in the valency pattern of these verbs, which is why the lexicon treats them as one - polysemic - lexeme. With the meaning 'inspect/see again' the verb is transitive (<vt>), with the meaning 'leak through' it is intransitive (<vi>). Passive constructions, and the simple fact that the participle is inflected (here: F S), disallow the intransitive reading:

V "to inspect"

A *casa* *foi* → *revista* *ontem.*
'ser+PCP' <vt> PCP

A *casa* *revista*

<vt>← **PCP @N<**

revista

<vt>← **PCP F (not PCP M)**

(The house was inspected yesterday)

Therefore, valency instantiation can be handled locally, by morphology alone, without context. With a finite inflexion, however, more context is needed to make the distinction. Obviously, one criterion is the presence or absence of a direct object (@<ACC). Another context check could be based on the H/h semantic atomic feature at the subject, since ‘inspect’ demands a *human* subject, while ‘leak through’ demands a *non-human* (more precisely, inanimate) subject:

O coronel *reviu* *a casa*

HUM-AG → V-SENSE

<vt> ← **@<ACC**

(The major inspected the house)

<u>V "to leak through"</u>

As suas verdadeiras intenções com o plano *reviram* .

not HUM-AG → V-PROCESS (NOT V-SENSE)

<vi> ← **not @<ACC**

(His true intentions regarding the plan leaked)

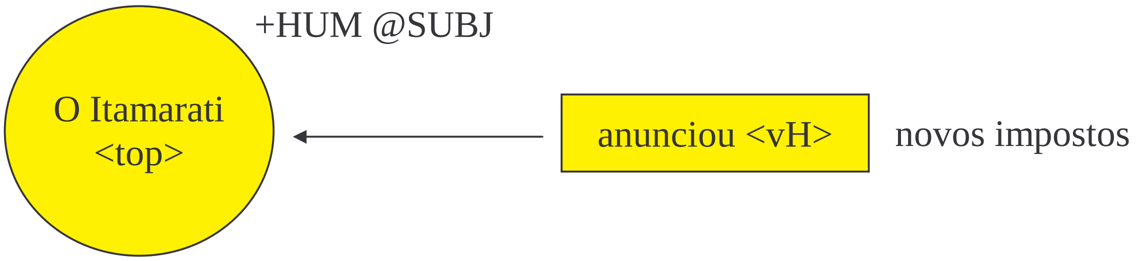
6.5.2 Metaphorical interpretation of semantic tag mismatches

In the previous chapters, we have seen how a variety of semantic and syntactic tags can be disambiguated in order to resolve polysemy in Portuguese. One inherent assumption has been that semantic disambiguation targets such and only such polysemy as has been specified, *a priori*, by a list of differing translation equivalents, in a bilingual lexicon with discrimination markers.

Metaphor in the source language, therefore, will only be treated (by CG rules), if it has lexical consequences in the target language. A good deal of the metaphorical potential of a language, of course, is *live*, i.e. has never been lexicalised (bypassing the dictionary), has maybe never even been used before, and exploits the listener’s/reader’s knowledge of the world rather than his lexical competence. Such metaphors are usually not a real MT-problem, since the metaphorical reading does not affect translation *per se* - rather, the reader/listener has to perform some kind of *conceptual* translation in her head, largely independent of the *language* in question.

Moving beyond the translational level, I would like to hold that even such metaphorical usage can be detected and decoded, to a certain degree, within the suggested framework of my parser's semantics. One of the main tools for semantic disambiguation, as described above, have been rules trying to "mismatch" semantic prototype or atomic feature tags (on nouns, e.g. <H>, <mad>) with the semantic context established by verbs (semantic valency, e.g. <vH>, <+ACC-hum>, V-MOVE) and adjectives (as modifiers, e.g. <jH>, <nH>). Now, such semantic mismatches have been introduced as *tools*, but *could* - in the spirit of progressive level parsing - be interpreted as "primary" information themselves. Then, with semantic mismatches becoming part of the system of analysis proper, there would be a certain trade-off between (interpretable) semantic mismatches and semantic disambiguation: Either, discarding some tag readings out of a polysemic range of tags will resolve the mismatch (and yield the correct analysis and translation), or - if there is no polysemy stated in the lexicon - a given semantic mismatch can be *interpreted.* Simply, in the context of a semantic CG-parser, non-lexical metaphors should be defined as *those semantic mismatches that survive semantic disambiguation.* Better still, a classification of metaphors could be based on combining the semantic type (tag) *expected* by the lexicon with the reading *forced* by the mismatching constituent slot, with metaphoric transfer moving from the latter to the former.

Thus, <vH> verbs (verbs unambiguously asking for +HUM subjects), if used with an unambiguously -HUM subject, will lead to a (nominal) metaphor built on non-human->human transfer. One could say that the verb projects a +HUM reading onto any filler of its subject valency slot, the semantic properties of the slot being stronger than those of the filler. Consider the following example:



The Itamarati [palace of government] announced new taxes

Here, the verb *anunciar* ('proclaim') creates a "subject space" that is destined as +HUM. Provided that the syntactic CG module establishes a correct subject-predicator link, this semantic projection can have two effects:

(a) With a polysemous filler, it may help resolve the polysemy. In *O leão penalizou a especulação* ('The **tax-lion** punished speculation'), *leão* is a Brazilian symbol for the finance department, and as such does not belong to the semantic prototype <A> (animal), but to <inst> (human institution), which is the reading selected by the +HUM subject projection of *penalizar*.

(b) With a semantically unambiguous filler, that is -HUM, the semantic projection of *anunciar* (or *penalizar*) creates +HUM as a metaphoric reading. Here, a place name is metaphorically turned into an institution (metonymical transfer).

Incidentally, some metaphors would be treated in this way *in spite of* having become "dictionary-worthy" due to their frequency, simply because a metaphor may be so universal that it will not materialise in a lexicon that defines polysemy in terms of bilingual discrepancy. In *A estrela hesitou*, for instance, *estrela* ('star') would be read as +HUM rather than <star>, but still be translated in the same way.

In the examples, semantic projection is interpreted as metaphoric transfer in the direction of valency, from head to dependent. This is quite common also in the case of non-valency projections: In *um dia triste* ('a sad day'), it is the -HUM head *dia* that projects a semantic change in the modifier *triste*. A day is not sad the same way a human being is, but the projection is still far more acceptable than imagining a "human-sentient" kind of "day"[230].

Still, the conclusion that metaphoric transfer is preferably induced by head-to-dependent projection, is questionable: Dependent-to-head projection is not rare at all, as in *o coronel explodiu* ('the colonel exploded'), where *explodir* prototypically is <sN> (i.e. does not take human subjects), but is interpreted as in a humanoid manner when the feature <sH> is forced upon it by a semantically unambiguous +HUM subject, with the resulting concept being one of "explosive behaviour" rather than a soldier being torn to pieces like a bomb. Rather than dependency-direction, it seems to be semantic criteria

[230] As a matter of fact, my tag set provides for this case with a feature <X> stating the "normality" of "human" modifiers with nouns of a certain class. <X> would not be used with 'dia', though, but with certain more typical prototype classes, such as <rule> and <occ>.

that make one projection direction more probable than another. Thus, it is my impression that - in general - there is a +HUM-bias in metaphoric transfer, i.e. the human part of a semantic mismatch “wins” over the non-human part, irrespective of which one of the two syntactically functions as head or dependent, respectively. Another regularity seems to be that concrete-to-abstract metaphorical transfer in nouns is more common than the inverse:

um ***monte*** *de possibilidades (‘a mountain of possibilities’, ‘et hav af muligheder’)*
exauriu *as suas finanças (‘he exhausted his funding’, ‘han udtømte sine midler’)*

Finally, let’s return to the trade-off between disambiguation and metaphorically interpretable semantic mismatches. Both are the results of the CG interaction between lexical assumptions (hand crafted or corpus derived lexical tags) and CG matching (SELECT) or mismatching (REMOVE) rules. The difference is that polysemy resolving disambiguation is the active product of a semantic CG, whereas surviving lexical/valency mismatches are a passive by-product, to be interpreted (as metaphorical) *a posteriori*. The interesting thing is, that - in a CG framework - semantic selection restrictions can be exploited whether they work or not: If they work, polysemy is resolved, if they don’t, we get metaphor. This is radically different from traditional generative grammar implemented in a declarative programming language, where a mismatch of selection restrictions conventionally is interpreted, in “all-or-nothing” terms, as a “no-parse” situation meaning that the sentence in question is not part of the language system described by the grammar. Creative, *new* metaphors are not easy to capture in classical generative systems with semantic selection restrictions, since every possible combination has to be provided for already in the lexicon part of the system. The English sentence ‘The sea ate the coast’, for instance, would fail the rewriting rules, since there is no edible match (<mad>, ‘food’) for the *eating* verb (V-EAT). In a CG system, on the other hand, the analysis is first successfully assembled at the syntactic level (working on verbal transitivity and word order), and then fails gracefully at the semantic level - only to produce a metaphoric reading exactly by doing so. Thus, contrary to a commonly held view regarding the existence of metaphor as a reason for *not* using selection restrictions, they do work quite well in a CG framework. In the example, the *eating* verb projects “foodhood” (<mad>) onto its direct object (‘coast’ @<ACC), - actually *helping* metaphorical understanding rather than hindering it[231].

[231] Only in the case of metaphors involving highly polysemous words, there is a danger of disambiguation getting the better of metaphor. For the sentence above, one could imagine a pun with a direct object that among its meanings had *both* that of a place *and* that of some edible substance, or a verb, that besides ‘eat’ could mean ‘reach’. This having been said, it is not easy to come up with many such sentences introspectively, and they are probably even rarer in actual corpus statistics. And, methodologically more relevant, the disambiguation-interpretability trade-off problem is inherently the same for every level of CG analysis, and not specific of the treatment of metaphor. With the “right” combination of ambiguous words, CG rules may discard readings that appear incompatible at a lower level of analysis, though they might be interpretable at a higher level. For example, in the erroneous English sentence *‘women talks more than men*, ‘women’ and ‘talks’ disagree in number,

and a morphological CG module may choose to discard the verb reading of ‘talks’ in favour of a plural noun reading, before the *syntactic* CG module gets a chance to make ‘women’ a subject (@SUBJ), ‘talks’ a main verb (@MV), - and the sentence itself a case of learner’s English.

7

The applicational level: Teaching and translating on the internet

At http://ling.hum.au.dk and http://visl.hum.sdu.dk (current) an integrated interactive user interface is presented for the automatic analysis of running Portuguese text and for teaching grammatical analysis through the Internet medium. Though the system's internal grammatical tools - for reasons of robustness, efficiency and correctness - are based on the Constraint Grammar formalism, users are free - at the applicational level - to choose from a variety of notational filters, supporting different descriptional paradigms. The original kernel of programs was built around the multi-level parser for Portuguese described in this text. A similar system has since been implemented for English and Spanish as part of the ongoing VISL-project at Odense University.

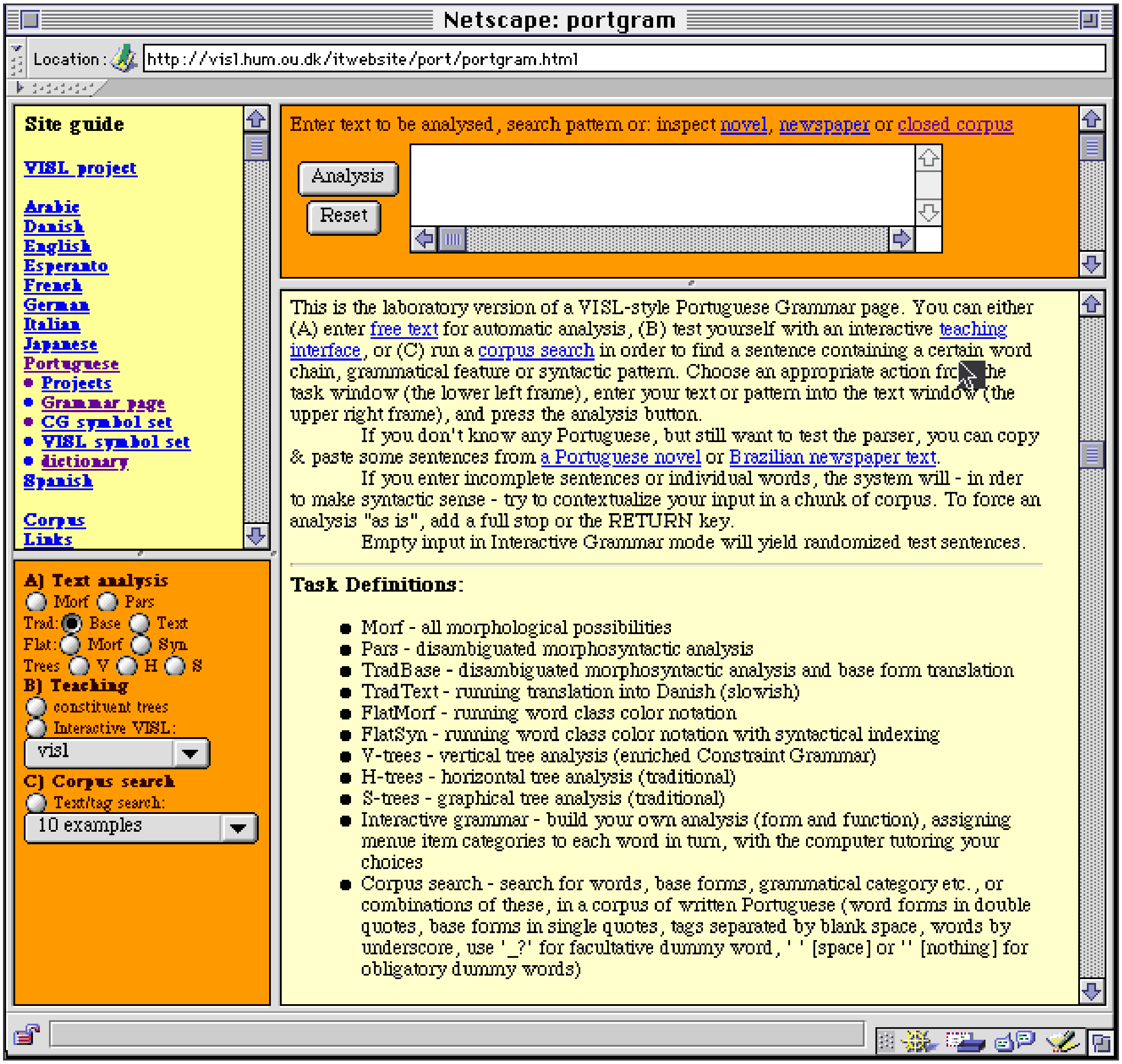
7.1 Progressive Level Parsing as a real time tool on the Internet

One obvious application of a Constraint Grammar based parser is as a real time tool: The technique is so robust and fast that “live” analysis is possible, and so error-free that post-editing becomes dependable for many purposes. Once started, the parsing programs handle text at many times reading speed (hundreds or thousands of words per second, depending on the level of analysis), which is an important condition for applications where the parser is to be integrated in other programs, for instance text processors with spelling checkers, search engines or grammar-tutors. Real-time performance also allows internet-applications, and as a first step, I have made the different parsing stages themselves (as described in chapter 1) available on the internet. Full morphological analysis, disambiguated PoS-tagging, syntactic flat dependency parsing, tree structure parsing, and bilingually oriented semantic disambiguation can thus be run individually:

- **Morf -** all morphological possibilities (preprocessor, lexicon and analyser)
- **Pars** - disambiguated morphosyntactic analysis (Constraint Grammar)
- **TradBase -** disambiguated morphosyntactic analysis plus base form translation
- **TradText** - running translation into Danish
- **FlatMorf -** running word class colour notation
- **FlatSyn** - running word class colour notation with syntactic indexing

- **V-trees** - vertical tree structure analysis (line-based enriched CG notation with constituent mark-up)
- **H-trees** - horizontal tree analysis (traditional constituent tree structures)
- **S-trees** - graphical constituent tree analysis (with notational filtering choices)

(1) The Portuguese grammar page



Enter text to be analysed, search pattern or: inspect novel, newspaper or closed corpus

A) Text analysis
○ Morf ○ Pars
Trad: ● Base ○ Text
Flat: ○ Morf ○ Syn
Trees ○ V ○ H ○ S
B) Teaching
○ constituent trees
○ Interactive VISL:
visl
C) Corpus search
○ Text/tag search:
10 examples

This is the laboratory version of a VISL-style Portuguese Grammar page. You can either (A) enter free text for automatic analysis, (B) test yourself with an interactive teaching interface, or (C) run a corpus search in order to find a sentence containing a certain word chain, grammatical feature or syntactic pattern. Choose an appropriate action from the task window (the lower left frame), enter your text or pattern into the text window (the upper right frame), and press the analysis button.

If you don't know any Portuguese, but still want to test the parser, you can copy & paste some sentences from a Portuguese novel or Brazilian newspaper text.

If you enter incomplete sentences or individual words, the system will - in rder to make syntactic sense - try to contextualize your input in a chunk of corpus. To force an analysis "as is", add a full stop or the RETURN key.

Empty input in Interactive Grammar mode will yield randomized test sentences.

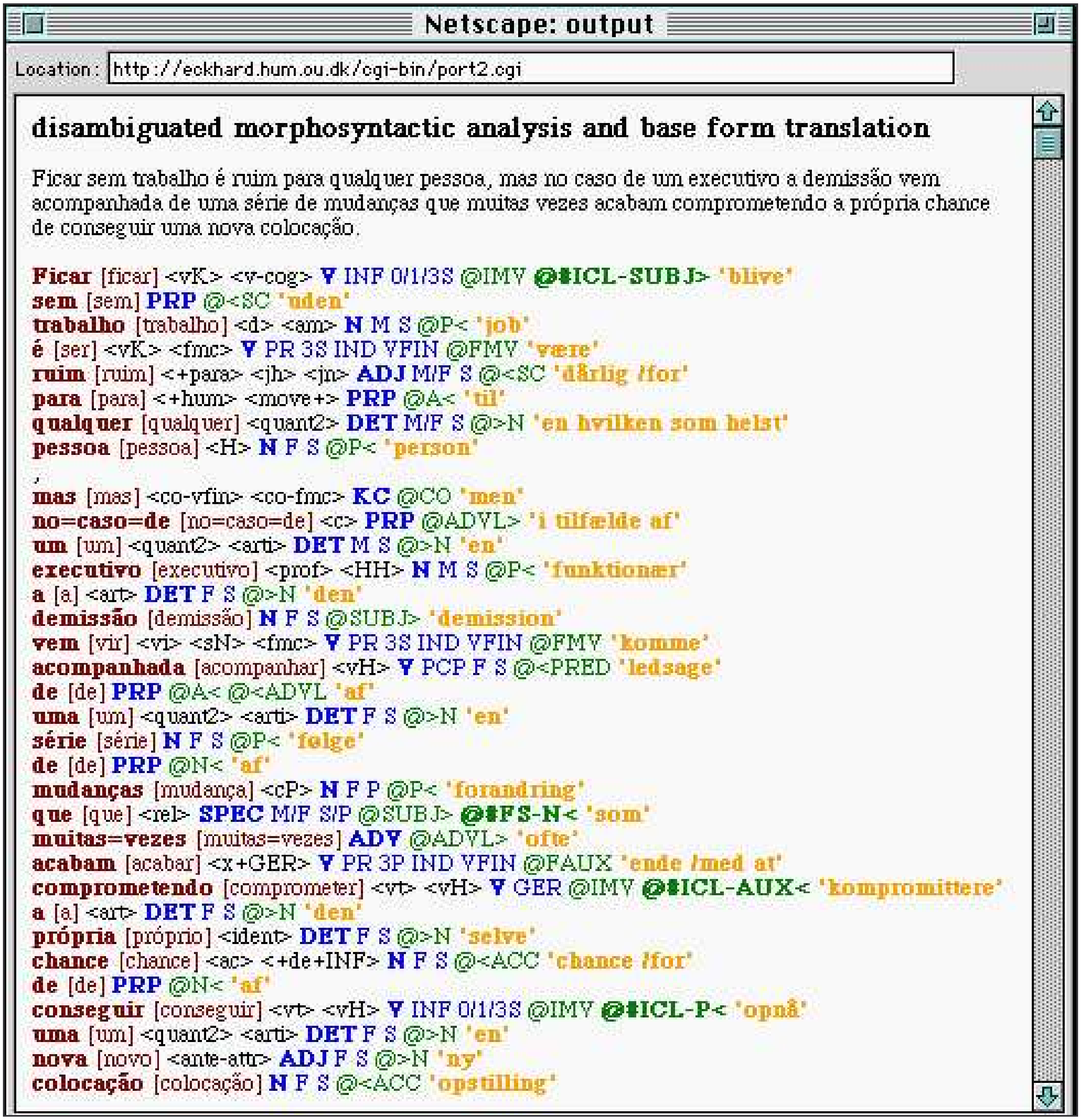
Task Definitions:

- Morf - all morphological possibilities
- Pars - disambiguated morphosyntactic analysis
- TradBase - disambiguated morphosyntactic analysis and base form translation
- TradText - running translation into Danish (slowish)
- FlatMorf - running word class color notation
- FlatSyn - running word class color notation with syntactical indexing
- V-trees - vertical tree analysis (enriched Constraint Grammar)
- H-trees - horizontal tree analysis (traditional)
- S-trees - graphical tree analysis (traditional)
- Interactive grammar - build your own analysis (form and function), assigning menue item categories to each word in turn, with the computer tutoring your choices
- Corpus search - search for words, base forms, grammatical category etc., or combinations of these, in a corpus of written Portuguese (word forms in double quotes, base forms in single quotes, tags separated by blank space, words by underscore, use '_?' for facultative dummy word, ' ' [space] or '' [nothing] for obligatory dummy words)

The page prompts the user to enter Portuguese text, and less inventive souls (or curious people without any deeper knowledge of Portuguese) are offered a default example as well as sample text or newspaper links for cutting and pasting. Next, there is a choice between the different tasks and levels of analysis, from simple tagging ('Portmorf') over morphosyntactic disambiguation ('Portpars') to bilingually motivated polysemy resolution ('Porttrad'), between different notational conventions (verticalised word based CG notation, enriched text ['flatmorf' and 'flatsyn'] with meta-tagging as well as tree-structures ['V-trees' and 'H-trees]') and between notationally filtered levels of descriptional complexity: 'full tag set' and 'traditional tag set' in flat text mode, and ‘CG-style’, ‘VISL-style’ and ‘simplified’ in tree-diagram mode [S-trees].

A "raw" Constraint Grammar analysis with full disambiguation, for instance, yields the following output for the default sentence given:

(2) Full multi-level Constraint Grammar analysis (‘Porttrad’)



disambiguated morphosyntactic analysis and base form translation

Ficar sem trabalho é ruim para qualquer pessoa, mas no caso de um executivo a demissão vem acompanhada de uma série de mudanças que muitas vezes acabam comprometendo a própria chance de conseguir uma nova colocação.

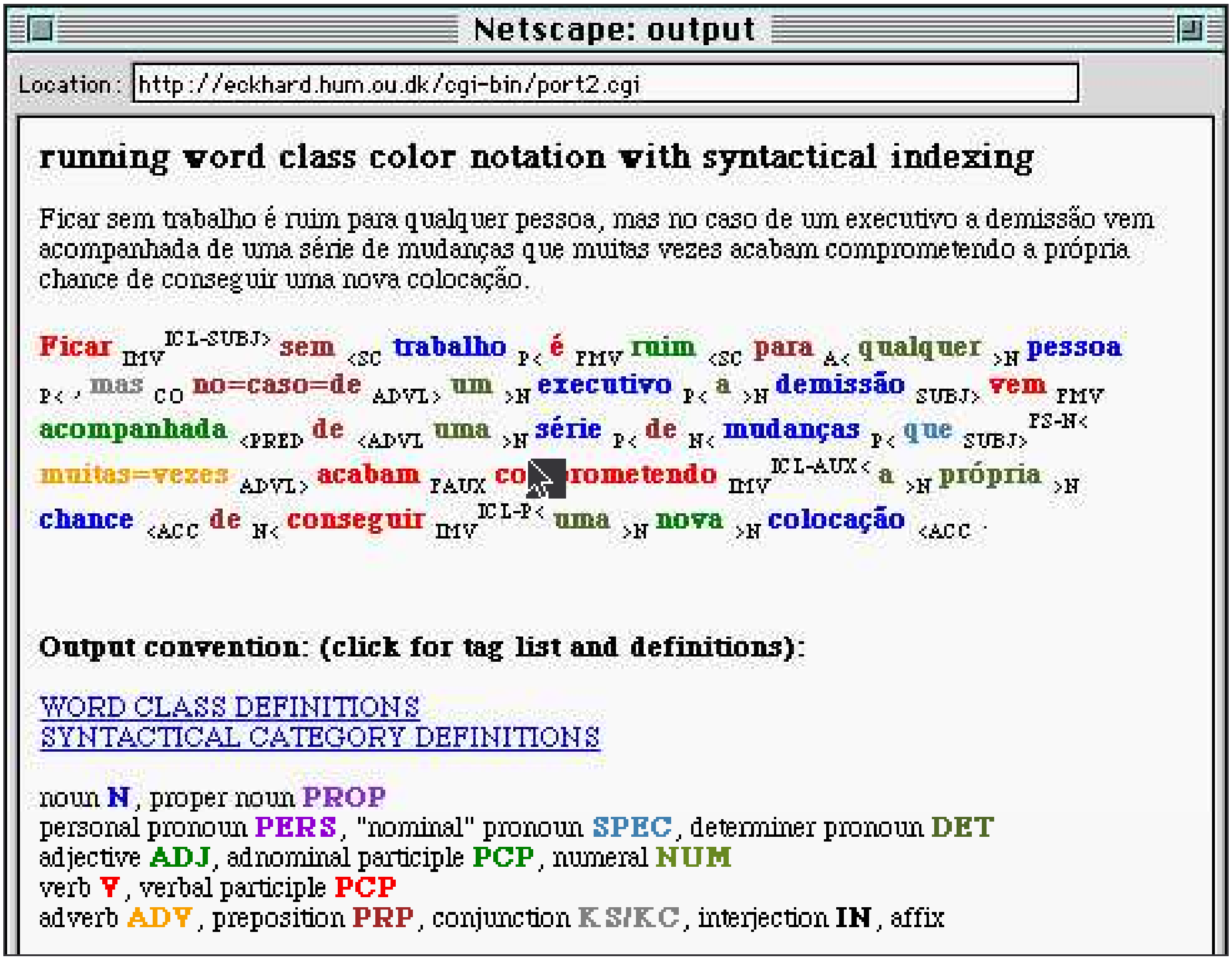
```
Ficar [ficar] <vK> <v-cog> V INF 0/1/3S @IMV @#ICL-SUBJ> 'blive'
sem [sem] PRP @<SC 'uden'
trabalho [trabalho] <d> <am> N M S @P< 'job'
é [ser] <vK> <fmc> V PR 3S IND VFIN @FMV 'være'
ruim [ruim] <+para> <jh> <jn> ADJ M/F S @<SC 'dårlig /for'
para [para] <+hum> <move+> PRP @A< 'til'
qualquer [qualquer] <quant2> DET M/F S @>N 'en hvilken som helst'
pessoa [pessoa] <H> N F S @P< 'person'
,
mas [mas] <co-vfin> <co-fmc> KC @CO 'men'
no=caso=de [no=caso=de] <c> PRP @ADVL> 'i tilfælde af'
um [um] <quant2> <arti> DET M S @>N 'en'
executivo [executivo] <prof> <HH> N M S @P< 'funktionær'
a [a] <art> DET F S @>N 'den'
demissão [demissão] N F S @SUBJ> 'demission'
vem [vir] <vi> <sN> <fmc> V PR 3S IND VFIN @FMV 'komme'
acompanhada [acompanhar] <vH> V PCP F S @<PRED 'ledsage'
de [de] PRP @A< @<ADVL 'af'
uma [um] <quant2> <arti> DET F S @>N 'en'
série [série] N F S @P< 'følge'
de [de] PRP @N< 'af'
mudanças [mudança] <cP> N F P @P< 'forandring'
que [que] <rel> SPEC M/F S/P @SUBJ> @#FS-N< 'som'
muitas=vezes [muitas=vezes] ADV @ADVL> 'ofte'
acabam [acabar] <x+GER> V PR 3P IND VFIN @FAUX 'ende /med at'
comprometendo [comprometer] <vt> <vH> V GER @IMV @#ICL-AUX< 'kompromittere'
a [a] <art> DET F S @>N 'den'
própria [próprio] <ident> DET F S @>N 'selve'
chance [chance] <ac> <+de+INF> N F S @<ACC 'chance /for'
de [de] PRP @N< 'af'
conseguir [conseguir] <vt> <vH> V INF 0/1/3S @IMV @#ICL-P< 'opnå'
uma [um] <quant2> <arti> DET F S @>N 'en'
nova [novo] <ante-attr> ADJ F S @>N 'ny'
colocação [colocação] N F S @<ACC 'opstilling'
```

Here, the running word forms in a sentence (bold face brown) are verticalised and their relevant tags coloured for lexeme (light brown), word class (bold face blue), inflexion (light blue), clause internal syntactic function (light green), clause function (bold face green) and base form translation (yellow). The angle bracketed tags provide additional "secondary" information (which has also been partially disambiguated) about, for instance, subclasses like 'relative', 'interrogative', 'demonstrative' and so on for pronouns, as well as valency patterns used in the context given. By following the relevant links at the bottom of the page the student can find help with regard to category definitions, abbreviations and the like. Contentwise, a Danish user can deduce a rough translation from the Danish equivalents offered as part of the tag string, or he may ask for additional help in the form of a "real", running translation of the sentence ('Portdan', cp. chapter 7.4):

Output as in (2) is close to the grammatical core of the system and combines most of the CG advantages listed in chapter 3. For many purposes, however, this very detailed notation may seem too heavy a tool, especially if the user has no prior experience with Constraint Grammar. According to the principle of naturalness[232], one would prefer a notation as close to ordinary text as possible. That way, sentence context will be easier to grasp, and the interface will feel less "technical" (as intended). I believe to have found such a notation in what I call "enriched text", where running text is "meta-tagged":

(3) "Enriched text" (running text with meta-tagging)

[232] Four basic principles have guided the design of the VISL interface: flexibility, interactivity, naturalness and tutoring.



running word class color notation with syntactical indexing

Ficar sem trabalho é ruim para qualquer pessoa, mas no caso de um executivo a demissão vem acompanhada de uma série de mudanças que muitas vezes acabam comprometendo a própria chance de conseguir uma nova colocação.

Ficar $_{\text{IMV}}^{\text{ICL-SUBJ>}}$ sem $_{\text{<SC}}$ trabalho $_{\text{P<}}$ é $_{\text{FMV}}$ ruim $_{\text{<SC}}$ para $_{\text{A<}}$ qualquer $_{\text{>N}}$ pessoa $_{\text{P<}}$, mas $_{\text{CO}}$ no=caso=de $_{\text{ADVL>}}$ um $_{\text{>N}}$ executivo $_{\text{P<}}$ a $_{\text{>N}}$ demissão $_{\text{SUBJ>}}$ vem $_{\text{FMV}}$ acompanhada $_{\text{<PRED}}$ de $_{\text{<ADVL}}$ uma $_{\text{>N}}$ série $_{\text{P<}}$ de $_{\text{N<}}$ mudanças $_{\text{P<}}$ que $_{\text{SUBJ>}}^{\text{FS-N<}}$ muitas=vezes $_{\text{ADVL>}}$ acabam $_{\text{FAUX}}$ comprometendo $_{\text{IMV}}^{\text{ICL-AUX<}}$ a $_{\text{>N}}$ própria $_{\text{>N}}$ chance $_{\text{<ACC}}$ de $_{\text{N<}}$ conseguir $_{\text{IMV}}^{\text{ICL-P<}}$ uma $_{\text{>N}}$ nova $_{\text{>N}}$ colocação $_{\text{<ACC}}$.

Output convention: (click for tag list and definitions):

WORD CLASS DEFINITIONS
SYNTACTICAL CATEGORY DEFINITIONS

noun **N**, proper noun **PROP**
personal pronoun **PERS**, "nominal" pronoun **SPEC**, determiner pronoun **DET**
adjective **ADJ**, adnominal participle **PCP**, numeral **NUM**
verb **V**, verbal participle **PCP**
adverb **ADV**, preposition **PRP**, conjunction **KS/KC**, interjection **IN**, affix

Here, each line of the CG-notation is condensed into its text kernel, the word form as such, which is all that is left *on* the line. Thus, the impression of running text is recreated. Of the original tags, only syntactic function is retained, with clause internal function as sub-scripts, and clausal function as super-scripts. Word class is retained as meta-notation, too, in the form of colour codes[233] (which are explained at the bottom of the page). Nominal material is tagged in different shades of blue so as to retain NP-coherence in a visual, pedagogically intuitive way. Thus, nouns are blue, proper nouns dark violet and adjectives green. Pronouns match what they are pro-forms for - personal pronouns are coloured light violet, independent “np-substituting” (non-inflecting) pronouns turquoise, and determiner (inflecting) pronouns olive-green. With a grass-green shade, numerals belong in the adnominal modifier (adjective) camp, too. Verbs receive an entirely different ("active") colour, red, so as to make them stick out from the rest of the sentence. Since also infinitives and gerunds are coloured red, the whole verb chain is easily detected. Participles, being a morphological class capable of both

[233] With regard to the colour notation of word classes categories I have been inspired by a similar notation, Gratex, for manually pre-analysed Danish text, described in (Lytje & Donner, 1996).

"verbal" and "adjectival" function, are tagged according to syntactic function - as part of a verb chain they are red, but in adnominal position they become as green as ordinary adjectives. The non-inflecting particle classes, finally, divide the remaining colours among themselves, - adverbs, for instance, are yellow, and prepositions brown.

7.2 Grammar teaching on the Internet: The VISL system

The grammar teaching application described here, has been developed and integrated into the parsing site during my time as project leader for the VISL-project (Visual Interactive Syntax Learning, 1996-1999) at Odense University, where the Portuguese system was used as a point of departure for similar systems in other languages. The multilinguality and practical orientation of the VISL project has raised healthy discussions about notational paradigms and flexibility, which in the case of Portuguese led to the introduction, for teaching purposes, of different levels of grammatical tag sets and syntactic tree types. Being distinct from the original CG tags and dependency markers, this notational variation put the transformational potential of the CG-parser to a test.

7.2.1. Introducing and designing IT-based teaching tools

When trying to introduce IT-based tools into a teaching environment, apart from the hardware problem of there never being enough (compatible) machines at the right place and time, there is the very central problem of psychological resistance against the new medium, simply because it may feel too "technical". Things technical traditionally have a very low acceptance rate in the Humanities, which is where language teaching belongs. Text processors, for example, were widely shunned until the day when they started to use a "non-technical", i.e. graphical, interface. In the same vein, there is the fundamental difference between a human teacher and a computer terminal, - the latter lacks the teacher's naturalness, interactivity, flexibility and tutoring capacities. On the other hand, computers do have evident teaching advantages - they can integrate the senses, making use of colours, pictures and sounds in a more flexible and impressive manner than paper can. Also, a program can "know" more - in terms of facts and data and within a well defined, specific field - than a human teacher. And last, but not least, a computer program can teach an infinite number of students at the same time in an individual manner, if it is installed on as many machines, or accessible through as many terminals in Internet country.

Given these advantages, it makes sense to invest some effort into addressing the above mentioned “technicality disadvantage” of the human-computer interface. Here, my grammar teaching interface tries to make advances with regard to the following four principles:

(a) Flexibility

The interface is notationally flexible, i.e. the user can choose one of several notational conventions (e.g. flat dependency grammar, enriched text, meta text notation, tree structures). According to the student's background, the analysis' complexity can be modified, - for example, by increasing or decreasing the number of distinct word class tags used. At the same time, such filtering permits a choice between the use of traditional word class concepts on the one side, or , for instance, purely morphologically motivated ones, on the other side. In order to make a session more colourful, it is also possible to move between corpus text, live newspaper text, randomised test sentences and one's own creative idiolect.

(b) Interactivity

A set of CGI-controlled programs reacts instantly to those user choices advocated by the flexibility principle, and the interface changes accordingly in an interactive way, permitting, for example, to move back and forth between levels and notational conventions. When a sentence proves problematic or incomprehensible, the user can modify it, or ask for the computer’s opinion. Grammatical analysis can be run in interactive mode, where a sentence is analysed step by step, with the student suggesting form or function readings for words or constituents, and the computer checking and commenting the choices. In this mode the text in question will gradually be coloured for word class and indexed for function as the student's analysis progresses.

(c) Naturalness

A major draw back of most language teaching software (or, for that matter, language analysis software) is that they do not run on free, natural language, but on a small set of predefined sentences that cannot be modified or replaced. Usually "toy lexica" and "toy grammars" are used that can handle only a narrow range of built-in structures. In my interface the underlying lexica and grammar modules cover the whole language, and the user can thus manipulate the text to be analysed in much the same way as in an ordinary text processor.

The second aspect of naturalness concerns, as mentioned above, "untechnicality", and as much keyboard-interaction as possible has therefore been replaced by graphical and mouse based means, like menu choices and clickable radio buttons and help windows. Being internet based, the system automatically takes advantage of a browser's navigation tools, scroll bars, page memory and cut'n'paste functionality.

(d) Tutoring

Tutoring is probably that human teacher feature that is hardest to emulate. A teacher's intuitive understanding of a student's problems is inherently difficult to build into a program. A certain minimum of tutoring can be achieved by providing readily available (i.e. "clickable") definitions of grammatical terms, and examples of their usage and the phenomenon's distribution in the language. For the latter purpose, a powerful corpus searching tool has been crafted to find examples of user-defined grammatical structures in automatically (and, in theory, simultaneously) annotated corpora at the system's disposal. After acquiring some basic notational skills a student (or researcher) can search for any combination and sequence of word forms, lexemes, word classes, syntactic function and so on. Ultimately, "guided tours" could be designed for certain topics by blending the definition and corpus example tools.

Another aspect of tutoring, which appears useful in *foreign* language teaching, and has been tentatively introduced for Portuguese, is translational help, either in the form of dictionary enquiries, translational tagging[234] or even rough sentence translation.

7.2.2 The grammatical base: Why Constraint Grammar?

The grammatical backbone of both the Portuguese and the English VISL systems is the Constraint Grammar framework as it has been discussed elsewhere in this dissertation. Constraint Grammar was chosen for a number of practical reasons:

♦ Constraint Grammar is robust. A language teaching system based on natural text must be very stable, and be able to provide *some* analysis to all input. A "no parse"-message window would destroy the illusion of a real teacher, and - if frequent - ultimately result in student frustration. Since CG works by adding and removing information, the *correct* reading will crystallise in an indirect way - simply by being the last surviving analysis. Thus, in the CG formalism, even unusual or partial sentences will receive *some* analysis, and an ill-formed sentence will not prevent correct *lower* level analysis, for instance, correct word-class and noun phrase analyses.

♦ Constraint Grammar is tag based, and adds tag strings to word forms. First, string based information is easy to port and easy to manipulate in a computer, and second, this way different kinds of information, lexical, morphological and syntactic, form, function and structure, can be handled within the same formalism[235], which allows easy notational transformation. Thus, tags can be fused into more general Portmanteau-tags (downward compatibility), and split up into subcategories by using higher lever information from other tags in the same string (upward compatibility). An example for the first is the fusion of adverbial adjuncts, adverbial objects and prepositional objects into a Portmanteau-tag 'adverbial', and an example for the latter is the function-based distinction between "adjectival" (adjective-like) and "substantival" (noun-like) pronouns. With sufficiently detailed dependency markers, CG-notation can even be transformed into constituent based tree structure notation (Bick, 1997-1).

♦ CG-notation has elegant ways of underspecifying ambiguity. Postnominal PP-attachment, for example, is expressed as "nominal attachment to the left" (@N<), so

[234] Here, base form translation equivalents are given as the last tag on the tag line in verticalized CG-notation. Some polysemy resolution is performed, based on valency instantiation and the disambiguation of atomic semantic features by Constraint Grammar rules.

[235] Of course, this is true of the *internal working* of the grammar, too. Constraints can be worded in much the same way whether they are morphological, syntactic or semantic, and information from different levels can interact in disambiguation.

that the Chinese origin in *The man with the bicycle from China* can be applied to both 'man' and 'bicycle'. In cases of ambiguous functional analysis, CG can add several (competing) function tags to the same word, so that the ambiguity can be expressed in *one* analysis. Especially with long sentences this is pedagogically superior to having to scroll through several pages with tens or hundreds of possible analyses. Also, it becomes easier to judge the student's analysis - if the tag suggested is a substring of the ambiguous tag string, then the suggested reading will be accepted by the computer, even if it is not the only one. For the same reason, if the computer fails to resolve some ambiguity, this will not impair the student-computer interaction, - as long as the correct reading is among the ones "surviving" the CG-treatment (which can be geared to prefer ambiguity to errors), the robot teacher may be over-indulgent, but it will not harshly criticise a justifiable student choice.

♦ Due to the modularity of the underlying CG-based progressive level parsing system, it is possible to manage a growing system *in flux* and to choose those modules that already have achieved a sufficiently high level of correctness and coverage, and make them accessible to the student community. In the VISL system, this approach has led to the development of certain modules for one language that later could be applied to other languages without major modifications. The incrementality of the CG-parser lets the system grow like a holographic picture - the object is visible all the time, and only its granularity improves with the amount of time and work put into it. Once the user interface is in place as such (and hardware and wiring technology permits), there is a teaching and demonstration dividend, even if the parser is still to be improved. Thus, with a CG-parser, the time lapse between grammatical work and pedagogical implementability can be reduced to a minimum.

7.2.3. The pedagogical base

Word based tags (after, under, over, indexed or - as colour code - "inside" the words in question, with or without underlining, in the form of abbreviations or symbols) are pedagogically intuitive and close to "basic" grammar, - not only for marking word class, but also in syntax, as can be told from the "cross-and-circle" grammar (‘kryds og bolle’) used in Danish primary schools, or the corresponding colour-underlining system used in Germany. A special advantage of CG's dependency notation is that it mirrors children's semantically based intuition making the head of a phrase the bearer of its syntactic weight. For the sentence *"Pia's stupid rabbit ate the flowers I collected for mother"* the quick answer to the subject question (*"Who ate ...?"*) is *"The rabbit!"* and, even more surely, the answer to the object question (*"What did the rabbit eat?"*) will be *"The flowers"*. It usually takes additional syntactic curiosity from the teacher's part to elicit answers as to <u>whose</u> rabbit and <u>which</u> flowers the sentence was about. Apart from articles (that are necessary to state a noun's definiteness, something which can be achieved in Danish by <u>morphological</u> means), most other modifiers seem to be outside the reach of "subject"-ness or "object"-ness. Most strikingly so in the case of parenthetic relative clauses: *Ann, who hadn't slept for two nights, wanted to go home - Who wanted to go home? - Ann.* Here, word based dependency analysis seems to be mentally more basic than constituent analysis, which becomes secondary: It is the primary subjects 'rabbit' and 'Ann' that grow into complex constituents by absorbing modifiers like 'stupid' or 'who hadn't slept for two nights', - and not a (larger) subject constituent that breaks down into several sub-constituents. I believe that it is pedagogically important to start from the (concrete) referent centre (i.e. 'the rabbit' and 'Ann') and work from there by adding more and more bricks (each of them still as small and as concrete as possible), creating a - larger - whole that is still concrete in the child's mind, instead of starting with an abstract unit (a subject constituent) that will not be made concrete but several layers of analysis further down (i.e. at word level).

Therefore, the current internet teaching system offers the choice of a CG-derived interactive grammar module, where this thought is matched both by notation and procedural sequentiality: Functions are tagged at a phrase's head word, and it is possible to correctly click and identify, say, a subject head as "subject", even *before* possible modifiers have been attached by additional clicks and menu-choices. This contrasts strongly with a traditional constituent based approach, where there can be no subject without a subject *constituent.*

Still, while advocating a head-driven and bottom-up analysis for pedagogical and psycholinguistic reasons, the flexibility principle is applied to this matter, too, and students do have the choice of a tree-structured constituent analysis (which, for Portuguese, is automatically derived from the flat dependency notation), thus facilitating a top-down perspective where desired[236].

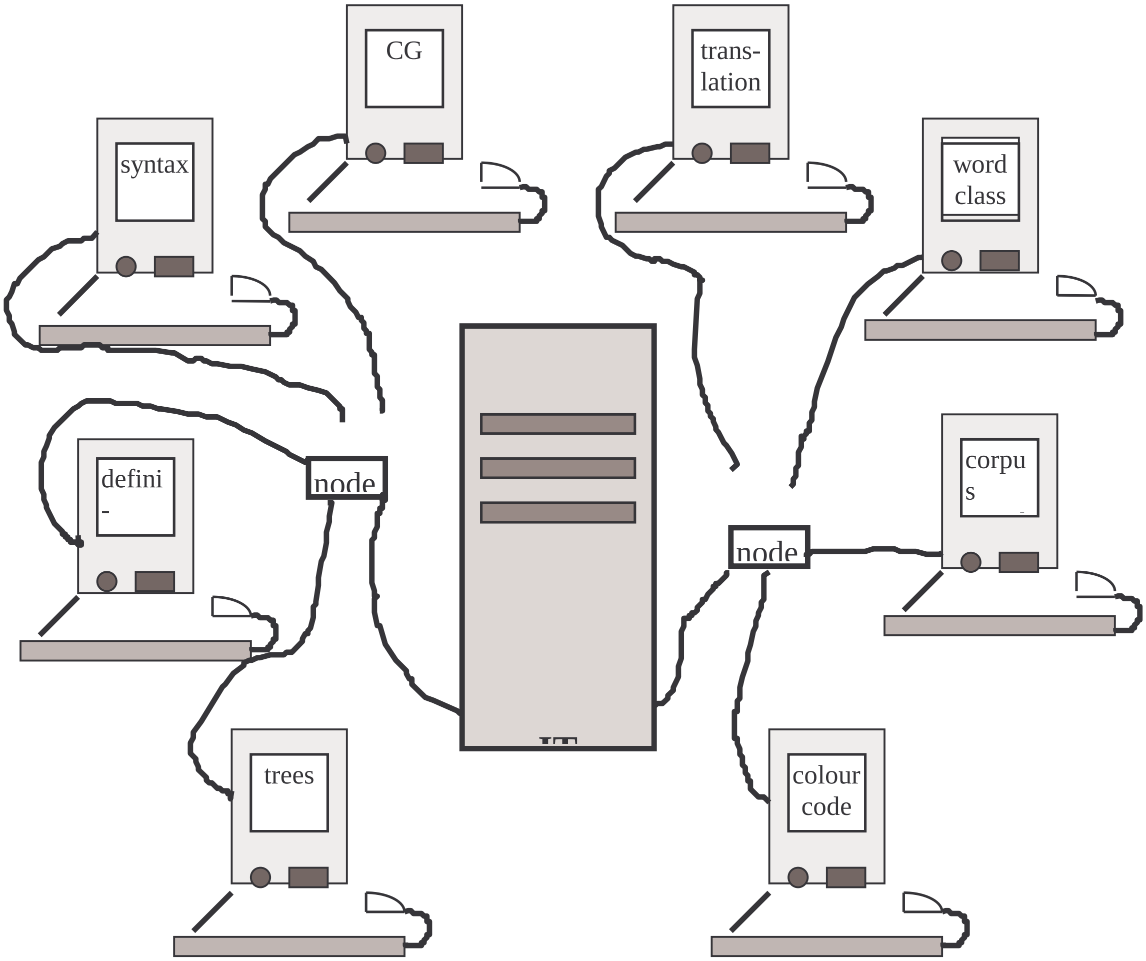
7.2.4. The interface

The system is implemented as a free-for-all distributed teaching environment, with one or more servers running the grammar software and the CGI-programs necessary to interact through the internet with users at their school, university or home computers (1). A central IT-grammar server handles - in parallel - a large number of student terminals that may focus on different languages, different levels of analysis or different training tasks, representing different notational or grammatical systems.

[236] The choice of options integrated into the *interactive* , i.e. student driven, analysis is not the same in the tree notation scheme as in the flat dependency grammar notation, partly for pedagogical reasons - because the two notations focus on different structural and functional aspects of sentence analysis - partly for technical reasons, since the former is java-based, while the latter is entirely based on iteratively cgi-generated standard html pages. Both, however, depend of the flexibility of the underlying CG-analysis, and one of the spin-offs of the VISL-project at OU will - hopefully - be a more flexible CG-compiler allowing the integration of notationally conditioned changes (like for instance constituent boundaries) into the rule system of the disambiguation grammar itself, as well as the mapping of add-on corrections and tutoring comments.

A first version of a tree-drawing java program was written by Thomas Larsen and later modified and extended by Martin Carlsen for the VISL-project at Odense University.

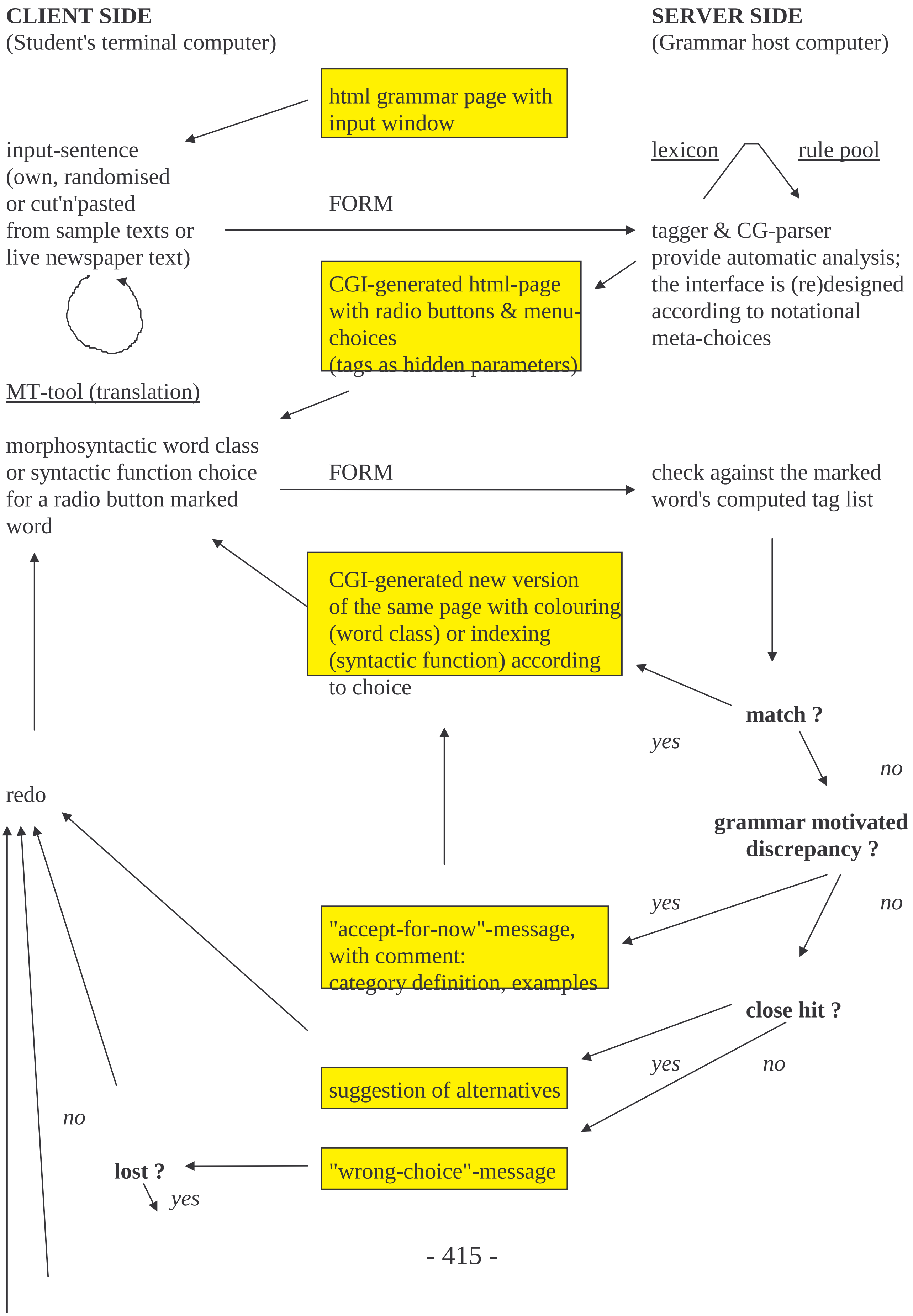
(1) **Distributed grammar teaching environment**:

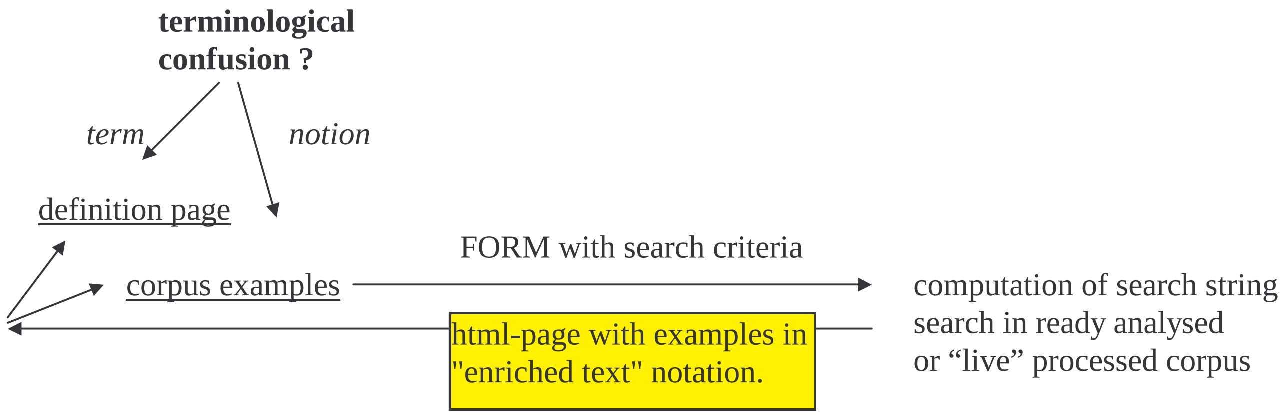


All computation is done *at server side*, and user input and choices are managed through server-updated html-forms, either directly or by providing parameters for a java-applet. This method has a number of advantages over traditional *user side* based software: First, no software has to be sold or distributed and, consequently, copy-right problems are minimized; second, the age and quality of the user's computer is of less importance (as long as it can run a java-enabled browser), and - not entirely unimportant for multi-language applications - incompatibilities with regard to software, character set, machine type etc. are circumvented; third, interaction is speedy, since only short html-texts are sent back and forth, while programs proper are run by those machines that are good at it - heavy computation intensive grammar programs by the server, light keyboard, mouse and text manipulation by the terminal machine whose language and other preferences remain customised by the user.

The flow chart diagram (2) illustrates the interaction between student and grammar server in a sequential way, pointing out where and how information is provided and what choices can be made by the student in order to navigate through the teaching system.

(2) **flow chart of student - server interaction**:


CLIENT SIDE
(Student's terminal computer)
SERVER SIDE
(Grammar host computer)
html grammar page with input window
input-sentence (own, randomised or cut'n'pasted from sample texts or live newspaper text)
lexicon
rule pool
FORM
tagger & CG-parser provide automatic analysis; the interface is (re)designed according to notational meta-choices
CGI-generated html-page with radio buttons & menu-choices (tags as hidden parameters)
MT-tool (translation)
morphosyntactic word class or syntactic function choice for a radio button marked word
FORM
check against the marked word's computed tag list
CGI-generated new version of the same page with colouring (word class) or indexing (syntactic function) according to choice
match ?
yes
no
redo
grammar motivated discrepancy ?
yes
no
"accept-for-now"-message, with comment: category definition, examples
close hit ?
yes
no
suggestion of alternatives
no
"wrong-choice"-message
lost ?
yes



terminological confusion ?

term

notion

definition page

corpus examples

FORM with search criteria

computation of search string
search in ready analysed
or “live” processed corpus

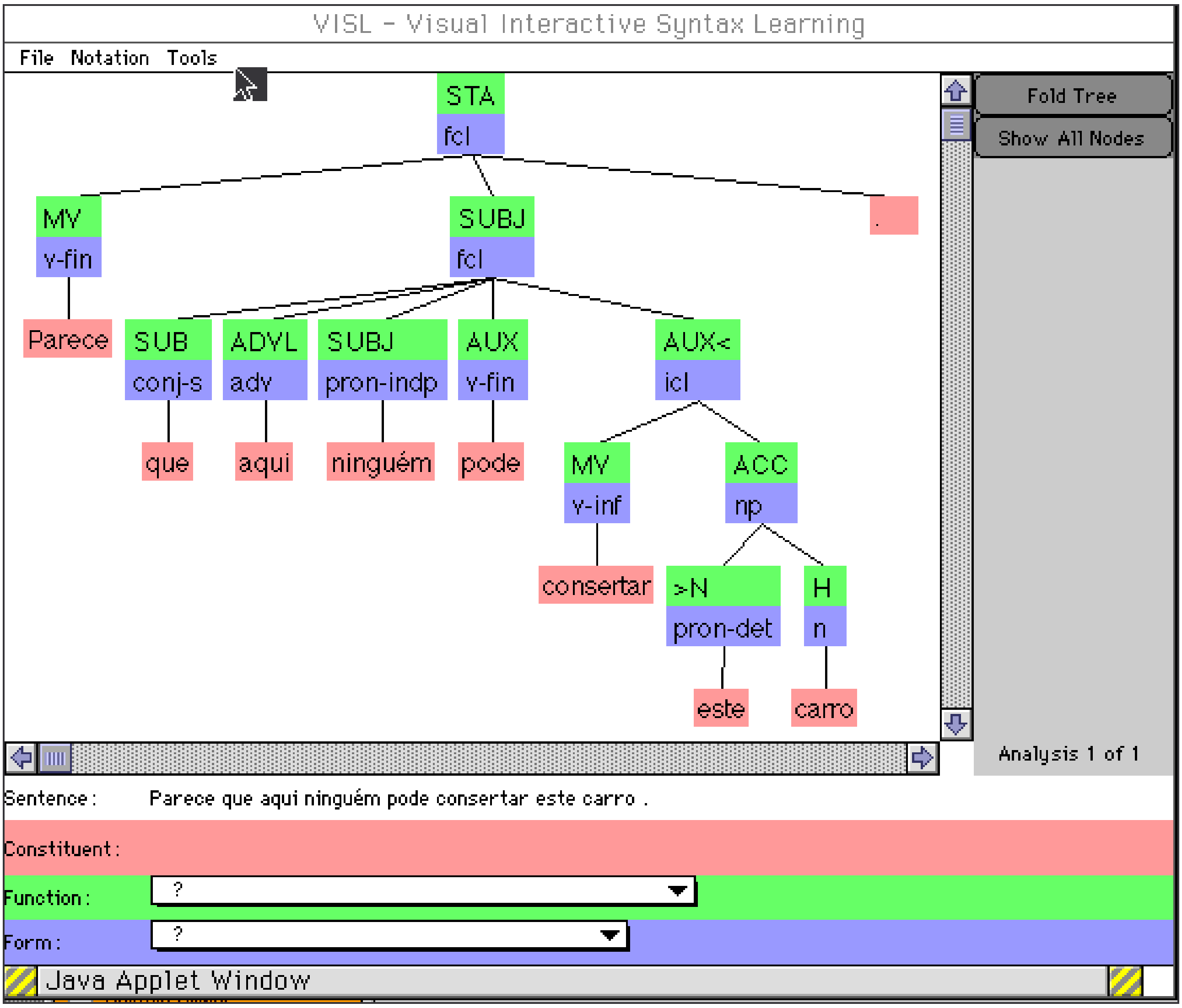
html-page with examples in "enriched text" notation.

7.2.5. Syntactic tree structures

While the "enriched text" notation (cp. illustration (3) in chapter 7.1.) is ideal for combining the natural flow of running text with word class and function information in a graphical way, it does not emphasise constituent structure. Rather, the latter is expressed indirectly, and in a flat way, by mounting heads and dependents into constituents with the help of directed "dependency markers". Adnominals, for instance, are mounted on “N-words” (i.e., typically, nouns): @>N points to an np-head to the right, thus signalling a prenominal modifier, while @N< stands for a postnominal modifier (attaching left). Still, the dependency grammar embodied by the system's Portuguese Constraint Grammar rules is detailed and precise enough to permit automatic transformation into PSG-like tree structures (chapter 4.6.3). This is achieved by running a perl-compiled set of substitution rules on top of a detailed CG analysis[237], mapping constituent borders and deriving complex constituent function from function tags at the dependency heads in question.

(3) Automatic transformation into syntactic tree structures

[237] In the English VISL-system, I have written an ordinary generative grammar running on top of a CG-system enhanced with rules for subclause form and function. The - declarative - generative system is better at capturing tree structure ambiguity, but faces more serious time & space problems than the - procedural - substitution rule program.



Sentence: Parece que aqui ninguém pode consertar este carro .

Constituent:

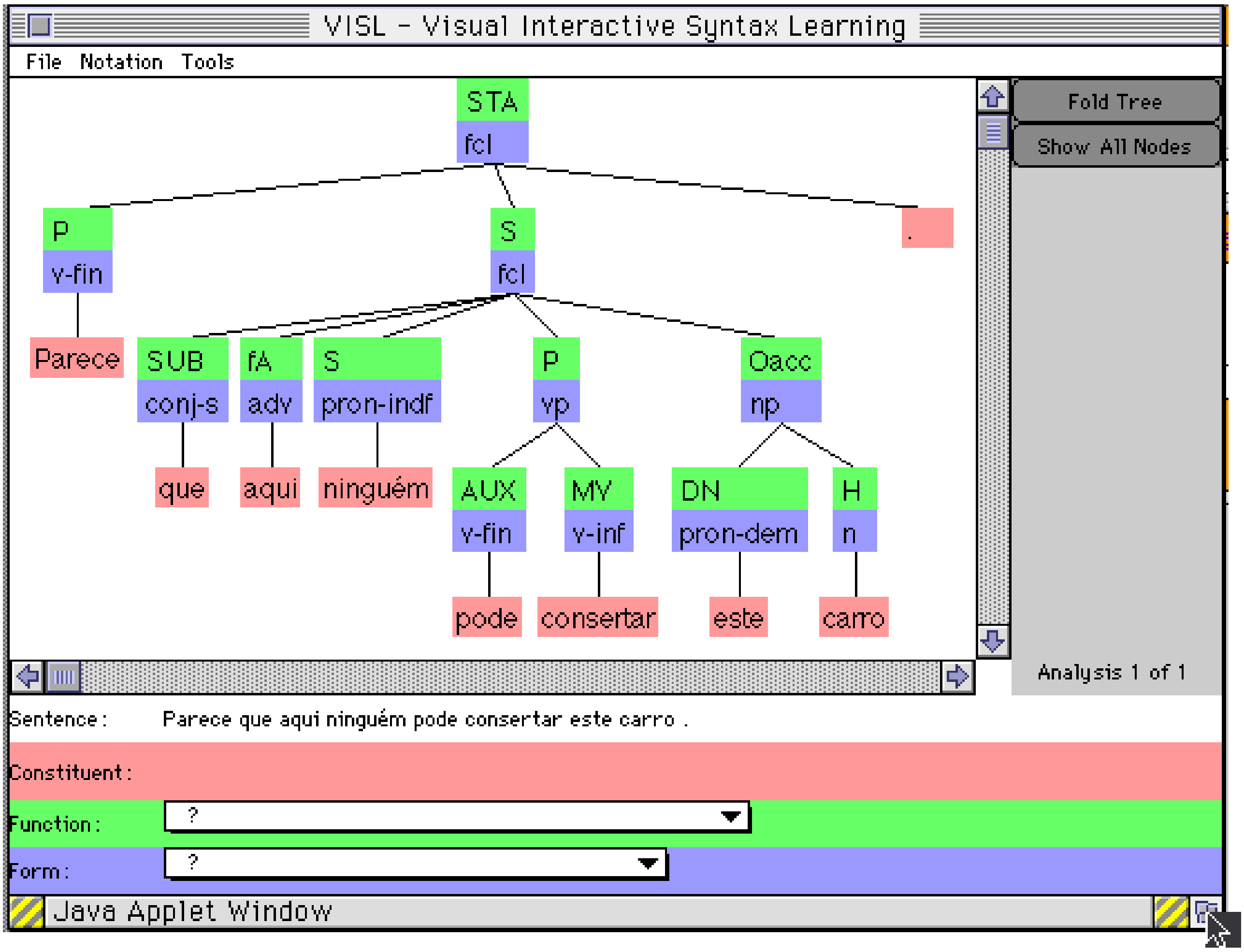
Function: ?

Form: ?

The tree structure of the sample sentence is 5 levels deep, and the symbols used are still close to the original CG notation. The tree provides one *function-form* pair for each constituent, also where the constituent is a node in the tree, and not an individual word.

In the alternative *VISL-notation*, an additional filter program adapts the "raw" CG output to a more traditional set of categories. Thus, traditional pronoun classes are introduced instead of the parser's three morphologically defined pronoun classes. Also, the category of *complement of auxiliary* (AUX<) is abandoned in favour of a more traditional *group* treatment of verbal constituents, with predicator function:

(4) Syntactic tree structures (VISL notation)



Sentence: Parece que aqui ninguém pode consertar este carro .

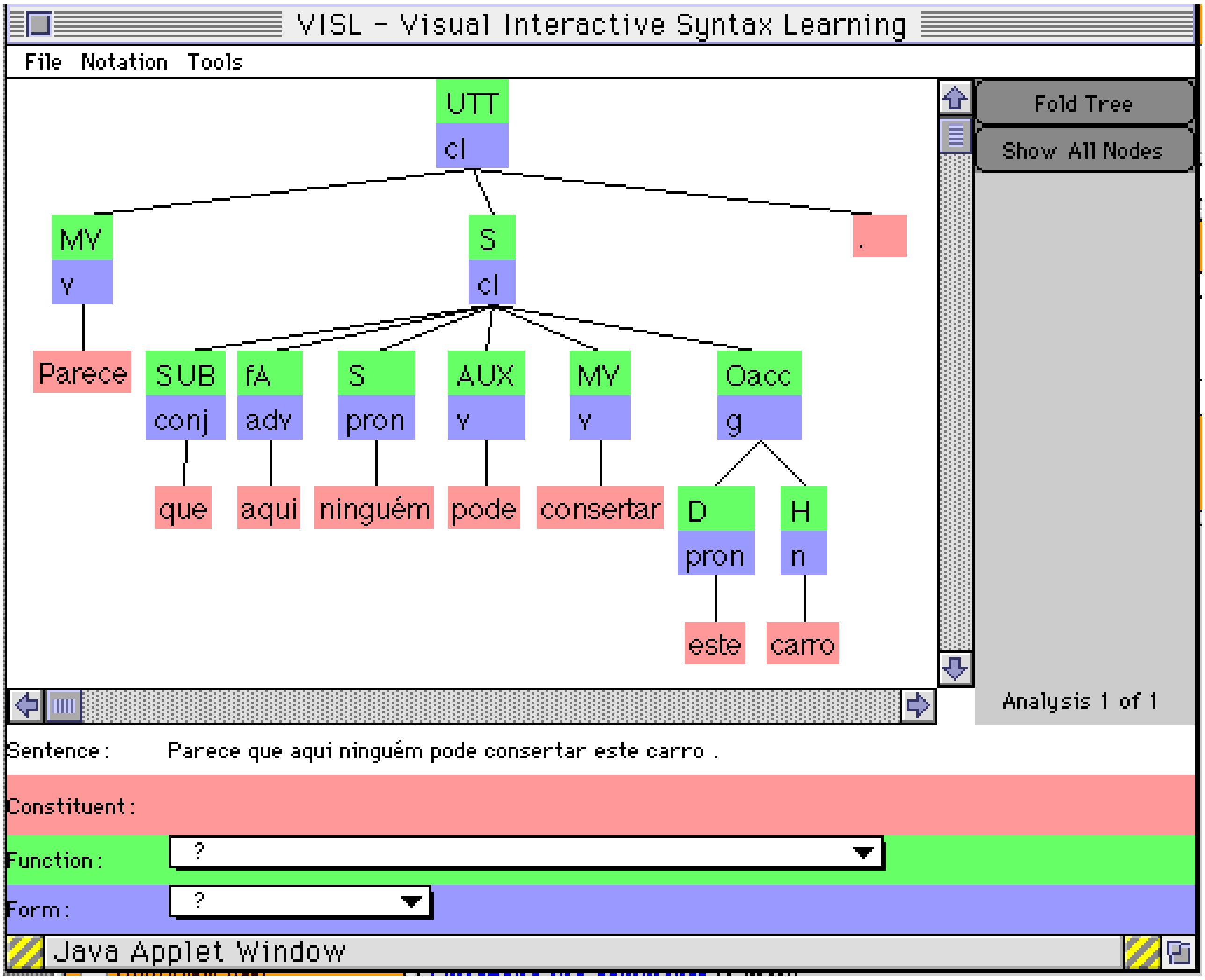
Constituent:

Function: ?

Form: ?

Finally, students can opt for a simplified analysis, where no subdistinctions are made in the form classes of pronoun, verb, group and clause, and where in-group functions are reduced to heads (H) and dependents (D). Verb chains are treated in a flat way, with auxiliaries and main verbs both functioning as clause level constituents:

(5) Syntactic tree structures (VISL simplified notation)



Sentence : Parece que aqui ninguém pode consertar este carro .

Constituent :

Function : ?

Form : ?

For the sake of pedagogical continuity, and in order to facilitate integration of internet tree-analyses into pre-existing grammar courses, corpora of text book sentences can be tagged manually or semi-automatically according to the needs of individual teachers or teaching institutions, creating “closed corpora” for easy (and error-free) reference[238]. Likewise, individual notational tag filters can be crafted “made to order”.

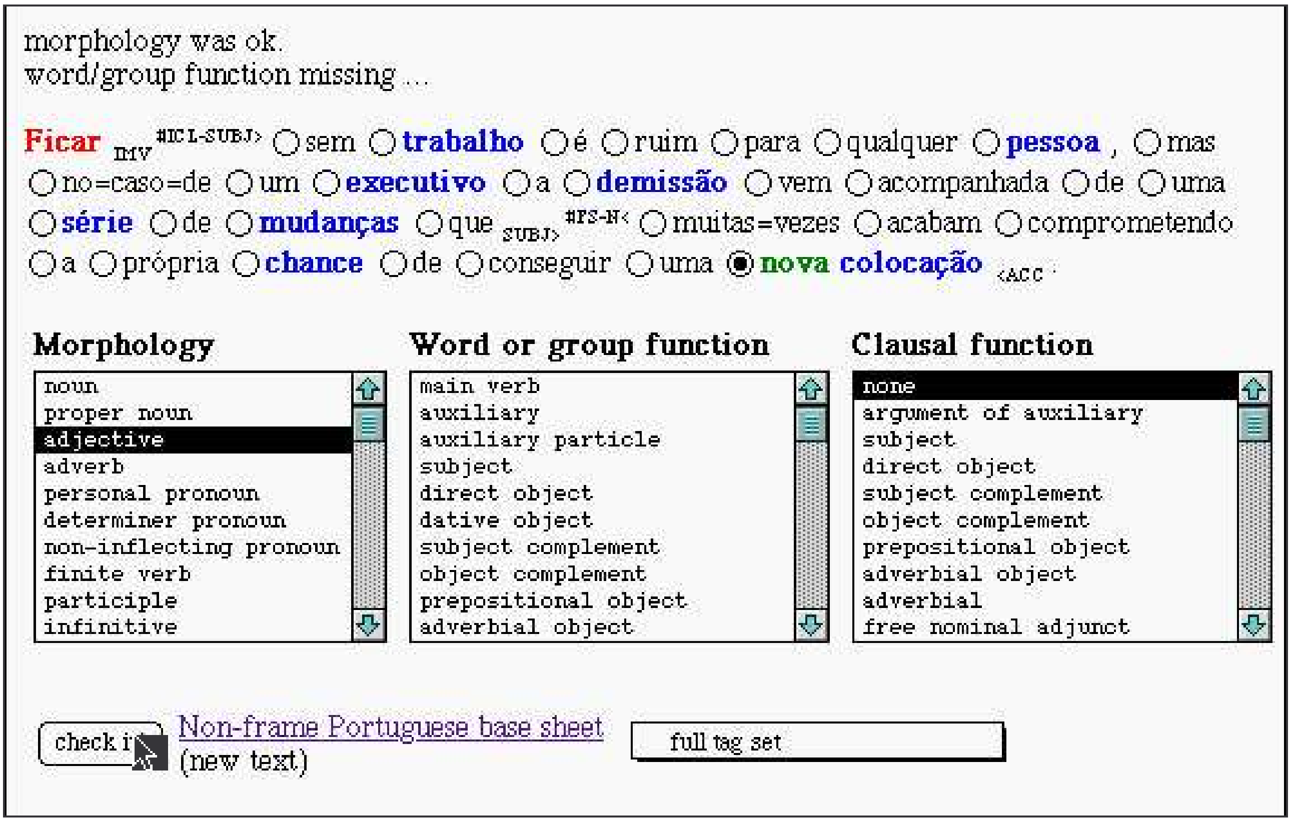
7.2.6. Student driven interactive analysis

In the interactive analysis mode, a full analysis is computed by the server, but the Constraint Grammar tags remain concealed as hidden parameters in the html-forms

[238] The “closed corpus” approach also allows integrating into the VISL system languages that have not yet Constraint Grammar systems at their disposal for live analysis (Danish, French, Italian), or lack the necessary syntactic CG module (German). The English VISL system is the one most closely integrated into an existing teaching program. Thus, the English VISL group at OU, supervised by John Dienhart, has hand-tagged a text book corpus containing all exercises from (Bache et. al., 1993), using the book’s tag sets for form and function, and a tree coding scheme compatible with the existing tree-designing module of the Portuguese system.

sent back and forth through the CGI-channel. Text is presented as running word forms with "clickable" radio buttons, and tag options are presented as menu choices:

(6) Text with radio buttons and tag-menus for progressive interactive analysis



morphology was ok.
word/group function missing ...

Ficar $_{\text{IMV}}^{\text{\#ICL-SUBJ>}}$ ○ sem ○ trabalho ○ é ○ ruim ○ para ○ qualquer ○ pessoa , ○ mas ○ no=caso=de ○ um ○ executivo ○ a ○ demissão ○ vem ○ acompanhada ○ de ○ uma ○ série ○ de ○ mudanças ○ que $_{\text{SUBJ>}}^{\text{\#FS-N<}}$ ○ muitas=vezes ○ acabam ○ comprometendo ○ a ○ própria ○ chance ○ de ○ conseguir ○ uma ● nova colocação $_{\text{<ACC}}$.

The first menu is primarily about word class, but makes - in addition - a morphological distinction between 3 types of non-inflecting verb forms (infinitives, participles and gerunds). The second menu selects word or group function, with the latter to be marked on the group's head word. The last menu, finally, allows to add subclause function, which is assigned to main verbs in non-finite subclauses, and to complementisers (conjunctions, relatives, interrogatives) in finite or verb-less (averbal) subclauses[239].

In the example, the student has just chosen to analyse the second last word of the sentence, ‘nova’, as an adjective by selecting the appropriate tag from the first menu. ‘Adjective’ being the right choice, ‘nova’ has been coloured in “adjective colour”, green. The student can now add a functional tag, or progress to another word. One of the most simple exercises, which can be carried out even by primary school children, would be to identify, say, *all* the nouns, with correct choices being “rewarded” by progressive colouring of the sentence, as seen here. Note, that the last noun's radio button has disappeared, since it also has been tagged for function (here, @ACC, direct object). For the leading verb (in the infinitive), a full analysis means

[239] Since complementizers are obligatory for these clause types in Portuguese, but not in English, this convention has been changed in the English VISL module, and subclause function is here always tagged on the clause's first verb, whether finite or not.

two function tags, since the student here also needs to identify the subject function of the infinitive-clause as a whole (shown as super-script). This way, the general appearance of the sentence will gradually change into the coloured "enriched text" notation (chapter 7.1., illustration 3), to which it becomes identical after full correct analysis.

One of the tags already assigned by the student is a (postnominal) subclause function label for the relative pronoun "que", but the word still lacks a form label. Since my grammar - somewhat unorthodoxically - defines 3 pronoun classes in a purely morphological manner by lexeme and word form categories, he may have found it difficult to decide on one pronoun subclass rather than another. He can now procede in one of three ways:

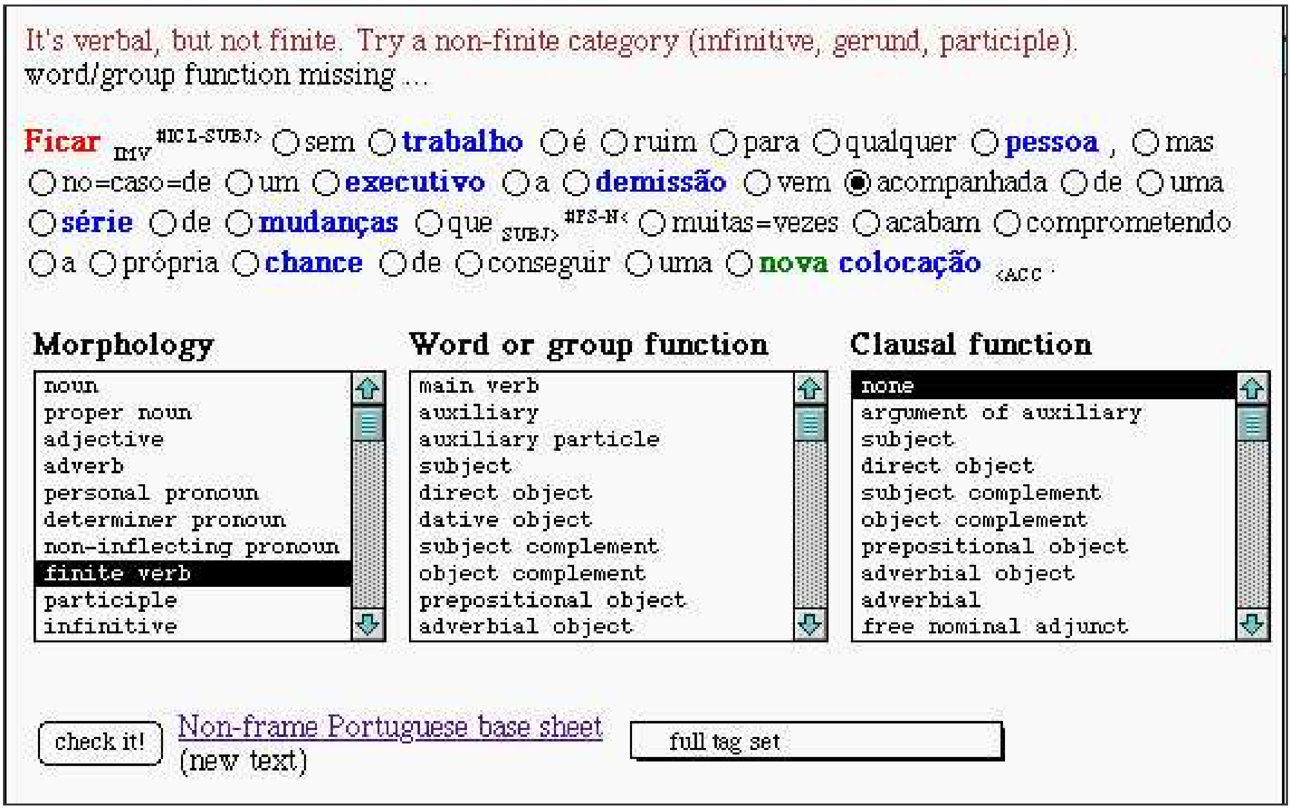
(a) He can choose (maybe at random) a pronoun subclass and wait for comments. In the case of a wrong choice, the system will act teacher, accepting his choice for being *within* the pronoun class (knoledge which should be duly honoured), while at the same time explaining why the system prefers another subcategory, and how this subcategory is defined. Here, the pedagogical strategy is to distinguish between "absolute errors" and errors originating from the clash of two conceptually different schools of grammar. In the latter case, - be the teacher human or not -, the student's view should be accepted for what it is, and the difference be explained.

(b) He can scroll to the bottom of the menu window and select the last item, "Show me!". The system will then show the correct analysis and colour/index the word in question accordingly. Especially for the second menu, the word or group function menu, it proved unavoidable to introduce this choice, due to the highly differentiated tag set used - and, of course, so as not to frustrate the student unnecessarily. Also, since "live" text is being used, there *is* a chance - though a tolerably small one - for the system being plain wrong, and the student's analysis being right.

(c) Finally, there is the possibility of switching to a smaller, more traditional tag set by means of a special meta-menu among the navigation buttons underneath (now showing "full tag set" mode). The tag-menus will then be simplified, and there will be only one pronoun class, with articles forming a new, independent class. Similarly, for function, *"adverbial object"* (@ADV, i.e. valency bound adverbial) and *"adverbial adjunct"* (@ADVL), or even *"prepositional object"* (@PIV), will be fused into the Portmanteau tag *"adverbial"* (@ADVL).

If an error is made, even if it does not originate from a different view on the categories of grammar, it may still be a "soft" error, where the student is fairly close to the correct answer. In (7), the participle "acompanhada" has been assigned a gerund reading by the student. Since both the participle and the gerund categories are clearly verbal (and even, both non-finite), the system does not simply reject the answer as "plain wrong", but accepts the "verbality" as correct and encourages further subdivision:

(7) Tutoring in the case of a "close miss"



It's verbal, but not finite. Try a non-finite category (infinitive, gerund, participle).
word/group function missing ...

Ficar $_{\text{IMV}}$ $^{\text{\#ICL-SUBJ>}}$ ○ sem ○ trabalho ○ é ○ ruim ○ para ○ qualquer ○ pessoa , ○ mas ○ no=caso=de ○ um ○ executivo ○ a ○ demissão ○ vem ● acompanhada ○ de ○ uma ○ série ○ de ○ mudanças ○ que $_{\text{SUBJ>}}$ $^{\text{\#FS-N<}}$ ○ muitas=vezes ○ acabam ○ comprometendo ○ a ○ própria ○ chance ○ de ○ conseguir ○ uma ○ nova colocação $_{\text{<ACC}}$.

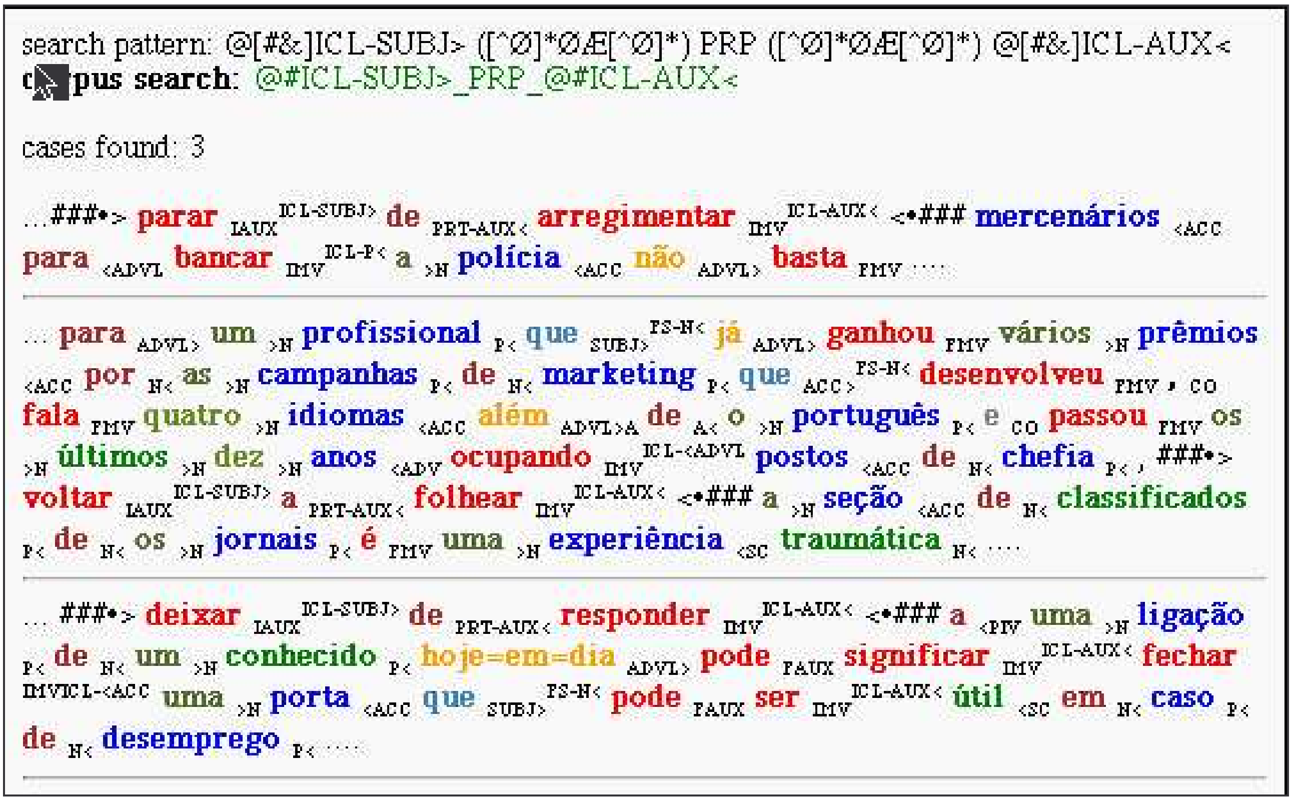
Though the possibility to work with free, real life text and to make up one's own examples is compelling proof of the efficiency of the system's underlying parser (or at least proves that such efficiency is courageously being claimed ...), not all students master a foreign language to such a degree that they enjoy inventing their own sentences, and they will not always come up with a *correct*. sentence, either. And even copying and pasting from corpus texts (the obvious solution, implemented in both the Portuguese and English modules) may become tedious in the end. On the other hand, many people enjoy a test match - at least as long as they are not being watched or judged. Therefore, I have integrated a sentence randomiser into the system, that offers corpus examples of its own[240] if the input window is left empty in interactive grammar mode. In order not to hit upon headlines and other unorthodox or "incomplete" text material, all random text choices are cut at sentence delimiters (full stop, colon etc.), and filtered out if they do not contain at least one finite verb.

[240] Presently, the text base for the randomizer is about 1,000 sentences large, but since it is based on automatic analysis, ten or a hundred times that number would not be a problem, either.

7.3 Corpus research

In the case of persisting difficulties with a particular grammatical topic, or for want of a satisfactory definition, a student may want to look at a few examples of how the feature in question is used in different sentences, something one would expect to achieve in traditional, text book based, exercises by referring back to a specific chapter in the grammar book. In the case of an IT interface with a live parser at its disposal, there is - in principle - no limit to the amount of corpus text to be searched for "typical" examples, and the illusion of a concise "chapter" can be created even with a chaotic "book" (corpus) with thousands of pages: While the grammar server searches the "book", the terminal will show the "chapter". Let's assume, for instance, that the student has a problem with Portuguese verb chains - he is in doubt just how and if prepositions can be integrated in auxiliary verb structures. He therefore clicks "open corpus search" in the task frame, and looks for prepositions preceded by auxiliaries (@FAUX or @IAUX) or followed by post-auxiliaries (@#ICL-AUX<). The system will then output a very long "chapter" on this topic, and he may get the impression, that, say, such verb chains do not occur in infinitive-subclauses. To look for counter-examples, all he has to do is add a form/function label for such subclauses. In (12), the search is for non-finite *subject* subclauses: @#ICL-SUBJ>_PRP_@#ICL-AUX<:

(1) searching for corpus examples



search pattern: @[#&]ICL-SUBJ> ([^Ø]*ØÆ[^Ø]*) PRP ([^Ø]*ØÆ[^Ø]*) @[#&]ICL-AUX<
corpus search: @#ICL-SUBJ>_PRP_@#ICL-AUX<

cases found: 3

...###•> parar IAUX ICL-SUBJ> de PRT-AUX< arregimentar IMV ICL-AUX< <•### mercenários <ACC para <ADVL bancar IMV ICL-P< a >N polícia <ACC não ADVL> basta FMV ....

... para ADVL> um >N profissional P< que SUBJ> FS-N< já ADVL> ganhou FMV vários >N prêmios <ACC por N< as >N campanhas P< de N< marketing P< que ACC> FS-N< desenvolveu FMV , CO fala FMV quatro >N idiomas <ACC além ADVL>A de A< o >N português P< e CO passou FMV os >N últimos >N dez >N anos <ADV ocupando IMV ICL-<ADVL postos <ACC de N< chefia P< , ###•> voltar IAUX ICL-SUBJ> a PRT-AUX< folhear IMV ICL-AUX< <•### a >N seção <ACC de N< classificados P< de N< os >N jornais P< é FMV uma >N experiência <SC traumática N< ....

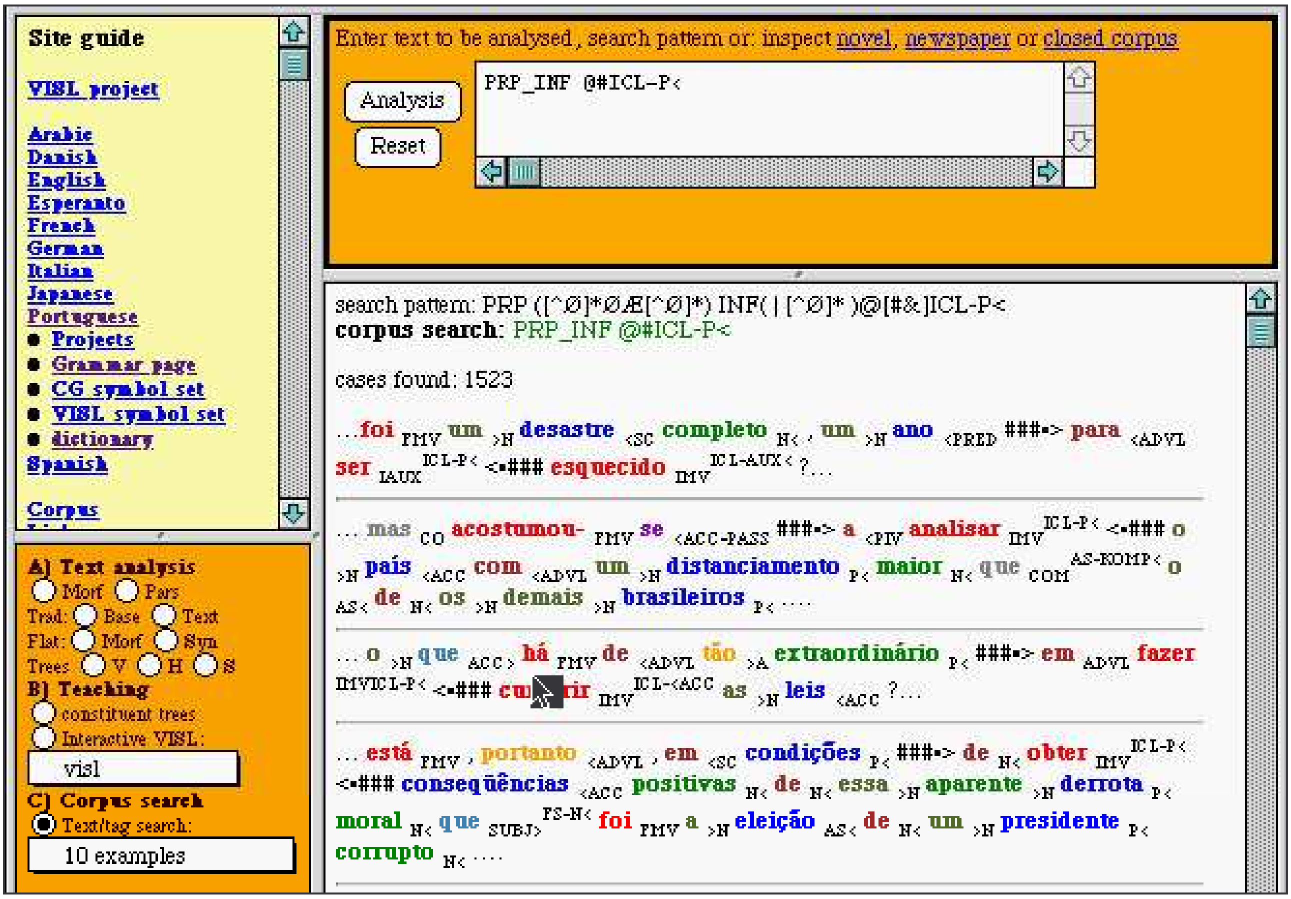
... ###•> deixar IAUX ICL-SUBJ> de PRT-AUX< responder IMV ICL-AUX< <•### a <PIV uma >N ligação P< de N< um >N conhecido P< hoje=em=dia ADVL> pode FAUX significar IMV ICL-AUX< fechar IMVICL-<ACC uma >N porta <ACC que SUBJ> FS-N< pode FAUX ser IMV ICL-AUX< útil <SC em N< caso P< de N< desemprego P< ....

Here, three results of this quite specific search task are given. Note that, again, the "enriched text" notation permits text coherence and facilitates context understanding. The particular structure looked for is marked by fat arrows, but thanks to the concise notation, the whole sentence context can be shown together with most of the tags.

Even tags *not* shown, like the inflexion category 'plural', the base form 'amigo' or the valency feature 'monotransitive', can be searched for: Virtually any combination of word forms, base forms, inflexion tags and syntactic function can be searched for each individual word in any combination of words as well as one ore more obligatory or optional dummy words. Obviously, the *real* search pattern for complex searches is much longer than the chain of tags entered by the user. The system automatically "translates" the search parameters into a regular expression string (cp., for instance, the search pattern line at the top of [1]) to be used by fast, specialised search algorithms running on the UNIX based grammar server.

Let's look at another example, where a student wants to write an essay on another aspect of infinitive function, - infinitive arguments after prepositions. A corresponding search (PRP_INF @#ICL-P<) will indeed yield a lexicographically interesting list of infinitive-governing prepositions.

(2a) Preposition-infinitive sequences in Portuguese

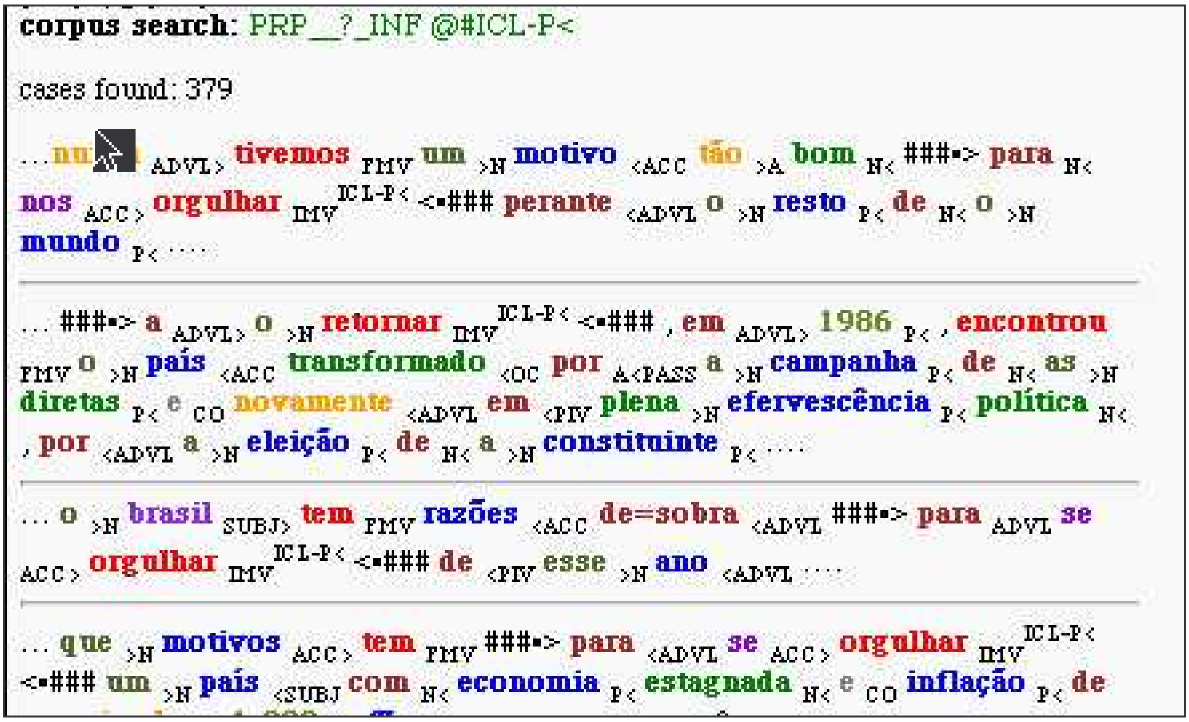


```
PRP_INF @#ICL-P<
```

```
search pattern: PRP ([^Ø]*ØÆ[^Ø]*) INF( | [^Ø]* )@[#&]ICL-P<
corpus search: PRP_INF @#ICL-P<

cases found: 1523

...foi FMV um >N desastre <SC completo N< , um >N ano <PRED ###•> para <ADVL
ser IAUX ICL-P< <•### esquecido IMV ICL-AUX< ?...

... mas CO acostumou- FMV se <ACC-PASS ###•> a <PIV analisar IMV ICL-P< <•### o
>N país <ACC com <ADVL um >N distanciamento P< maior N< que COM AS-KOMP< o
AS< de N< os >N demais >N brasileiros P< ....

... o >N que ACC> há FMV de <ADVL tão >A extraordinário P< ###•> em ADVL fazer
IMVICL-P< <•### cumprir IMV ICL-<ACC as >N leis <ACC ?...

... está FMV , portanto <ADVL , em <SC condições P< ###•> de N< obter IMV ICL-P<
<•### conseqüências <ACC positivas N< de N< essa >N aparente >N derrota P<
moral N< que SUBJ> FS-N< foi FMV a >N eleição AS< de N< um >N presidente P<
corrupto N< ....
```

If the student then wants to generalise his structural assumption he may try to admit interfering material between the preposition and the infinitive (PRP__?_INF):

(2b) Interfering material in preposition-infinitive sequences



corpus search: PRP__?_INF @#ICL-P<

cases found: 379

... nu ADVL> tivemos FMV um >N motivo <ACC tão >A bom N< ###•> para N< nos ACC> orgulhar IMV ICL-P< <•### perante <ADVL o >N resto P< de N< o >N mundo P< ....

... ###•> a ADVL> o >N retornar IMV ICL-P< <•### , em ADVL> 1986 P< , encontrou FMV o >N país <ACC transformado <OC por A<PASS a >N campanha P< de N< as >N diretas P< e CO novamente <ADVL em <PIV plena >N efervescência P< política N< , por <ADVL a >N eleição P< de N< a >N constituinte P< ....

... o >N brasil SUBJ> tem FMV razões <ACC de=sobra <ADVL ###•> para ADVL se ACC> orgulhar IMV ICL-P< <•### de <PIV esse >N ano <ADVL ....

... que >N motivos ACC> tem FMV ###•> para <ADVL se ACC> orgulhar IMV ICL-P< <•### um >N país <SUBJ com N< economia P< estagnada N< e CO inflação P< de

As might be expected, the most common interfering material are direct object pronouns, especially the reflexive “se”, but to his possible surprise, the student will now encounter the quite special Portuguese construction of a “nominal” infinitive with an article (second example), and - quantification making for good research instinct - he may continue with individual word class and function searches:

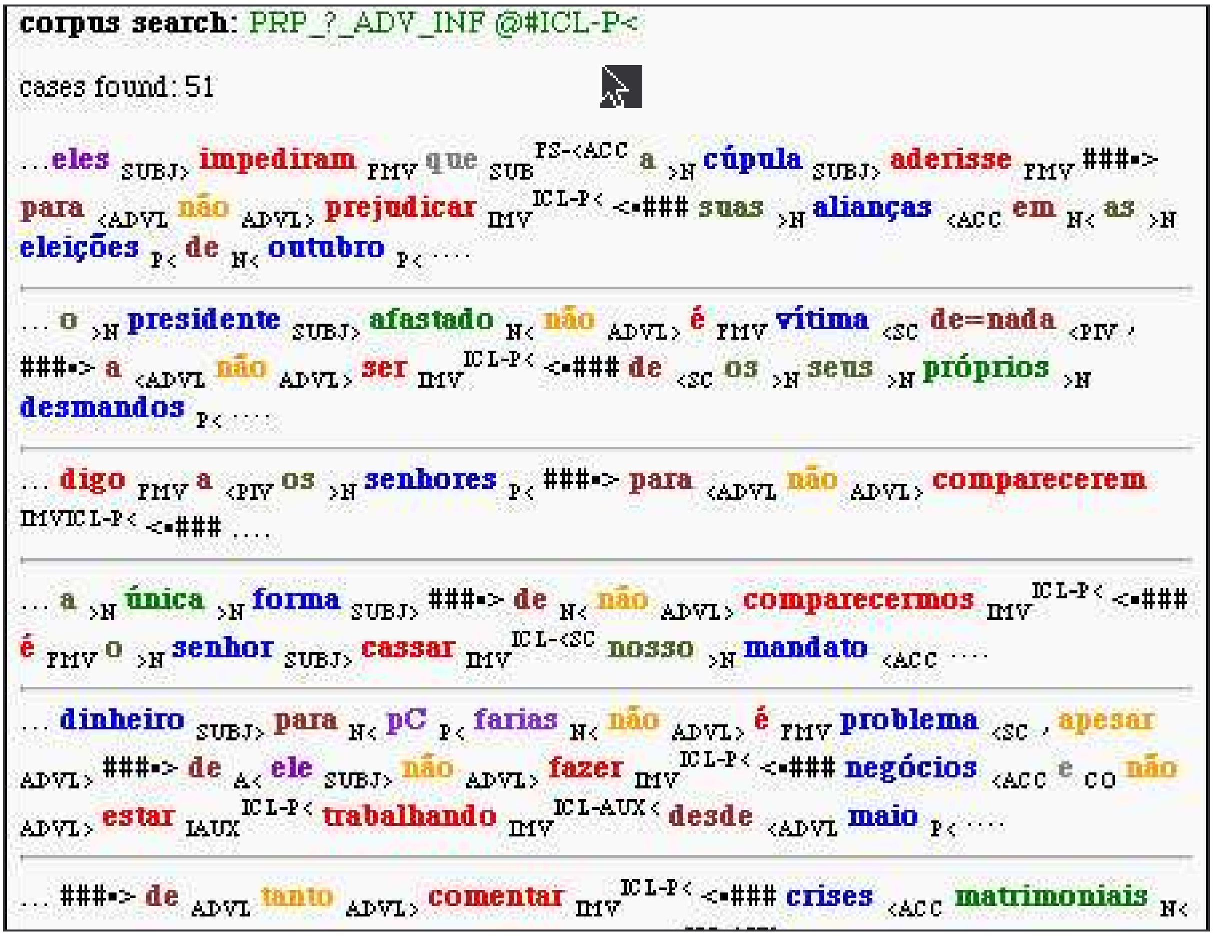
PRP_?_**PERS @ACC>**_?_INF @#ICL-P<	75
PRP_?_**PERS @SUBJ>**_?_INF @#ICL-P<	36
PRP_(@>N)?_**N @SUBJ>**_(@N<)?_INF @#ICL-P<	18
PRP_?_**ADV**_INF @#ICL-P<	51
PRP_**<art> DET @>N**_INF	80

A look at the 80 preposition-article-infinitive-sequences shows, that most are of the type “ao +INF” (a special construction translating as subclauses of the type ‘when VFIN’). Precise checks show that this case alone accounts for 76 of the 80 examples, and that articles before infinitives are all but nonexistent without a preposition, since removing ‘PRP’ from the search string only raises the number of hits by one, to 81:

“a” PRP_**<art> DET @>N**_INF	76
<art> DET @>N_INF	81

An inspection of the 51 cases with interfering adverbs suggests a closed list: “não” is by far the most common, with 1 example each of “tanto”, “melhor” and “também”, and 2 instances of “jamais”. Only in one case is there both a pronoun and an adverb.

(2c) Interfering adverbs in preposition-infinitive sequences



```
corpus search: PRP_?_ADV_INF @#ICL-P<

cases found: 51

...eles SUBJ> impediram FMV que SUB FS-<ACC a >N cúpula SUBJ> aderisse FMV ###▪>
para <ADVL não ADVL> prejudicar IMV ICL-P< <▪### suas >N alianças <ACC em N< as >N
eleições P< de N< outubro P< ....

... o >N presidente SUBJ> afastado N< não ADVL> é FMV vítima <SC de=nada <PIV ,
###▪> a <ADVL não ADVL> ser IMV ICL-P< <▪### de <SC os >N seus >N próprios >N
desmandos P< ....

... digo FMV a <PIV os >N senhores P< ###▪> para <ADVL não ADVL> comparecerem
IMVICL-P< <▪### ....

... a >N única >N forma SUBJ> ###▪> de N< não ADVL> comparecermos IMV ICL-P< <▪###
é FMV o >N senhor SUBJ> cassar IMV ICL-<SC nosso >N mandato <ACC ....

... dinheiro SUBJ> para N< pC P< farias N< não ADVL> é FMV problema <SC , apesar
ADVL> ###▪> de A< ele SUBJ> não ADVL> fazer IMV ICL-P< <▪### negócios <ACC e CO não
ADVL> estar IAUX ICL-P< trabalhando IMV ICL-AUX< desde <ADVL maio P< ....

... ###▪> de ADVL tanto ADVL> comentar IMV ICL-P< <▪### crises <ACC matrimoniais N<
```

The teaching domain is only one, quite specific, area where corpus data are of interest. In the case of research corpora, factors like size, coverage, diversity and annotation correctness are usually much more important than colourful interfaces.

So far, the morphological and syntactic modules of the parser have been used in the following corpus annotation tasks and tests (for a quantitative performance evaluation, cp. chapters 3.9 and 8.1):

The **ECI-corpus** (excerpt from the **Borba-Ramsey** corpus published on cd-rom by the European Corpus Initiative)

ca. 670.000 words used for internal research in the development of the parser

mixed genre Brazilian Portuguese texts (science, fiction, plays, conversation etc.)

This corpus has been re-tagged with the latest version of the parser, in collaboration with Diana Santos at SINTEF (Oslo), and will be made available at www.oslo.sintef.no/portug/.

VEJA articles (1996 editions, kindly provided by the editor)

ca. 600.000 words, used for internal research and teaching examples

Brazilian Portuguese news magazine texts (mixed topics)

The **NURC** speech corpus ("Norma urbana culta") [241], described in (Castilho, 1989)
ca. 100.000 words, for testing purposes only (Bick, 1998-2)
Brazilian transcribed interviews, monologue and conversation

Folha de São Paulo (1994-1996 running editions)
ca. 90.000.000 words, for a research project[242] at the University of São Paulo
Brazilian newspaper texts (all topics)

The **Tycho Brahe** corpus (17th century sample), cp. www.ime.usp.br/~tycho
ca. 50.000 words, for external use
historical Portuguese (Antonio das Chagas)[243]

To make automatic comparison possible, the system's morphological tag set was filtered into specific synthetic tags also recognized by the probabilistic tagger used in the Tycho Brahe project.

The **NILC** corpus (Núcleo Interinstitucinal de Lingüística Computacional, http://www.nilc.icmc.sc.usp.br/)[244]
ca. 39.000.000 words, used for testing purposes
ca. 100.000 words for external evaluation
journalistic, didactic and student essay texts

Originally, I tagged this corpus for internal purposes only, as a means of testing the robustness of the morphological part of the CG parser. However, part of the corpus (100.000 words of mixed science, literature and economy) also exists in a hand-tagged version established by NILC in order to train a probabilistic or hybrid tagging system. Like in the Tycho Brahe case, the CG morphological tag set proved rich enough to allow filtering into the specific synthetic tags preferred by the NILC team, making direct comparison possible. A special challenge in this case was the distinction between 6 different verbal "valency word classes", VAUX, VLIG, VINT, VTD, VTI, VBI, roughly matching the (instantiated) CG valency tags <x>, <vK>, <vi>, <vt>, <vp> and <vdt>/<vtp>, respectively.

As can be seen from the list, the parser can handle a fairly broad spektrum of Portuguese language data. The largest task, the tagging of 3 years of running newspaper text (Folha de São Paulo) for a research group at the Catholic University of São Paulo, took 50 hours of CPU processing time on a linux system, averaging a speed of 500 words per second, and demonstrated the robustness of the system not only in grammatical, but also in technical terms.

So far, no large scale *semantic* annotation has been attempted, and automatic post-CG tree structure annotation of running text has only be performed on test texts and a 20.000 word corpus of teaching sentences.

[241] I would like to thank professor Ataliba de Castilho for making the NURC corpus accessible to me in electronic form.
[242] In this connection, I would like to mention Tony Berber Sardinha who is having a great deal of to-be-rewarded confidence in my parser.
[243] This text and the Tycho Brahe tag set was kindly made available by Helena Britto.
[244] I would like to thank the NILC team for letting me have a go at their corpus, and Sandra Maria Aluisio for having patience in discussing tagging differences with me.

7.4 Machine translation

7.4.1. A teleological judgement perspective

Since parsing is not an independent goal in its own right, different parsing schemes should be judged not only in terms of inherent criteria, such as information content and error rate, but also from a teleological perspective. Ultimately, the crucial user based criterion will be *which uses* a certain parsing scheme is likely to be put to. We have already seen that different syntactic notational systems have to match certain theoretical backgrounds, like functional or generative grammar, and have different uses in teaching (tags vs. PoS colour notation, word based "kryds & bolle"[245] function vs. tree diagrams). Likewise, corpora using tag based flat annotation are easier to search with ordinary string manipulation tools than graphical trees.

From a **machine translation perspective**, the following traits of the Portuguese Constraint Grammar parser seem relevant:

♦ Detailed, word order independent, *function tags* make it easier to transform source language structure into target language structure, without too many complicated transformation rules. Especially where languages like Portuguese are involved, which - unlike English - permit a great deal of variation in the order of clause level arguments.

♦ It is of great importance for polysemy resolution to know which of a word's potential *valency patterns* has been instantiated in a given clause or phrase, and which semantic class fills a given valency slot. Therefore it is advantageous that the parsing formalism can handle the disambiguation of valency tags, selection restrictions and other lexicon derived (originally) secondary semantic tags in the same fashion used for morphology and syntax at the lower parsing levels.

♦ The Constraint Grammar formalism can further be used for the context dependent *mapping and disambiguation of translation equivalents* that are *not* listed in the lexicon or *not* linked to specific secondary tags.

♦ The before mentioned *underspecification*, in Constraint Grammar, of certain postnominals, co-ordination and free nominal adjuncts becomes an asset when seen from a machine translation perspective: - First, a large part of these cases is "true *syntactic* ambiguity", which can only be resolved by the fully contextualised listener/reader. - Second, some of these structural ambiguities (prepositional phrase attachment and co-ordination) are fairly universal, i.e. language independent, so that they can be preserved in translation. Making such ambiguity explicit would only put an unnecessary burden on the intermediate levels of the translation module.

7.4.2. The Progressive Level approach in Machine Translation

[245] "Cross & circle", the icons used in Danish primary schools to denote the (often word based) functions of subject and predicator.

I have shown, in chapter 6, how adding an additional (third) layer of Constraint Grammar disambiguation (called *portval*) can establish discriminators for the resolution of lexicon-specified polysemy, drawing on head-dependent and dependent-head valency instantiation, semantic prototype tags and atomic semantic features. In some cases, however, translation equivalents are so idiosyncratic that they cannot be based exclusively on (disambiguated) tags at the word itself, but must be chosen individually and in a context dependent way. This is achieved by progressing to yet another (fourth) level of CG rules, mapping - and, if necessary, disambiguating - *translation equivalents*. In my system, translations equivalents can be replaced (a), “post-appended” (b) or “pre-appended” (c). Replacement will normally replace a base form (a1), but can be forced to ignore the later TL[246] inflexion module (*sic*, a2). Likewise, appended translations (b,c) can be attached to the base form (=) or as an additional, non-inflecting, word (+).

(a1)

MAP (@komme_//af_/med) TARGET ("acabar" <com^vp>) (*1 PRP-COM BARRIER CLB/VFIN LINK 0 &<PIV LINK *1 &P< LINK 0 N-HUM) ; # choose the new translation if “acabar” has a <com^vp> valency for a +HUM object, e.g. ‘acabar com os bandidos’

MAP (@tænde_/for) TARGET ("ligar") (*1 &<ACC BARRIER &NON->N LINK 0 (<il>) OR (<mu>) OR (<ild>) OR (<lys>) LINK *1 CLB/SB OR VFIN BARRIER PRP-COM) ; # choose the new translation if “ligar” has a direct object that is a musical instrument <mu>, a light <lys>, a fire <ild>, or a tool or machine <il>, and if there is no potential “com”-argument in the same clause (in which case the translation would be ‘forbinde med’)

(a2)

MAP (@lad_os+sic) TARGET (V 1P) (*-1 >>> BARRIER NON-ADV/IN/KC/KOMMA) (*1 <<< LINK 0 EXCLAM-MARK) ; # handles 1. person plural “imperatives” in Danish by preposing ‘lad os’ to the verb’s base form

(b)

MAP (@=isk) TARGET (<attr> N) (0 &N< OR &<SC) (-1 &>A OR ("parecer")) ; # adds ‘isk’ to attributive <attr> nouns if they are *used* attributively: ‘um presidente um pouco iconoclasta - en lidt ikonoklastisk præsident’

MAP (@+-quote) TARGET (<v-cog> IND VFIN) (0 3S OR 1S OR 1/3S) (-1 >>>) (*1 CLB-ORD OR <<< BARRIER &<ACC OR &&ICL-<ACC LINK NOT 0 &&FS-<ACC) ; # marks quoting-verbs for the *permut* [247] program for Danish TL syntax, which is to produce VS word order in the quoting clauselet rather than SV.

(c)

MAP (@skulle_//=) TARGET (V SUBJ) (*-1 &&FS-P< BARRIER CLB OR &MV LINK 0 ("que") LINK -1 PRP-PARA) ; # used in the case of a Portuguese subjunctive after the composite conjuction ‘para que’, adds Danish modal in front of verb, marking the old verb form as detachable (//), thus moving inflexion onto the modal: ‘para que ela voltasse’

In applicational terms, what CG polysemy resolution and CG semantic mapping can achieve, in connection with a (Portuguese-Danish) lexicon look-up program (called *trad*), is a kind of in-text dictionary service which could be integrated into text

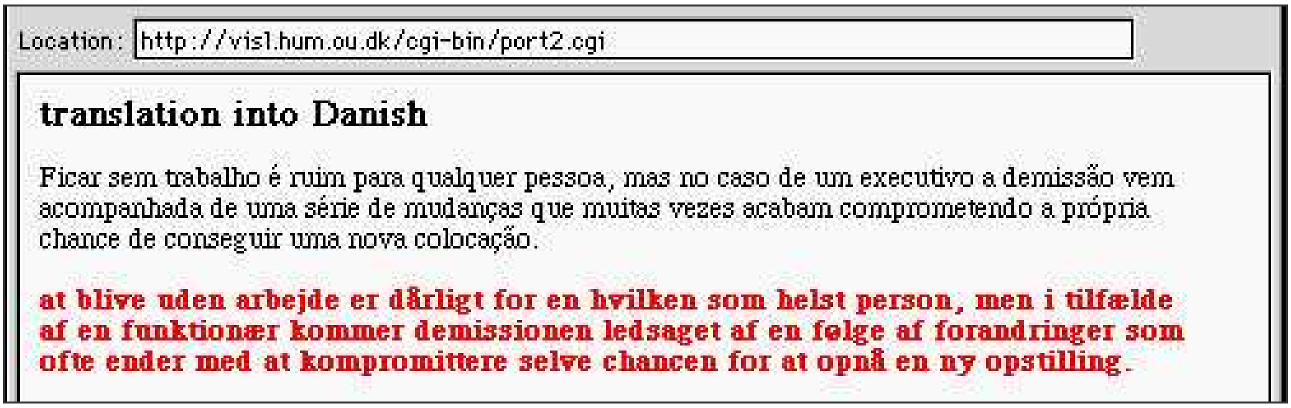
[246] In this chapter, the abbreviations SL and TL stand for *source language* and *target language*, respectively.

[247] The *permut* program module handles syntactic differences between Portuguese and Danish: general word order, complex tense marking, anaphora, VFIN - @SUBJ inversion etc.

processors used by translators or language students. In order to achieve full running machine translation, however, translating base forms obviously isn't enough. I have therefore written two additional programs, *permut* and *danmorf*, that handle the generation of (Danish) TL syntax and TL morphology, respectively. From a performative point of view, *permut* is not too different from a CG system, since it (a) is compiled from a set of context sensitive grammatical rules, and (b) works by string manipulation, treating sentences as word-&-tag strings. The difference is that *permut*, unlike a CG, not only removes and adds information, but also replaces information and - primarily - changes the order of words, groups and clauses - in some cases even that of morphemes. *Permut* handles things like complex tenses, NP-agreement, enclitic articles and pronouns, incorporated (elliptic) pronouns and pronoun anaphora, group-clause and clause-group conversion, reflexivity removal and addition, prepositional "case", incorporating verbs, main- and subclause word order etc. *Danmorf* takes as input translated base forms (from *trad* and the translation mapping CG) as well as attached word class and inflexional information (modified by *permut*), and generates (Danish) target language word forms in the order specified by *permut*. To this end, *danmorf* integrates a Danish base form lexicon with PoS and inflexion class information.

Even without the use of the CG translation mapping module (i.e. only using the valency and semantic feature instantiation performed by *trad*), *permut* and *danmorf* can turn the parser's Portuguese output into intelligible running Danish text:

(1) live, CG based, machine translation



translation into Danish

Ficar sem trabalho é ruim para qualquer pessoa, mas no caso de um executivo a demissão vem acompanhada de uma série de mudanças que muitas vezes acabam comprometendo a própria chance de conseguir uma nova colocação.

at blive uden arbejde er dårligt for en hvilken som helst person, men i tilfælde af en funktionær kommer demissionen ledsaget af en følge af forandringer som ofte ender med at kompromittere selve chancen for at opnå en ny opstilling.

Though the system's present MT is often fairly crude for longer sentences, this is due to the fact that the semantic rule body is still quite small in comparison with the parser's morphological and syntactic disambiguation rules. Long term, MT perspectives seem promising, and in principle, the system can be made to handle all kinds of semantic and structural distinctions, provided that the necessary CG rules are added for mapping and feature instantiation.

7.5 The applicational potential of PALAVRAS: Summary

Experiments with notationally filtered Constraint Grammar analyses for (Portuguese or English) free, running text have shown that an efficient parser can transcend its traditional corpus annotation and research tasks. Among other things, a CG parser can be turned into a valuable grammar teaching tool, - especially if it can be accessed through a "non-technical" interface, which honours the four basic principles of "live" teaching: Interactivity, flexibility, naturalness and tutoring. By exploiting the distributed character of the Internet, one or more central grammar servers can service a large number of simultaneously active, individualised versions of the teaching interface, at the same time allowing easy up-dating and solving collateral problems of copy-right, compatibility and accessibility.

As to naturalness, students cam work with free language samples and use the tools they know from other "friendly" software, like windows, mouse and menus. In fact, the interface can be run "single-handedly", by mouse alone, without ever touching the keyboard. As to flexibility, one can choose from different levels of analysis and descripitional complexity, and even move between different schools of syntactic description. Users may either ask for a ready analysis or interactively build their own with the computer tutoring their choices, providing definitions, translating text and exemplifying concepts. Finally, more research-minded students can venture into the realm of corpus analysis and put grammatical notions to the test.

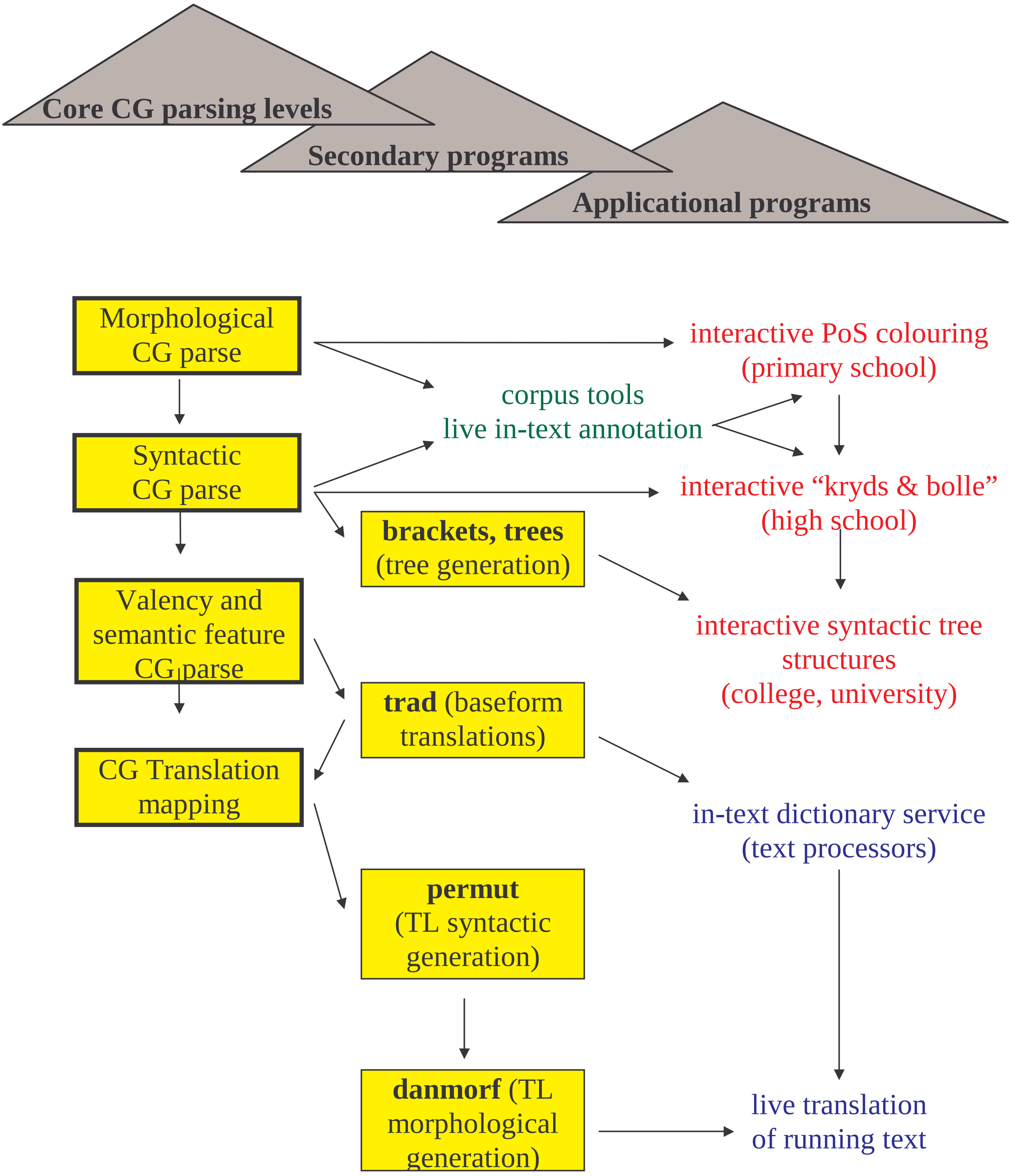
Pedagogically, I have advocated the advantages of word based form and function markers (tags), flat dependency syntax and in-text meta-notation in the form of colour codes and indexing. Ideally, in the case of "wrong" analyses, students should not be criticised for diverging choices if these are motivated by different grammatical backgrounds. Likewise, - unless the student explicitly asks for it -, testing should not focus on the quantification of errors ("scores"), but on the game aspect of the challenge, i.e. the process as such, not the result. In this vein, the interface features a sentence randomiser suggesting unknown sentences to the student for interactive analysis.

Finally, integration of the IT teaching tool into the broader context of ordinary, pre-existing language teaching is encouraged. Here, special notational filters on top of Constraint Grammar parsers, as well as text book based closed corpora are possible solutions. This way, given the inherent flexibility of the interface, it should not be too difficult to introduce similar Internet tools on all levels of language teaching, in universities as well as in secondary and primary schools.

As the most advanced of applications implemented so far, a CG based machine translation system has been designed, exploiting tagging information and disambiguation tools from different levels of the PALAVRAS parser for polysemy resolution, supplemented by a translation equivalent mapping CG, a syntactic transformation program and a Danish morphological generator.

The following diagram shows how different applicational modules are grafted onto the CG based progressive level parsing modules of PALAVRAS:

Applicational add-on program modules for PALAVRAS


Core CG parsing levels
Secondary programs
Applicational programs
Morphological CG parse
Syntactic CG parse
Valency and semantic feature CG parse
CG Translation mapping
corpus tools
live in-text annotation
brackets, trees (tree generation)
trad (baseform translations)
permut (TL syntactic generation)
danmorf (TL morphological generation)
interactive PoS colouring (primary school)
interactive “kryds & bolle” (high school)
interactive syntactic tree structures (college, university)
in-text dictionary service (text processors)
live translation of running text

8

Conclusion:
The advantages of incrementality

8.1 Evaluating the Portuguese parser: Specific conclusions

In chapter 3.9, I have shown that the Portuguese CG parser achieves correctness rates on free text of over 99% for morphology/PoS and 96-97% for syntax, which compares favourably (ch. 3.5) to both PSG-systems - which are not robust and do not usually run on free text -, and to probabilistic systems, which hover around the 97% correctness mark for PoS-tagging, and only rarely succeed in analysing even medium size sentences correctly in their syntactic entirety. I have suggested (following Chanod & Tapanainen, 1994) that the advantage of the lexicon and rule based CG approach over a probabilistic approach resides in the possibility of formulating rules for exceptions, individual lexemes or rare patterns without disturbing the functionality of the majority cases, and - as opposed to HMM-taggers in particular - in the frequent use of long range and unbounded context restrictions (cp. rule type statistics, ch. 3-7-3). On the other hand, I have striven to document, that a CG grammar's advantage over another major family of rule based systems, PSG-grammars, is not limited to the approach immanent robustness of a parser that expresses syntactic function by *tags,* and disambiguates rather than generates, but also can be made visible on the PSG-grammars' home turf, syntactic tree structures. Thus, in the Portuguese system, I have incorporated *dependency markers* on the clause level (as opposed to only using them on the group level, like in the "traditional" ENGCG system), and introduced subclause function tags for finite and non-finite subclauses. Also, as practical proof of the system's dependency information content, a compiler and a set of transformation rules were crafted to transform CG-output into PSG-style syntactic trees.

Within the growing family of CG-based taggers/parsers, the Portuguese system is the only fully developed parser for a Romance language, so a certain typological interest is justified in the degree to which the Portuguese system differs from or resembles other Constraint Grammars. Areas of interest are (a) the notational system as such, (b) ambiguity and rule set typology, and (c) performance.

At present, Constraint Grammar projects have been launched for a variety of languages, of which at least 5 (English, Portuguese, Swedish, Norwegian, Estonian)[248] have published morphological and/or syntactic tag sets. A comparison shows (below) that at least the Indo-European grammars share large parts of their

[248] In my Spanish Constraint Grammar, the morphological and syntactic tag sets are almost identical to the ones used for Portuguese.

annotation system (probably inherited directly from the English system described in Karlsson et.al. [1995]), allowing direct comparison in terms of parsing output and performance. While there is almost complete conceptual overlap with regard to PoS categories, there are more differences on the syntactic level. Thus, the tags for dependents of nominal heads differ as to which *type* of tag is used to convey tagging information. Thus, Portuguese does not add word class and semantic class information to the functional tags, as is the case for @QN>, @DN>, @AN>, @NN> etc. in the English and Swedish systems. A more important difference is that the Portuguese parser tags dependency direction on the *clause* level (e.g. @ACC> vs. @<ACC). A new English system, FDG (Functional Dependency Grammar), built by Pasi Tapanainen, Atro Voutilainen and Timo Järvinen on top of an improved ENGCG2, aims at specifying *all* dependency links, using a refined CG-compiler allowing the mapping and disambiguation of dependent-head relations. Similarly, the output of the Portuguese Constraint Grammar proper can be processed by a constituent assembling grammar *(brackets, trees)* to yield syntactic tree structures and constituent head tagging, a transformation which also implies full dependency specification.

(1) Table: CG word class categories across languages

	Portuguese	**English***	**Swedish**	**Norwegian (bokmål)**	**Estonian**
Noun	N	N	N	subst	S
Proper noun	PROP			subst prop	
Pronoun	SPEC, PERS	PRON	PRON	pron	P
Determiner	DET	DET	DET	det	
Article	<art> DET	DET ART	DET ART		
Adjective	ADJ	ADJ	A	adj	A
Adverb	ADV	ADV	ADV	adv	
Verb Finite	V V VFIN	V V VFIN	V	verb	V
Infinitive	V INF	V INF	V INF	verb inf	
Participle	V PCP	PCP2 (EN)	V SUPINUM A <PCP2>	verb perf-part adj <perf-part>	
Gerund	V GER	PCP1 (ING)	A <PCP1>		
Preposition	PRP	PREP	PREP	prep	K
Conjunction	CO	CC	CC	konj	?
coodinating					
subordinating	KS	CS	SC	sbu	?
Numeral	NUM	NUM	?	det kvant	
Interjection	IN	INTERJ	INTERJ	interj	
Abbrebiation	<abbr>	ABBR	ABBR	abbr ?	
Infinitive marker		INFMARK	INFMARK	inf-merke	
Negation particle		NEG-PART			

*For English, EngCG-2 (as presented by www.conexor.fi), uses EN instead of PCP2, and ING instead of PCP1.

(2) CG syntactic tag sets across languages

	Portuguese	**English***	**Swedish**	**Norwegian (bokmål)**	**Estonian**
Finite main verb	@FMV	@+FMAINV	@+FMV	@FV @IV	@FV @>FV
Non-finite main verb	@IMV	@-FMAINV	@-FMV		
Finite auxiliary	@FAUX	@+FAUXV	@+FCV		
Non-finite auxiliary	@IAUX	@-FAUXV	@-FCV		
Subject	@SUBJ> @<SUBJ	@SUBJ	@SUBJ	@SUBJ	@SN @SP @S
Formal subject		@FSUBJ	@FSUBJ		
Subject marker			@<SUBJM		
Direct object	@ACC> @<ACC @ACC>>	@OBJ	@OBJ	@OBJ	@ON @OG @OP @O
Formal object			@FOBJ		
Dative object	@DAT> @<DAT	@I-OBJ	@IOBJ	@I-OBJ	
Prepositional object	@PIV> @<PIV				
Subject complement	@SC> @<SC	@PCOMPL-S	@SCOMP	@S-PRED	@CN
Object complement	@OC> @<OC	@PCOMPL-O	@OCOMP	@O-PRED	@CP @C
Adverbial object	@ADV> @<ADV	@ADVL-O (np’s)			
Adverbial adjunct	@ADVL@ADVL> @<ADVL	@ADVL	@ADVL	@ADV	@Q
Interjection			@INTERJ	@INTERJ	
Free predicative	@PRED> @<PRED				
Stray NP-head	@NPHR	@NPHR	@NPHR	@LØS-NP	
Determiner premodifier	@>N DET	@DN>	@DN>	@<DET	@A
Adjective premodifier	ADJ	@AN>	@AN>	@ADJ>	
Genitive modifier		@GN>	@GN>		
Quantifier modifier	<quant>	@QN>	@QN>		
Noun premodifier	N	@NN> @A>*	@NN>	@SUBST>	
Postmodifier of noun	@N< N PRP @#ICL	@<NOM @<NOM-OF @<NOM-FMAINV	@<NN	[@<SUBST]	
Argument of adjective	@A<		@AOBJ		
Argument of participle	@A<PIV @A<ADVL @A<PASS				
Title	<+n>	@TITLE		@TITTEL	
Apposition	@APP @N<PRED	@APP	@<NN	[@APP]	
Adverbial premodifier	@>A	@AD-A>	@AD> @PA>	@ADV>	
Adverbial postmodifier	@A<		@<ADV	@<ADV	
Argument of preposition	@P<	@<P @<P-FMAINV*	@<P @P>>	@<P-UTFYLL	@P< @>P
Modifier of preposition	@>P				
Co-ordinator	@CO	@CC	@CC	@KON	
Subordinator	@SUB	@CS	@CS	@<SBU(-REL)	
Auxiliary particle	@PRT-AUX<	@INFMARK>	@IMCV		
Sentence apposition	@S<				
Vocative	@VOK	@VOC*			
Focus marker	@FOC> @<FOC				
Finite subclause	@#FS-				
Non-finite clause	@#ICL-				
Averbal subclause	@#AS-				
Argument of complementiser	@AS<				

*For English, conexor’s FDG (finite dependency grammar) uses @A> instead of @NN>, and has introduced the tag @VOC.

Sources:

English:

EngCG: Karlsson et. al. (1995)

FDG: http://www.conexor.fi/fdg.html#1 (14.3.1999), by conexor

Portuguese: Bick (1996)

Swedish: http://huovinen.lingsoft.fi/doc/swecg/intro/stags.html (23.12.1998), by lingsoft

Norwegian: http://www.hf.uio.no/tekstlab/tagger.html (23.12.1998), by Janni Bonde Johannesen

Estonian: http://www.cl.ut.ee/ee/yllitised/first/kailimyyrisep.html (23.12.1998) by Kaili Müürisep, University of Tartu

In their comparison between ENGCG and their newly developed FDG, Voutilainen and Tapanainen (http://www.conexor.fi) report morphosyntactic success rates (percentage of correct morphosyntactic labels present in the output) of 94.2-96.8% for ENGCG and 96.4-97% for FDG, with an ambiguity rate of 11.3-13.7% for the former, and 3.2-3.3% for the latter. The Portuguese parser compares favourably to this, achieving about the same success rate as FDG (96.4-97.5) even with an ambiguity rate close to zero.

For Estonian, Müürisep (1996), reports a syntactic error rate of 0.32%, but with an ambiguity rate of 32% (1.47 tags per word), making a direct comparison difficult.

On the morphological level, performance is evidently better than on the syntactic level, for all CG systems. The “classic” ENGCG can be regarded as a base line, with an error rate of only 0.3% at 3-7% disambiguation. For SWECG 1.0, a Swedish Constraint Grammar (where no performance data on the syntactic level could be obtained at the time of writing), morphological performance is about the same as for English, with an error rate of 0.3%, at an ambiguity rate of 5% (www.sics.se/ humle/ projects/ svensk/ projectPlan.html, by Mikael Eriksson, Björn Gambäck and Scott McGlashan, accessed on 23.12.98). The Portuguese parser, by comparison, has an error range between 0.3% and 1.2%, with 0% ambiguity.

With regard to the *dependency* performance of PALAVRAS, only a comparison with FDG makes sense. To compile table (4), a 5000-word text chunk was automatically analysed by a tree-generating version of PALAVRAS (the one used internally in the VISL grammar teaching programs), producing vertical tree output as in the example below (3).

(3)

```
STA:fcl
=SUBJ:np
==>N:art(<artd> M S)                      O               The
==H:n(M S)                                ministro        minister
=ADVL:adv                                 já              already
=MV:v-fin(PS 3S IND VFIN)                 patenteou       adopted
=ACC:np
==>N:art(<arti> M S)                      um              a
==H:n(M S)                                estilo          style
==N<:pp
===H:prp                                  de              of
===P<:np
====H:n(M S)                              trabalho        work
====N<:fcl
=====SUBJ:pron-indp(<rel> M/F S/P)        que             which
=====ACC:pron-pers(M 3S ACC)              o               him
=====MV:v-fin(PR 3S IND VFIN)             diferencia      distinguishes
=====PIV:pp
======H:prp                               de              from
======P<:np
=======>N:pron-det(<poss 3S/P> <si> M P)  seus            his
=======H:n(M P)                           antecessores    predecessors
=.
```

Recall and precision were calculated for individual function tags, which here refer to tree-nodes, not words. Unlike the flat CG-notation, the tree-notation makes all dependency attachments visually explicit. In addition, for better or worse, some dependencies underspecified in CG (especially postnominals) are resolved in tree-notation. In the case of true (in-sentence) ambiguity, attachment was judged as correct, if *either one* of the correct readings made it into the tree.

For comparison, FDG numbers are quoted from http://www.conexor.fi (accessed on 14.3.99). For both systems, the numbers given concern newspaper texts.

(4) **Table: syntactic tree structure analysis - performance**

	Portuguese CG - PALAVRAS (with full disambiguation) 4937 words, 6674 nodes							English FDG	
	cases	precision			recall			preci -sion	re- call
		tag alone	tag/ attach - ment	attach-ment alone	tag alone	tag/ attach-ment	attach-ment alone	tag/ attachment	
@SUBJ	351	97.3	**97.3**	100	93.4	**93.4**	98.0	*95*	*83*
@ACC	368	95.7	**95.4**	99.7	97.2	**97.0**	98.6	*94*	*88*
@PIV	88	93.1	**93.1**	100	92.0	**92.0**	100		
@ADV	19	84.2	**84.2**	100	84.2	**84.2**	100		
@SC	113	92.2	**92.2**	100	94.7	**94.7**	99.1	*92*	*96*
@OC	17	100	**100**	100	82.4	**82.4**	88.2		
@MV	596	99.3	**99.3**	100	99.7	**99.7**	100		
@AUX	87	98.9	**98.9**	100	100.0	**100.0**	100		
@AUX<	96	98.9	**98.9**	100	97.9	**97.9**	100		
@ADVL	518	92.5	**91.9**	99.4	95.4	**94.8**	96.7		
@PRED	48	87.0	**84.8**	97.7	83.3	**81.3**	87.5		
@APP	20	84.2	**84.2**	94.7	80.0	**80.0**	90.0		
@P<	911	99.3	**99.3**	100	98.9	**98.9**	99.0		
@>A	45	92.7	**92.7**	100	84.4	**84.4**	84.4		
@A< (PCP)	43	97.0	**97.0**	100	76.7	**76.7**	79.1		
@A< (other)	26	100	**100**	100	88.5	**88.5**	88.5		
@KOMP<	10	100	**100**	100	90.0	**90.0**	90.0		
@>N	1029	98.9	**98.9**	100	99.5	**99.5**	100		
@N<	749	98.6	**97.1**	98.5	95.7	**94.3**	94.5		
@SUB	71	100	**100**	100	98.6	**98.6**	100		
@COM	17	94.1	**94.1**	100	94.1	**94.1**	100		
@NPHR	60	75.0	**75.0**	100	100	**100**	100		
@AS<	25	95.7	**95.7**	100	88.0	**88.0**	100		
unnamed								*95.3*	*87.9*
all of above	5307	97.1	**96.8**	99.5	96.9	**96.6**	97.9		

In order to retain comparability with ordinary CG performance numbers, and because they have no “upward” dependency links, @H (head of group) and the sentence top node @STA (statement) are not used in the above table. Of 1988 @H-tags and 264 @STA-tags, only 2 were wrong, and including them in the performance calculations would “improve” both recall and precision in a meaningless way. Rather, only those function tags are measured that would have appeared in a pure CG-analysis, too. The discrepancy between a word count of 4937 and a function node count of 5307 is due to the fact that, in my CG-notation, subclause function is tagged as a number two tag onto complementisers or main verbs, but will appear as its own node in constituent tree analysis (on which the table is based). Though PALAVRAS tags co-ordinators for what they co-ordinate (e.g. <co-acc> for direct object co-ordination), the CG-to-tree transformation program used in the evaluation, did not yet handle co-ordination, so paratactic attachment was not quantified. A distinction was made between clause- and group-level @PRED, but appositions (@APP) were regarded as clause level constituents by the tree-generator.

The shaded columns (tag only) contain data directly reflecting CG tag output, while the bold face columns (tag/attachment) show the decrease in performance if attachment errors are counted, too, even where function tags are correct. The third column type (attachment only), finally, reflects pure attachment performance, judging the tree as such, without taking function tag errors into account.

For the system as a whole, PALAVRAS' recall and precision converge on the 97% syntactic tag correctness mark known from other text samples (cp. chapter 3.9)[249]. The fact that not much (0.3%) is lost when pure attachment errors are included is encouraging proof that CG-to-tree transformation *is,* in fact, feasible. A recall and precision for dependency *per se*[250] of over 97.9% and 99.5%, respectively, suggest that the attachment information contained in PALAVRAS' output is actually *more* robust than its function tag information.

There is a fair deal of variation in the specific performance data for individual constituents. By comparison to the English FDG, PALAVRAS has a better recall for @SUBJ and @ACC, and a slightly worse one for @SC.

Interestingly, subjects have a high precision and a (relatively) low recall, while direct objects have a (relatively) low precision, but a high recall, suggesting that the present rule set could be biased in favour of @ACC and against @SUBJ.

Apart from the @ACC - @SUBJ discrepancy, performance was best for verbal function and subordinators (where the real disambiguation task resides in the morphology), as well as prenominals (@>N) and arguments of prepositions (@P<). Most problematic were @OC, @APP, clause level @PRED, and @A< adjects after attributive participles. The last case covers cases where participles with group-level function govern what would amount to adjuncts and objects on the clause level (e.g. 'um artigo <u>publicado, em 1992, por seu amigo'</u>). In all four cases precision is higher than recall, but for @OC precision is all of 100%, and for post-participle @A< still 97%, illustrating the fact that both are heavily dependent on valency entries in the

[249] Normally, for the tag set as a whole, recall and precision should be identical at 100% disambiguation. The small difference here (0.2%) is due to partially unrecovered complementizer tags and ICL-tags in words that in my system bear both an in-clause function tag and the subclause's own function tag. The text chunk in question had a relatively low *morphological/PoS* error rate, of only 0.2%, reflecting - probably - chance features like the absence of foreign nouns heuristically mis-tagged as verbs. To avoid such text type fluctuation, much larger and balanced samples would have to be annotated and proof-read - generating a work load beyond the scope of this dissertation.

[250] To come up with values for *dependency* recall and precision, the number of *"recovered" correct attachments* was calculated as the number of possible ("original") category X attachments (equal to the the number of correct function tags X), minus unrecovered instances of X that in addition had wrong attachment, <u>and</u> minus instances of X that were correctly tagged, but wrongly attached. The number of *all surviving attachments*, on the other hand, was calculated as possible attachments of X minus unrecovered and wrongly attached instances of X (but <u>not</u> minus correctly tagged and wngly attached X).

lexicon which ensure removal of false positive argument mappings, but are inadequate to capture the full productive range of, for instance, time-, place- and manner-modifiers, that by morphosyntactic form are indistinguishable of clause-level adjuncts.

It is notable that attachment-only errors were almost completely restricted to post-nominals and adverbials, plus the odd free predicative or apposition, all of which are not bound by the uniqueness principle. Adverbials can be ambiguous with regard to position in the clause, and post-nominals are frequently ambiguous as to order of attachment.

Though there are, of course, marked typological differences between Portuguese and English on both the morphological level (inflecting vs. isolating) and the syntactic level (optional subject vs. obligatory subject, relatively free word order vs. fixed word order), I have found no major differences in terms of CG-typology. Average morphological ambiguity before disambiguation, for instance, is about 2 for both languages, and even V - N ambiguity, typical for an inflexion-poor language like English, is a sizeable problem in Portuguese, too, due to the fact that the masculine and feminine noun/adjective singular endings *-o*, and *-a* both occur in most verbal paradigms, too. On the syntactic level, neither English nor Portuguese case-marks noun subjects or noun objects, but Portuguese has the additional disadvantage of allowing VSO and VOS word order (OV is mostly restricted to pronoun objects, which *are* case marked). With regard to finite subclauses, Portuguese - unlike English - demands an obligatory complementiser (conjunction, relative or interrogative), which facilitates clause boundary resolution in the Portuguese CG. On the other hand, Portuguese allows rich *non*-finite subclauses (usually *without* a complementiser), which complicate matters.

Concluding from Portuguese, I would like to suggest the following CG universals, some of which mirror similar findings published for ENGCG, and might thus be thought to hold across languages:

- A full-grown Constraint Grammar needs thousands of rules for each additional level of analysis (in the case of Portuguese, morphology, syntax and valency/semantics), though of course the number of rules (or even contexts) can not be used to predict the recall or correctness of a given grammar. Neither (but less obviously) does the number of rules directly reflect a grammar’s precision, since disambiguation gain depends heavily on the word class or function targeted (ch. 3.2.2, table 6). For Portuguese PoS-tagging, noun disambiguation pays best, since nouns have a good disambiguation gain, and no particular ambiguity bias (unlike infinitives, which are highly ambiguous word forms, but at the same time very likely to be, in fact, infinitives).

- REMOVE rules are more typical for Constraints Grammars than SELECT rules, the proportion in the Portuguese Grammar being 2:1 - and rising, as new rules are added. In particular, REMOVE rules increase the robustness of the grammar since alternative readings cannot be removed without being target directly. For the Portuguese grammar, in particular, it is important that REMOVE rules allow the simultaneous management of *different kinds* of “syntactic” (@-) tags (word function, clause function, valency instantiation) that are to survive in varying and multiple combinations), a feature that is inherently relevant for the progressive level approach.

- Rule complexity, measured by number of context conditions and the percentage of rules with unbounded contexts, is higher for syntax than for morphology/PoS. Thus, an average of 3.37 contexts was needed for morphological disambiguation, 4.22 for syntactic mapping and 5.28 for syntactic disambiguation. The proportion of rules with unbounded contexts and rules with only local contexts (the ones that could be expressed by HMM taggers) is 10 times higher for syntactic than for morphological rules (the numbers being 2.0 and 0.2, respectively).
- CG rules are “left-leaning” in the sense that left hand context conditions are more common in rules than right hand contexts, reflecting the linear and sequential composition of language. Thus, on all levels and for both absolute and unbounded contexts, the percentage of left contexts is about 60% (i.e. 40% for right hand contexts), as opposed to 81% left contexts for unbounded and 42.6% for absolute contexts in ENGCG (Karlsson, 1995, p. 352).

With its specific focus on Portuguese Dependency Grammar, and its notational distinction between **form** (N-, V-, S-, P-tags etc. as well as attachment markers) and **function** (@-tags), PALAVRAS represents not only a technical parsing solution, but also a comprehensive description of Portuguese morphology and syntax. The parser and its notational system have constantly been tested on authentic corpora ensuring that no large area of Portuguese syntax remained unconfronted.

8.2 CG Progressive Level Parsing: General conclusions

There does not appear to be any obvious limit as to which level of grammatical distinction can be handled by Constraint Grammars. Rather, performance depends on the amount and type of information available from the lexicon, and on the quality of tags disambiguated on the preceding (lower) level(s) of analysis.

For example, identification of direct objects (@ACC) is rather difficult with only (undisambiguated) morphological information to draw upon, as is the case when morphological CG rules try to PoS-distinguish the Portuguese accusative pronoun ‘a’ (to the left of an N/V-ambiguous word) from the determiner ‘a’ and the preposition ‘a’. The task turns easier if lexical information about transitive valency is provided, and after verbs, for instance, have been PoS-disambiguated: (NOT 1 <vt>) suggests ‘a’ is *not* a pronoun, and (1C VFIN) is a very strong context condition for discarding the determiner and preposition readings. The simultaneous treatment, finally, of the @ACC tag together with *other* syntactic tags (like *other* @ACC objects, or @SUBJ subjects etc.), allows a high degree of correctness: for example, the uniqueness principle can be exploited by adding the context conditions **(i)** (NOT *1 @<ACC) or **(ii)** (*1C @<ACC BARRIER @NON->N) to decide whether an NP-head between a transitive verb to the left and another NP to the right be tagged @<ACC (i) or @<SUBJ (ii).

In the same vein, valency class is difficult to treat as more than lexical potential on the morphological level, but allows easy instantiation *after* syntactic parsing. Semantic (sense) disambiguation, too, becomes accessible only after syntactic function is assigned. Thus, <+HUM> will be selected in a noun that has been “proven” subject (by the syntactic CG-level) of a cognitive or speech-verb. This does not, of course, mean, that <+HUM> cannot itself be a useful *secondary* tag on the syntactic level. In fact, such promotion from secondary to primary tag is - as I have shown in chapters 5 and 6 - typical of and essential for the concept of Progressive Level Parsing as employed in the Portuguese parser. As a bonus, working with different levels of parsing allows the postponement of difficult disambiguation tasks to a later, more information-rich, stage. For instance, I have chosen to retain the word class of adverb for ‘como’, ‘onde’ and ‘quando’ even where they function like “conjunctions”, postponing the assignment of “conjunctionhood” to the syntactic level, where complementiser relatives will be marked as clause headers anyway (by a @#FS-tag). Similarly, the distinction between definite article and demonstrative pronoun (‘o’, ‘a’, ‘os’, ‘as’) is postponed to the valency level where it can easily be resolved by checking for prenominal function (@>N) and discarding the <art>- or <dem>-tags, respectively, turning them from primary into secondary tags.

Another advantage of Progressive Level Parsing is that different linguistic systems of classification can be kept apart and “pure”. Thus, it was possible to define word class largely in morphological terms (by inflexion category inventory, cp.

2.2.5.1), *without* ultimately losing the syntactic or semantic information[251] residing in traditional word class definitions (like the categories of *demonstrative, possessive, indefinite*, *interrogative* and *relative* pronouns, for what I have morphologically defined as *determiners* [DET, number/gender-inflecting] and *independent specifiers* [SPEC, non-inflecting]). Likewise, sense-distinctions alone were not regarded as a sufficient criterion for lexeme-distinctness, reserving, for instance, the distinction of the etymologically unrelated, but in Brazilian Portuguese homographic, 'fato' and 'fa(c)to' for the semantic level - to be performed by sense tag disambiguation. This way, a purely morphological approach to lexeme identity was possible.

One of the main objectives of this dissertation has been to show that the Constraint Grammar approach, which from the beginning has stressed the importance of a parsing lexicon and grown from morphology into syntax (Karlsson, 1995, p.11), is ideally suited for such progression, not alone towards more "delicate" syntax, but also with respect to notational filtering (constituent or dependency trees) and, ultimately, semantics and semantics-based applications (like MT), - provided the lexicon is upgraded along the way. It appears justified to say that the progression on the syntactic level, involving subclause function, clause level dependency markers and constituent tree transformation, has yielded quantifiable results comparable to what has been achieved for the "benchmark" surface syntax of ENGCG (cp. chapters 3.9 and 8.1).

Though here I have only sketched the outlines of a semantic CG-level (in chapters 5 and 6), a working system has been implemented for the *entire* lexicon, and disambiguation rules have been written for comprehensive valency instantiation, as well as selected areas of polysemy resolution, showing that - at least in principle - Constraint Grammar can be used to address parsing tasks at these levels. Valency instantiation, though it could be performed in many other ways, shows how the CG formalism can be made to handle what basically is a unification process, exploiting *unambiguous* syntactic information that has already been established (tagged). Unlike ordinary disambiguation rules, most of the valency unification rules are relatively simple, REMOVE rules often involving only the target valency tag and two unbounded context conditions, one left, one right, checking for the presence of a matching argument[252]. On the other hand, rules targeting semantic features or semantic prototype tags cannot be based on established syntactic tags alone, but need to take into account other (semantic) *ambiguous* information, resulting in a rule type more reminiscent of the morphological and syntactic levels. Finally, there is ongoing work involving a translation-equivalent Constraint Grammar for Portuguese-Danish machine translation, which is basically a context conditioned *mapping* grammar, refining and correcting the translations obtained by using morphological, syntactic, valency- and semantic tags as polysemy discriminators. And though such a statement should be worded carefully, nothing indicates insurmountable difficulties on the immediate Constraint Grammar application horizon ...

[251] This information, no not morphologically explicit, is retained by means of secondary and function tags, which can be exploited to recreate the traditional pronoun classes in a given application.

[252] For instance: REMOVE (<vt>) (NOT *1 @<ACC OR @#FS-<ACC) (*-1 SB/CLB/VFIN BARRIER @ACC>)

Methodologically, it is commonly claimed that the Constraint Grammar concept is robust as a *parsing technique* (among other things, because one reading always survives disambiguation). In addition, I would like to claim that Constraint Grammar, as it is practised by its present research community, is also quite robust as a *grammatical system*.

First, Constraint Grammars are written in a corpus based environment with "quantitative" control. The process of writing and rewriting rules on the background of constantly reiterated corpus-performance-checks ensures that a Constraint Grammar remains close to "real" language, confronting every conceivable niche of syntactic variation, derivational productivity etc. A CG system is at no stage sheltered by a "toy lexicon" or a "laboratory grammar".

Second, I have illustrated that word and tag based flat dependency grammar, while being a robust starting point for transformations into other grammatical systems, also has notational robustness advantages in its own right: As such, CG-style dependency notation is a more robust system of syntax than a constituent tree analysis, since certain attachment ambiguities (for instance, co-ordination and PP-attachment) are left underspecified at the syntactic level, whereas constituent analysis forces distinctions that often are not meaningful except on the perceptual, humanly contextualised, level.

Third, a grammatical description handled and implemented by a CG parser, is in its very nature empirical in a unique way, ensuring a valuable and interesting kind of authenticity. Since new sets and secondary tags are introduced into the grammar along the way, corpus data and corpus "needs" are allowed to actually shape the grammar itself. This is entirely different from the purely statistical, lexicographic and stylistic uses ordinarily made of corpora. For instance, the category of *"cognitive verb"* (<vcog>[253], as a hybrid syntactic category with semantic interpretation) was added along the way, growing from the disambiguational need to tag for a valency selection restriction concerning direct object que-clauses (that-clauses). Likewise, the category of *"ergative verb"* was not defined a-priori, but derived from corpus cases where verbs are particularly likely to precede their subject (@<SUBJ). And in the face of newspaper corpus data, a set of speech verbs (V-SPEAK) had to be defined in order to tag correctly post-verbal subjects after quotations. This way, while necessary as a point of departure, the a-priori grammatical concepts of the CG grammarian are constantly checked, updated and modified by the needs of the system. Ultimately, like with the physical laws of nature, simplicity and workability are allowed to become strong arguments for preferring one theory over another[254].

[253] This category is also used in the English CG described in Karlsson et. al. (1995). For Portuguese, the list of "cognitive" verbs was compiled from corpus-excerpts where verbs are followed by conjunctional 'que' or interrogatives.

[254] Of course, such empirically and process-dictated distinctions may be in conflict with the applicational intentions or grammatical background of the grammarian, and should not, therefore, be allowed "to get out of hand". However, much can be achieved simply by filtering and postprocessing CG output with other programs, leaving the CG system itself untouched, as it works best.

Appendix: The tag set

WORD CLASS TAGS

N	Nouns
PROP	Proper names
SPEC	Independent specifier pronouns (defined as non-inflecting pronouns, that can't be used as prenominals): e.g. indefinite pronouns, nominal quantifiers, nominal relatives
DET	Determiner pronouns (defined as inflecting pronouns, that can be used as prenominals): e.g. articles, attributive quantifiers
PERS	Personal pronouns (defined as person-inflecting pronouns)
ADJ	Adjectives (including ordinals, excluding participles[255] which are tagged V PCP)
ADV	Adverbs (both 'primary' adverbs and derived adverbs ending in '-mente')
V	Verbs (full verbs, auxiliaries)
NUM	Numerals (only cardinals)
KS	Subordinating conjunctions
KC	Co-ordinating conjunctions
IN	Interjections

Only used on the morphological level (now obsolete in the parser as such):

EC	Morphologically "visible" affixes (elemento composto, e.g. "anti-gás")

INFLEXION TAGS

Gender:	**M** (male), **F** (female), **M/F** [for: N', PROP', SPEC', DET, PERS, ADJ, V PCP, NUM]
Number:	**S** (singular), **P** (plural), **S/P** [for: N, PROP', SPEC', DET, PERS, ADJ, V PCP, V VFIN, INF, NUM]
Case:	**NOM** (nominative), **ACC** (accusative), **DAT** (dative), **PIV** (prepositive), **ACC/DAT, NOM/PIV** [for: PERS]
Person:	**1** (first person), **2** (second person), **3** (third person), **1S, 1P, 2S, 2P, 3S, 3P, 1/3S, 0/1/3S** [for: PERS, V VFIN, V INF]
Tense:	**PR** (present tense), **IMPF** (imperfeito), **PS** (perfeito simples), **MQP** (mais-que-perfeito), **FUT** (futuro), **COND** (condicional) [for: V VFIN]
Mood:	**IND** (indicative), **SUBJ** (subjunctive), **IMP** (imperative) [for: V VFIN]
Finiteness:	**VFIN** (finite verb), **INF** (infinitive), **PCP** (participle), **GER** (gerund) [for: V]

(In this table, " ' " after a category means, that the category in question for this word class is a lexeme category, and thus derived directly from the lexicon. No " ' " means, that the category in question is a word form category for this word class, and thus expressed by inflexion.)

<*> the asterisk stands for capitalisation (<*>) and quotes[256] (<*1> and <*2>)

[255] The “adjectivity” of past participles can be concluded from the adnominal (@>N, @N<) function tag, and is lexically marked <adj> for the most common cases (not least from a bilingual perspective, i.e. where translation equivalents in Danish would be adjectives). Participles *not* used “verbally”, are recognizable by means of their @V function tag (@IMV, @IAUX).

[256] <*1> means a left quote («), or a neutral quote attached at the beginning of a word (“xxx), without interfering blanks, <*2> means a right quote (»), or a neutral quote attached at the end of a word (xxx”), without interfering blanks. Free neutral quotes, not attached to words, are marked $”.

$ the dollar sign is used to mark non-word items, like punctuation marks ('$.' for a fullstop, '$,' for a comma) and numbers ('$1947')

SYNTACTIC TAGS

@SUBJ> @<SUBJ subject

@ACC> @<ACC accusative (direct) object

@DAT> @<DAT dative object (only pronominal)

@PIV> @<PIV prepositional object

@ADV> @<ADV adverbial object (place, time, duration, quantity, cp. <va>, <vta>)

@SC> @<SC subject predicative complement

@OC> @<OC object predicative complement

@ADVL> @<ADVL adjunct adverbial

(All above clause arguments [@SUBJ, @ACC, @DAT, @PIV, @ADV, @SC, @OC] and the adverbial complements [@ADVL] attach to the nearest main verb to the left [<] or right [>].)

@ADVL stray adverbial (in non-sentence expression)

@NPHR stray noun phrase (in non-sentence expression without a top-node verb)

@VOK 'vocative' (e.g. 'free' addressing proper noun in direct speech)

@>N prenominal adject

(attaches to the nearest NP-head to the right, that is not an adnominal itself)

@N< postnominal adject

(attaches to the nearest NP-head to the left, that is not an adnominal itself)

@>A adverbial pre-adject (intensifier adject)

(attaches to the nearest ADJ/PCP/ADV or to a attributively used N <attr> to the right)

@A< adverbial post-adject (rare as modifier: *“caro demais”*, more common as argument of adjective: *“rico em”*, or participle)

@A<PIV/ADVL/SC arguments or adjuncts of an attributively used PCP

@APP identifying apposition (always after NP + comma)

@PRED> 'forward' free predicative adjunct

(refers to the following @SUBJ, even when this is incorporated in the VP)

@<PRED ‘backward’ free predicative adjunct or predicative adject

(refers - as adject - to the nearest NP-head to the left, *or* - as adjunct - to the nearest main verb and its subject to the left)

@N<PRED predicate in small clause introduced by 'com/sem' (rare, e.g. 'com a mão *na bolsa*', 'sem o pai *ajudando*, não conseguiu'); also used - in constituent grammar transformation - to indicate adject predicatives (i.e. group level @<PRED)

@P< argument of preposition

@>P modifier of prepositional phrase *(“até em casa”, “muito de propósito”)*

@FAUX finite auxiliary (cp. @#ICL-AUX<)

@FMV finite main verb

@IAUX infinite auxiliary (cp. @#ICL-AUX<)

@IMV infinite main verb

@PRT-AUX< verb chain particle (preposition or *"que"* after auxiliary)

@CO co-ordinating conjunction

@SUB subordinating conjunction

@KOMP< argument of comparative (e.g. *"do que"* referring to *“melhor”*)

@COM direct comparator without preceding comparative (e.g. “word *like*”)

@PRD role predicator (e.g. "work *as*", "function *as*")

@#FS- finite subclause (combines with clausal role and intraclausal word tag, e.g. @#FS-<ACC @SUB for "não acredito *que* seja verdade")

@#ICL- infinite subclause (combines with clausal role and intraclausal word tag, e.g. @#ICL-SUBJ> @IMV in "*consertar* um relógio não é fácil")

@#ICL-AUX< argument verb in verb chain, refers to preceding auxiliary

@#AS- 'absolute' (i.e. verbless) subclause
(combines with clausal role and intraclausal word tag,
e.g. @#AS-<ADVL @ADVL> in "ajudou *onde* possível")

@AS< argument of complementiser in absolute subclause

@S< statement predicative (sentence apposition)
(refers back to the whole preceding statement: *“não venceu o que muito o contrariou”)*

@FOC focus marker (*“gosta é de carne”*)

VALENCY TAGS and FUNCTIONAL SUBCLASS TAGS

Valency tags functional subclass tags are (somewhat idiosyncratic) lexicon tags that are used by the parser for the disambiguation of “ordinary” primary tags, i.e. morphological word class and syntactic function. Valency tags and functional subclass tags themselves are only partly disambiguated on the morpho-syntactic level, an important example for full disambiguation being the interrogative and relative subclasses of adverbs and pronouns. Tags expressing functional and valency potential can, however, be instantiated at a later stage, drawing on (then) established syntactic information.

The list below defines the most common tags used:

Verbal valency:

<vt> monotransitive verb with accusative (direct) object

<vi> intransitive verb (ideally, inergative)

<ve> ergative (inaccusative) verbs[257]

<vtd> ditransitive verb with accusative and dative objects

<PRP^vp> monotransitive verb with prepositional object (headed by PRP)

<PRP^vtp> ditransitive verb with accusative and prepositional objects

<vK> copula verb with subject predicative complement

<vtK> copula verb with object predicative complement

<va> transitive verb with adverbial argument relating to the subject: <va+LOC>, <va+DIR>,

<vta> transitive verb with adverbial argument relating to the object: <vta+LOC>, <vta+DIR>

<vt+QUANT> transitive verb with NP as quantitative adverbial object (e.g. *"pesar")*

<vt+TEMP> transitive verb with NP as temporal adverbial object (e.g. *"durar")*

<vU> "impersonal" verbs (normally in the 3S-person, e.g. *"chove")*

<x> governs infinitive (as auxiliaries tagged @(F)AUX - @#ICL-AUX<)

<x+PCP> governs participle (all are auxiliaries, tagged @(F)AUX - @#ICL-AUX<)

<x+GER> governs gerund (all are auxiliaries, tagged @(F)AUX - @#ICL-AUX<)

<PRP^xp> governs preposition mediated infinitive
(as auxiliaries tagged as @(F)AUX - @PRT-AUX< - @#ICL-AUX<)

<xt> governs infinitive clause with subject in the accusative case (e.g. ACI- and causative constructions, tagged as @(F)MV - @SUBJ> - @#ICL-ACC)

<PRP^xtp> governs accusative object and prepositional object containing an infinitive clause with its (unexpressed) subject being identical to the preceding accusative object, tagged as @(F)MV - @<ACC - @<PIV - @#ICL-P<)

<vr> reflexive verbs (also <vrp>, <xr>, <xrp>)

<vq> "cognitive" verb governing a 'que'-sentence

<PRP^vpq> "cognitive" verb governing a prepositional phrase with a 'que'-sentence

<qv> "impersonal" verb with 'que'-subclause as subject predicative ("parece que")

<+interr> "discourse" verb or nominal governing an interrogative subclause

Nominal valency:

<+n>noun governing a name (PROP) (e.g. *“o senhor X”*)

<+num> noun governing a number (e.g. *"cap. 7", "no dia 5 de dezembro")*

[257] <ve> was introduced for corpus data reasons, to capture verbs with likely post-positioned “internal” (patient) subject, absolute participle constructions etc. The <ve> group may grow further on the expense of the <vi> group, and therefore, <vi> and <ve> can not (yet) be regarded as disjunct concepts (of inergative and inaccusative, respectively).

<num+>	"unit" noun (e.g. "20 metros")
<attr>	attributive noun (e.g. *"um presidente comunista*")
<mass>	mass noun (e.g. *"leite", "água")*
<+INF>	nominal governing infinitive (N, ADJ)
<+PRP>	nominal governing prepositional phrase headed by PRP, e.g. <+sobre>
<PRP+>	(typically) argument of preposition PRP
<+que> <+PRP+que>	nominal governing a 'que'-subclause (N, ADJ)

Syntactic and semantic subclasses of pronouns, adjectives, adverbs and numerals

<art>	definite article (DET)
<arti>	indefinite article (DET)
<quant0/1/2/3>	quantifier (DET: <quant1>, <quant2>, <quant3>, SPEC: <quant0>)
<dem>	demonstrative pronoun (DET: <dem>, SPEC: <dem0>)
<poss>	possessive pronoun (DET)
<refl>	reflexive ("se" PERS ACC/DAT, "si" PERS PIV)
<diff>	differentiator (DET) (e.g. *"outro", "mesmo")*
<rel>	relative pronoun (DET, SPEC)
<interr>	interrogative pronoun (DET, SPEC)
<post-det>	typically located as post-determiner (DET @N<)
<post-attr>	typically post-positioned adjective (ADJ @N<)
<ante-attr>	typically pre-positioned adjective (ADJ @>N)
<pre-attr>	obligatorily pre-positioned adjective (ADJ @>N, e.g. *“meio”*)
<adv>	can be used adverbially (ADJ @ADVL)
<post-adv>	adverb occurring in post-nominal position (@N<, @A<, e.g. *“demais”, “lá”*)
<KOMP> <igual>	"equalling" comparative (ADJ, ADV) (e.g. *"tanto", "tão")*
<KOMP> <corr>	correlating comparative (ADJ, ADV) (e.g. *"mais velho", "melhor")*
<komp> <igual>	"equalling" particle referring to comparative (e.g. *"como", "quanto")*
<komp> <corr>	"correlating" particle referring to comparative (e.g. *"do=que")*
<quant>	intensity adverb (e.g. *"muito")*
<setop>	operational adverb (e.g. *"não", "nunca", "já", "mais"* in *"não mais")*
<dei>	discourse deictics (e.g. *"aqui", "ontem")*
<ks>	conjunctional adverbs (e.g. "pois")
<prp>	prepositional adverbs (e.g. *“conforme”, “segundo”, “como”)*
<card>	cardinal (NUM)
<NUM-ord>	ordinal (ADJ)
<NUM-fract>	fraction-numeral (N)

Textual meta-information

<cif>	cipher (<card> NUM, <NUM-ord> ADJ)
<sam->	first part of morphologically fused word pair ("de" in "dele")
<-sam>	last part of morphologically fused word pair ("ele" in "dele")
<*>	1. letter capitalised
<*1><*2>	left and right parts of quotation mark bracket
<hyfen>	hyphenated word
<ABBR>	abbreviation

SEMANTIC TAGS

The tables below provide definitions (third column) and examples (last column) for all semantic noun prototypes (first column) used in the - ongoing - work described in chapter 6. The figures in column 2 refer to the number of different lexicon entries that bear a given feature. An asterisk after that number indicates that the category in question is an umbrella category to be phased out in favour of more specific (sub)categories.

ANIMATE HUMAN

H	2144*	+HUM-noun, human being -> cf. more specific categories below	*achacador 'kidnapper', acionário 'shareholder', inimigo 'enemy'*
HM	223	mystical or religious entity, constellations in astrology	*anjo 'angel', duende 'goblin', hidra 'Hydra', tauro 'Taurus'*
N	341	national -> cp. ADJ <n>, N <ling>	*cigano 'gypsy', escandinavo 'Scandinavian'*
prof	1333	professional	***-or:** escritor 'author', filósofo 'philosopher'*
fam	94	family member	*pai 'father', mãe 'mother'*
title	273	+HUM, often governing name (a) regular title (b) others, apart from <prof> or <fam>	*(a) rei 'king', presidente 'president', senhor 'mister', (b) moça 'girl', colega 'collegue'*
+n	31	-HUM, potentially governing name	*restaurante 'restaurant', plano 'plan', rua 'street'*
attr	580	attributive +HUM noun, often used as @N<	*comunista 'communist'*
HH	507	group of H	*companhia 'company', equipe 'team'*
parti	20	(political) party	*PT 'Labour'*
inst	498	institution [also topological]	*igreja 'the church', polícia 'the police', **-ria:** padaria 'bakery'*
h	2372	+HUM-adjective	*jubiloso 'jubilant', louco 'crazy'*
n	4308	nationality-adjective	*dinamarquês 'Danish'*

ANIMATE NON-HUMAN, MOVING

A	70*	+ANIM, -HUM	
AM	*not yet*	mythological animal	*pégaso 'Pegasus', licorne 'unicorn'*
AB	40	bacteria, cells	*macrófago 'macrophage'*
zo	705	animal, including mammals	*aligátor 'alligator', tênia 'tape-worm', babuíno 'baboon'*
D	278	mammal, especially domestic	*cavalo 'horse', vaca 'cow'*
orn	504	bird	*bem-te-vi 'Pitangus-bird', canário 'canary'*
ent	171	insect	*saúva 'ant', formiga 'ant'*

ich	222	fish	*perca ‘perch’, lobo-marinho ‘sea-lion’*
AA	81	group of animals (experimental: 4 AAorn, 5 DD)	*manada ‘herd’, matilha ‘pack’, cria ‘brood’, vacada ‘cow-herd’*

ANIMATE NON-HUMAN, NON-MOVING

bo	1307	plant (general)	*madressilva 'honeysuckle'*
B	132	tree or bush *(under, near, in)*	*macieira 'apple tree'*
BB	218	group of plants (cp. topologica)	*faial 'beech forest'*

TOPOLOGICALS (mostly, CONCRETA, NON-MOVABLE)

top[258]	1115	toponym or natural topological	*Brasília, monte 'mountain'*
BB	218	group of B, place as defined by what grows on it	*floresta 'forest', roça 'field', caniçal 'cane thicket'*
agua	218	body of water (where one swims)	*rio 'river', ma 'sea', laguna 'lagoon'*
sky	not yet	sky, space, air space (where one flies)	*céu 'sky'*
vej	196	path, road (where one walks or drives)	*rua 'street'*
topabs	439	abstract topological	*fim 'end', curvatura 'curvature'*
spids	22	point, tip	*farpa 'barb', pico 'sharp point, peak', cume 'peak'*
area	39	area, region	*área 'area', terreno 'area'*
hul	225	hole or cavity, notch or groove	*poro 'pore', valeta 'gutter', rasgão 'tear, gash'*
ejo	714	functional place	*quarto 'room', banheiro 'bathroom',* ***-douro***
hus	264	building [also cc]	*casa 'house', casebre 'hut', torre 'tower'*
by	88	group of houses, town, administrative unit country, state	*vila 'village', cidade 'town', estado 'state'*
inst	498	institution [also animate human]	*igreja 'church', polícia 'police',* **-ria:** *padaria 'bakery'*
ta	466	arquitectural feature	*trave 'beam', janela 'window'*
tm	348	piece of furniture [also cc]	*cadeira 'chair', mesa 'table'*
fælde	28	trap, snare	*nassa 'wicker basket', ratoeira 'mouse trap'*
kovr	150	blanket, carpet, curtain, cover, lid (what things can be *under*)	*tampa 'lid', manta*[259] *'blanket'*
ujo	663	container	*copo 'cup', garrafa 'bottle'*
rør	96	tube	*tubo 'tube', oleoduto 'pipeline'*
bild	125	picture	*pintura 'picture', grafiti 'graffiti'*
r	406	things you can read and touch[260]	*livre 'book', jornal 'newspaper'*
bar	34	fence or hedge, dike, dam (s.th. you pass over)	*fronteira 'border'*
dir	81	direction	*lés-sueste 'eastsoutheast'*
stil	77	position (you hold)	*presidência 'presidency'*

[258] The <top> feature is used not only for common nouns, but also in connection with names (e.g. *Brasília*).

[259] 'manta' is polysemic, it can also mean a manta fish <ich>, a scarf <tøj>, a furrow in agriculture <hul> and a saddle cloth <tøjzo>

[260] All r (touchable readables) are also rr (readables), but not vice versa, as with *poema 'poem'*.

sit	490	situation, state of affairs *(in+)*	*caos 'chaos', circunstâncias 'circumstances'*
vejr	66	weather	*gelada 'hoar frost'*
vind	102	wind	*furacão 'hurricane'*
regn	36	rain	*chuvisco 'drizzle'*
an[261]	369	anatomical site (HUM)	*dorso 'back', entranhas 'guts'*
anmov	164	movable anatomical part -> +PL (hand, finger, head)	*sobrancelha 'eyebrow', tentáculo 'tentacle'*
anorg	137	anatomical organ -> +PL *(in+)*	*coração 'heart'*
anost	79	bone	*fêmur 'thighbone'*
anfeat	117	(uncountable) anatomical "wearable" (grimace, tan, hair) -> <anfeatc> (countable anatomical feature=	*queixo-duplo 'double chin', riso=amarelo 'forced smile'*
anzo	159	animal anatomy	*colmilho 'tusk', focinha 'snout'*
anorn	15	bird anatomy	*pluma 'feather', rostro 'beak'*
anich	9	fish anatomy	*barbatana 'fin, flipper'*
anent	13	insect anatomy	*rostro 'proboscis'*
anbo	133	plant anatomy	*drupa 'drupe', estame 'stamen'*
star	51	star	*planeta 'planet', Venus 'Venus'*
surf	3	surface (2-dimensional, *on+*)	*chão 'floor, ground', superfície 'surface', face 'front, side'*
DIST	6	distance (after <vt+DIST>)	*légua 'mile'*

CONCRETA, MOVING

V	216	vehicle	*carro 'car', bicleta 'bicycle'*
skib	161	ship	*navio 'ship', iate 'yacht'*
fly	47	plane	*teco-teco 'little plane', pára-quedas 'parachute'*
VV	6	group of vehicles	*armada 'fleet', comboio 'convoy'*
or	114	machine	*britadeira 'stone crusher'*

CONCRETA, NON-MOVING (MOVABLE)

cm	734	physical mass nouns (+CONCRETE, +MASS)	*adubo 'fertilizer', ar 'air', breu 'tar', espuma 'foam'*
liqu	181	liquid	*petróleo 'oil', saliva 'saliva'*
mat	173	material	*madeira 'tree', silicone 'silicone'*

[261] The <an>-subcategories can be divided into two groups, between which combinations are possible. The first denotes anatomical "topics" and consists of <anmov>, <anorg>, <anost> and <anfeat>, the second indicates the biological category of the "owner" of the piece of anatomy in question, and consists of <anzo>, <anorn>, <anich>, <anent> and <anbo>. *tentáculo*, for instance, translates as 'tentacle' or 'feeler', depending on whether <anmov> combines with <anzo> or <anent>. The subdistinctions of the <an>-group have not yet been subjected to major disambiguation efforts in the parser's rule set.

stof	136	fabric	*seda ‘silke’*
mad	106	food, unprocessed -> <madc> (piece of unprocessed food)	*carne ‘meat’ alho ‘garlic’*
kul	413	food, processed -> <kulc> (piece of processed food)	*chocolate ‘chocolate’*
drik	176	drink	*chope ‘draught beer’, leite ‘milk’*
rem	69	remedy (medicine)	*morfina ‘morphine’, penicilina ‘penicilline’, vitamina ‘vitamin’*

cc	412	concrete objects (+CONCRETE, -MASS)	*pedra ‘stone’*
ar[262]	128	stack, heap, pile, bundle, row	*feixa ‘bundle’, fila ‘row’*
er	44	(countable) piece or part or group member	*ingrediente ‘ingredient’, lasca ‘chip, splinter’, parte ‘part’*
sten	117	stone (you can throw, cf. <mat>)	*pedra ‘stone’, rubim ‘ruby’*
stok	91	stick, plank, board	*vara ‘rod’, galho ‘twig’*
star	51	star (cf. topologica)	*estrela ‘star’, planeta ‘planet’*
ild	49	(1) fire, spark etc. (2) pipe, bonfire (all that can be lighted) -> cf. <lys> (light tools)	*chama ‘flame’, chispa ‘spark’, fogueira ‘bonfire’, relâmpago ‘lightning’, cachimbo ‘pipe’*
vejrc	6	countable weather phenomena	*nuvem ‘cloud’*
madc	154	piece of unprocessed food, eg. fruit	*ervilha ‘pea’, cebola ‘onion’*
kulc	190	piece of processed food, eg. burger	*pão ‘bread’*
il	1428	tool	*garfo ‘fork’, plectro ‘plectron’*
kniv	119	knife, sword, spear (bundling of features in prototype: sharp, pointed, cutting, tool)	*faca ‘knife’, canivete ‘pocket knife’, enxada ‘spade’*
fio	225	thread, rope	*estrém ‘anchor cable’, fio ‘thread’, cabo ‘cable, rope’*
klud	55	piece of cloth	*guardanapo ‘napkin’, toalha ‘håndklæde’*
sejl	40	sail	*bujarrona ‘jib, gib’*
paf	30	gun	*canhão ‘canon’, pistola ‘pistol’*
lys	65	lamp, torch etc. (all that gives light) -> cf. <ild> (fire-words)	*lanterna ‘lantern’, tocha ‘torch’*
ten	37	handle, mouth-piece, hilt	*maçaneta ‘doorknob’, hastil ‘shaft [of a lance]*
mu	209	musical instrument	*violão ‘guitar’, flauta ‘flute’*
tøj	390	garment (what you wear)	*saia ‘skirt’, camisa ‘shirt’*
sko	32	shoe	*chinela ‘slipper’*
hat	51	hat	*coroa ‘crown’, chapéu ‘hat’*
smyk	26	jewels etc.	*brinco ‘earring’*
tøjzo	31	what animals wear	*brida ‘bridle’, xairel ‘saddle cloth’*

ABSTRACTA

am	2154	quantifiable feature, abstract mass nouns (-CONCRETE, +MASS)	***-eza, -idade***
amh	not yet	+HUM quantifiable feature (ele tem mais ...ade)	

ac	623	abstract countables	*método ‘method’, módulo*

[262] The <ar> and <er> categories are semantically “transparent”, i.e. np’s headed by such words can acquire semantic features contributed by postnominal pp’s introduced by *de* ‘of’. The abstract sister-category is <meta>, a subcategory of <akc>, which is countable like <ar> and <er>, but can’t be touched.

		(-CONCRETE, - MASS, +COUNT)	*‘module’, onda ‘wave’*
featc	81	countable features (mark, spot)	*cor ‘colour’, listra ‘stripe’*
anfeatc	85	countable anatomical feature	*verruga ‘wart’, cabelo ‘hair’*
sygc	128	countable item of disease (boil, scar)	*terçolho ‘sty’, abcesso ‘abcess’*
p	115	what you think (thought)	*idéia ‘idea’, suspeita ‘suspicion’*
pp	88	plan, concept (product of thought)	*projeto ‘project’, estratégia ‘strategy’, noção ‘notion’*
reg	8	rule, law	*regra ‘rule’*
right	17	rights, habits (one has) (= reg +ADJECTIVAL)	*direito ‘right’, prerrogativa ‘privilege’*
emne	*not yet*	topic (both ac and ak)	*assunto ‘topic’*
l	288	what you hear (e.g. natural or artificial sound, noise)	*aplauso ‘applause’, berro ‘shout’* *som ‘sound’*
ll	150	song, piece of music, type of musik (product for hearing-listening)	*bossa=nova, canto ‘song’, hino ‘hymn’*
w	82	what you see (e.g. bubble, shadow, a light)	*vislumbre ‘glimpse, glimmer’, ilusão ‘illusion’, arco-iris ‘rainbow’*
ww	67	what you watch, e.g. movie, piece of theater (product for watching)	*filme ‘film’, sonho ‘dream’, novela ‘tv-series’, comédia*
s	187	what you say, short utterance (e.g. word, question, answer)	*pergunta ‘question’, salamaleque ‘salem aleikum’*
ss	196	speech, joke, lie, rumour, nonsense, gossip, boast (speech product)	*sermão ‘sermon’, testemunho ‘testemony’*
sd	260	speech act: tease, mockery, fun (you make of s.b.), insult, request (doing by saying), intrigue, proposal, settlement, appointment, ruling, judgement	*reza ‘prayer’, ultimato ‘ultimatum’, veto ‘veto’, queixa ‘complaint’*
ret	130	rhetoric terms (non-rhetoric, e.g parts of speech like ‘noun’, ‘subject’ -> **akc**)	*hipérbole ‘hyperbole’, coloquialismo ‘colloquialism’*
o	30	what you smell (odour) or taste	*bodum ‘smell of non-castrated goat’, bufa ‘fart’*
f		what you feel (whish, burst of pain, - where not **am**)	*cócegas ‘itching’, deleite ‘delight’, desejo ‘wish’*
rr	129	what you read (unlike **r**, which is **cc**: *livre, jornal)*	*romance ‘novel’, dissertação ‘dissertation’*
tegn	311	what you write (sign, character, icon, printed symbol, playing card (ace, king etc.) also: <NUM> (numbers as nouns)	*vírgula ‘comma’, dáblio ‘w’, emblema, gatafunhos ‘scribbles’, o cinco ‘number five’*
geom	102	geometric shape (circle, globe, angle) -> + **top**	*elipse ‘ellipse’, heksaedro, retângulo ‘rectangle’*
line	*not yet*	line, stripe, streak (now: **ac**)	*linha ‘line’, raio ‘ray, radius’*
d	326	what you do or make: mistake, error (test: *dar/fazer*, e.g. *dar uma lavadela*), also -> **CP**	*crime ‘crime’, espiada ‘glance’, gafe ‘blunder’*
num+	80	(syntactic, not semantic, therefore preferably in combination with <unit>,	

		<qu>, <qus>)	
unit	-	unit (always with -> **num+**)	*ano 'year', cruzeiro, acre 'acre'*
qu	166	quantity +*de*[263] (always with -> **num+**)	*montes de 'lots of', carradas de 'loads of'*
qus	189	"title"-quantity +*de* (always with -> **num+**) [in fact, a Danish MT motivated category, permitting Danish N N sequences]	*metro 'meter', litro 'liter', garaffa 'bottle'*
mon	358	quantity of money	*achádego 'finder's reward', bolsa 'scholarship'*
akc	98	countable category (rhetoric terms rather -> **ret**)	*verbo 'verb', substantivo 'noun', numeral 'numeral'*
meta	11	kind of, type of, group member (semantically transparent like **ar**, **er**)	*tipo 'type', espécie 'kind'*

[263] The <qu> category retains the preposition in the Danish translation, unlike <qus> where the Danish translation equivalents allow direct nominal valency. It remains to be shown whether the Danish distinction reflects a semantic universal (thus justifying the use of the categories even for *Portuguese*), - or just some syntactic idiosyncracy of Danish.

ax	25 *plus below:*	(-CONCRETE, -MASS, -COUNT, +FEATURE ["adjectival content"])	*antigüidade ‘antiquity’, neutralidade ‘neutrality’, simbiose ‘symbiosis’*
state	84	state (s.th. is in, distinct from -> **sit** - state of affairs)	*estupor ‘stupor’, incandescência ‘incandescence’, sossego ‘calm’*
sh	271	*human* state (e.g. health, ecstasy)	*apatie ‘apathy’, cansaço ‘fatigue’*
feat	273	non-quantifiable feature	*clima ‘climate’, enormidade ‘enormity’, textura ‘texture’*
fh	198	*human* non-quantifiable feature, human capacity, skill	*atitude ‘attitude’, calvície ‘baldness’, inocência ‘innocence’*
featq	114	quantifiable non-count non-mass (inherent) feature	*tamanho ‘size’, massa=atómica, circunferência ‘circumference’, cumprimento ‘length’*
fhq	*not yet*	*human* quantifiable feature	*tensão=arterial ‘blood pressure’*
ak	283	category	*mistura ‘mixture’, modo=maior ‘major key’, pretérito ‘past tense’*
akss	58	speech product category: nonsense, gossip, boast (similar to **ss**, but not countable)	*farelório ‘chit-chat’, galimatias ‘gibberish’, mexerico ‘gossip’*
ism	281	ideology, religion	*comunismo ‘communism’*
ling	207	language[264]	*esperanto, gíria ‘slang’, inglês ‘English’, **-ês***
syg	587	disease	*psitacose ‘parrot fever’*
col	232	coulour	*roxo ‘violet’, rubente ‘ruby-red’*

ACTIVITY-, ACTION-, PROCESS- AND EVENT-NOUNS (+V [verbality] feature)

CI	661	activity (+CONTROL, imperfective -> -PL), often derived from <vi>	*cavadela ‘digging’, circulação ‘circulation’, boicote ‘boycot’*
lud	91	game	*pôquer ‘poker’, roleta ‘roulette’*
sp	57	sport	*badminton, canoagem ‘canoeing’*
fag	319	subject (to learn), profession (to practice)	*cardiologia ‘cardiology’, ciência ‘science’, culinária ‘cookery’*
terapi	11	therapy	*acupunctura ‘acupuncture’*
dans	84	dance (also -> **ll**)	*mambo, polca, samba*
tæsk	24	beating	*sova ‘beating’, surra ‘thrashing’*

CP	2307	action (+CONTROL, perfective), often derived	*pacificação ‘pacification’, partida ‘departure’*

[264] As in English, Portuguese language names commonly have an ambiguity overlap with the noun denoting the nationality of the speaker of the language <N>, and the relatied “national” adjective <nat>.

		from <vt> (possibly, <ve>)	
CPS	*not yet*	-PL -actions	*-ação*
CPP	*not yet*	+PL -actions	*tiro 'shot'*
d	326	often, but not always deverbal, cp. -> **ac** (what you do or make)	dar uma *lavadela 'light washing'*
kneb	76	trick, cheat, fraud	*jeito 'trick', dica 'tip', engano 'cheat', intriga 'intrigue'*

cI	199	process (-CONTROL, imperfective -> -PL), often derived from <ve>	*crescimento ‘growth’, decaimento ‘decay’, decomposição ‘rotting’*

cP	696	event (-CONTROL, perfective)	*impacto ‘impact’, nascimento ‘birth’, ovulação ‘ovulation’*
cPS	*not yet*	-PL -events	*queda ‘fall’*
cPP	*not yet*	+PL -events	*explosão ‘explosion’, boléu ‘crash’, acidente ‘accident’*
snak	63	talk[265]	*debate ‘debate, bate-papo ‘little talk’, discussão ‘discussion’*
strid	45	fight, quarrel	*rixa ‘row’, briga ‘fight’*

TIME FIELD

There is a certain overlap between the time categories. *Guerra*, for instance, can be used as both a period <per>, a (historical) point in time <temp> and an occasion <occ>.

Nevertheless, a number of operational criteria makes prototype membership distinction possible: **<dur>** words are “measuring words” and take numeral premodifiers *(‘7 semanas’)*, **<occ>** is a “time-institution” and can be “taken part in” *(participar em/de* or similar verbs*)*, and **<per>** is a “time-place” and can be measured by <dur> words. **<temp>**, finally, covers time landmarks that often are governed by the “delimiter prepositions” *desde (‘from’, ‘since’)* and *até (‘until’)*, and cannot be measured by <dur> words.

temp	178	point in time [can be part of **occ**, -*V event*]	*início ‘beginning’, instante ‘moment’*
per	350	period in time [part of **cI/process**, *time-place*]	*fase ‘phase’, guerra ‘war’*
dur	28	measure of time [part of **num+**, *time-unit*]	*hora ‘hour’, semana ‘week’*
occ	514	occasion [part of **cP/event**, +HUM, *human place-event*]	*concerto ‘concert’, guerra ‘war’, Natal ‘Christmas’*

[265] <snak> and <strid> have been allocated the -CONTROL feature, because they “happen” in the sense that they are not fully controlled by the individual participant. <snak> thus means the situation of communal talking rather than a talk one gives (which is a speech product <ss>)

Appendix: PALMORF program architecture

PALMORF-preparations -> *makelistsuffix, makelistprefix*

allocates memory, establishes data-structures in RAM, assigns pointer-names.

PALMORF-main

direct input -> direct analysis
-> text file analysis
-> translation module

takes direct word or word sequence input from the key-board and sends it to direct analysis.

"file" prompts file name input, followed by automatic text file analysis (with only problematic words shown on the screen). Output is then written to a _pars file.

"trad" changes output to translation mode (in direct analysis only)

"slut" ends session

direct analysis -> inflexion analysis, -> prefix, -> direct output

analyses keyboard word input directly; unlike text file analysis, this does not make use of any major preprocessing, but is useful for fast checking of individual word forms or short sentences.

direct output

writes output to screen, each target word is followed by all its morphological readings, one per line, ordered by lexical root. If the optional "trad"-mode is on, every new root in the reading cohort is followed by a list of its syntactic word class possibilities, one per line, each followed by Danish translation equivalents. Thus syntactic word class provides a first, rough polysemy grouping.

findword

searches the lexicon for whole word items, abbreviations, polylexical items etc., called mainly by the preprocessor for fast reference

inflexion analysis

whole word search

looks the input word up in the lexicon, and, if found, stores word class and - if irregular - inflexion information for the output module. Whatever the outcome of the whole word search, the program proceeds to inflexion morpheme analysis.

inflexion morpheme analysis -> *root search*

looks for all possible inflexional ending morphemes in the input word (starting from its right hand lexical border), cuts them off, standardises the remaining trunk according to the inflexion morphemes base condition, and sends it to root search.

root search

lexicon search

takes a word trunk as input, searches for it in the lexicon, and - if positive - checks the root for "outward compatibility" (word class and combination rules) with any inflexion or suffixation element to its right. If compatible, the root is stored together with its derivational path for the output module.

If no root is found, or if is not compatible, the program tries to cut a (further) suffix of the trunk.

suffix analysis -> *root search (recursive)*

cuts a suffix off an input string, checks this suffix for "outward word class compatibility" and "inward phonologic compatibility" and - if both are positive - sends the remaining trunk in standardised or phonologically adapted form to root search, thus allowing for recursion and increasing "depth" in relation to the number of successive derivational elements involved.

If no suffix is found, or if it is not compatible, the program progresses to prefix analysis.

prefix -> *inflexion analysis*

cuts possible prefixes off the word stem, and - if phonologically compatible - sends the remaining word trunk to normal analysis, both inflexion and suffixation. Performed when none or only suffixed readings are found for a given word.

makelistsuffix

establishes pointer tree for suffix-searching

makelistprefix

establishes pointer tree for prefix-searching

preprocessor

polylexical structures *whole word analysis (<-> findword)*

looks up all word sequences of up to 4 elements length in the lexicon; if found, they are marked by '=' ligation between words. In the case of polylexicals only listed as incorporables - i.e. without another, autonomous reading - ligation is only performed if a form of the incorporating verb in question is found in the left hand context.

capitalisation

word initial capital letters are substituted by '*' + lower case letter

numbers

are marked '$'. If a string starts with a number, all of it becomes a numerical $-expression.

punctuation characters

are isolated as single characters and prefixed with '$'

abbreviations

strings ending in or containing '.', '-' or '/' are checked in the lexicon. If they do not figure there as abbreviations, the string is split up in ordinary word strings and $-prefixed punctuation marks. In ambiguous cases (e.g. sentence final abbreviations) a few simple context dependent rules are used.

hyphenation

hyphenated strings are split up, but the hyphen is retained as a suffixed marker at the end of the word originally preceding it. After individual inflexional analysis in the main program hyphenated polylexicals can thus be "reassembled" later on, and checked against the lexicon.

enclitics

as part of the hyphenation analysis pronominal enclitics are identified, isolated and morphologically standardised. If followed by inflexional elements, these are "glued" to the preceding verb (e.g. "dar-lhes-ei" -> darei- lhes)

text file analysis

next word -> *inflexion analysis (includes suffix module),* -> *prefix*

sends all non-$-strings to the main analysis module (punctuation marks, numbers etc. have been marked $ by the preprocessor)

orthographic variation*

changes oi/ou digraphs, brazilises European Portuguese spelling

accentuation errors*

removes, changes or adds accents in unanalysed words

spelling errors*

corrects a few common errors in unanalysed words, mostly ASCII problems (e.g. c -> ç, ao -> ão)

propria heuristics

assigns the PROP tag to unanalysed or heavily derivated capitalised words (restricted after full stop and by certain context sensitive rules searching for name chains and pre-name contexts.)

non-propria heuristics

assigns word class, inflexional and derivational tags by trying to do partial analyses of as large as possible a right hand chunk of any unanalysable word, recognising inflexion morphemes, suffixes and word class specific endings and attaching them to hypothesised 'xxx' roots.

local disambiguation

all but the least complex derivational readings are discarded

output

writes the remaining analyses to the _pars file, root, derivation, word class and inflexion

cohort statistics

ambiguity distribution analysis

translation,

prepares bilingual lexicon (Portuguese - Danish) in RAM (optional, not related to the parsing project). This module has now been replaced for the running text option by a full-fledged MT system featuring polysemy resolution, root translation and a bilingual syntax transformation module. The MT program can be run on top of the parser as a chained UNIX-program. Lexicon queries for individual words can be performed through a special html-form.

** orthographic intervention is used only where no analysis has been found, and the altered word forms are marked 'ALT' , so they can be identified later, for example for output statistics, and for the sake of general corpus fidelity.*

Appendix: CG-rules for proper nouns

Addendum to chapter 2.2.4.4: Context sensitive Constraint Grammar rules for the disambiguation of proper noun candidates (for an explanation of the rule formalism, cf. chapters 3.5.3 and 3.6)

A) Non-heuristic CG rules:

Choose the proper noun reading,

if the alternative is a singular noun and there is no preceding matching prenominal:

SELECT (PROP) (0 NS) (NOT -1 PRE-NS) (*1C VFIN) (NOT 0 N-UDEN-DET);

SELECT (PROP) (0 NS) (NOT -1 PRE-NS OR >>>) (NOT 0 <title> OR N-UDEN-DET) ;

if the word to the left is a pre-name word:

SELECT (PROP) (-1 <+n>) (NOT 0 ATTR OR HEAD-ORD&) ; # not: Advogado Importado

if the word before also is a name:

SELECT (PROP) (-1C PROP) (NOT 0 VFIN OR HEAD-ORD&) ; # not: Collor $Enloqueceu o Brasil

SELECT (PROP) (-1C PROP) (*1C VFIN BARRIER CLB) (NOT 0 HEAD-ORD&) ;

SELECT (PROP) (-1C PROP) (NOT 0 V3 OR ADV OR PCP OR HEAD-ORD&) ;

if the following word is a name:

SELECT (PROP) (1C PROP) (NOT 0 HEAD-ORD&);

SELECT (PROP) (1 PROP) (*2C N/A/V LINK NOT 0 <*>) (NOT 0 HEAD-ORD&) ;

SELECT (<title>) (1C PROP) ;

SELECT (<+n>) (1C PROP) ;

if the word is sentence-initial:

SELECT (PROP) (-1 >>>) (NOT 0 NP OR AP OR V) (*1 VFIN) ;

SELECT (PROP) (-1 >>>) (1 V3S) (NOT 0 NS OR AS) ;

SELECT (PROP) (-1 >>>) (1C V3S) ;

SELECT (PROP) (-1 >>>) (1 KC/VFIN/ADV) (NOT 0 SPEC OR N-UDEN-DET OR NOM OR NP) ; # Itamar só espera nomear

if the word is part of a typical "noble name chain":

SELECT (PROP) (-1C <art>) (-2 PRP-DE) (-3 PROP) ; # Tadeu do Amaral

SELECT (PROP) (-1 <artd>) (-2 PRP-DE) (-3C PROP) ;

if the word is co-ordinated with another proper noun:

SELECT (PROP) (-1 E-KC) (-2C PROP) (NOT 0 HEAD-ORD&); # Evandro Lins e $Silva

if the nearest safe content word (noun, adjective or verb) is not capitalised, and the target word is not sentence initial or post-attributive (like 'o século XXI')

SELECT (PROP) (*1C N/A/V LINK NOT 0 <*>) (NOT -1 >>>) (NOT 0 <post-attr>) ;

SELECT (PROP) (*-1C N/A/V LINK NOT 0 <*>) (NOT 0 <post-attr>) ;

if the word is preceded by another capitalised word and has another (derived) reading, as long as there are no head line words anywhere in the sentence:

SELECT (PROP) (0 DER) (-1 <*>) (**-1 ALL BARRIER HEAD-ORD LINK -1 >>>) (NOT *1 HEAD-ORD) ; # o Banco $Pactual in a typical van/von-name chain context

```
        SELECT (PROP) (-1 VAN/VON) (-2 PROP) ;
        SELECT (PROP) (1 VAN/VON) (2 PROP) ;
```

if the alternative non-PROP readings are derivated and the next neighbouring content word is lower case

```
SELECT (PROP) (0 DER) (*1C N/A/V LINK NOT 0 <*>) (NOT -1 >>>) (NOT 0 <post-attr\>) ; # not: Dias antes ...,
        not: o século XXI
        SELECT (PROP) (0 DER) (*-1C N/A/V LINK NOT 0 <*>) (NOT 0 <post-attr>) ;
```

Discard the proper noun reading,

if the word is sentence-initial and <u>not</u> followed by a co-ordinater, name, finite verb, adverb, non-name*de* or break

```
        REMOVE (PROP) (-1 >>>) (NOT 1 IT OR PROP OR VFIN OR ADV OR PRP-DE OR BREAK) ;
        REMOVE (<HEUR> PROP) (-1 >>>) (1 PRP-DE) (NOT 2 PROP) ; # not: Francisco de Melo
```

if there is a better defined (i.e. lexical) name reading for the same word:

```
        REMOVE (PROP M/F S/P) (0 PROPS) ; # Collor-$PC, Benito $Gama
```

if there is another (singular) name alternative, that has correct agreement with a safe prenominal

```
        REMOVE (PROP M S) (0 PROP-FS) (-1C DETFS) ;
        REMOVE (PROP F S) (0 PROP-MS) (-1C DETMS) ;
```

if the alternative is a transitive verb, the following word a safe determiner, and the preceding word a noun or name.

```
        REMOVE (PROP) (0 <vt>) (1C DETA/B/C) (-1C N/PROP) ; # Collor $Enloqueceu o Brasil
```

if it is a place-name competing with a person name, and is followed by another name not being a toponym:

```
        REMOVE (<top> PROP) (0 PROP-NAME) (1C PROP) (NOT 1 PROP-LOC) ;
```

if it also could be a closed class word (for instance, in a head-line or sentence-initial) or a "real" derivated proper noun

```
        REMOVE (<HEUR> PROP) (0 HEAD-ORD& OR (-inho PROP)) ;
```

After ordinary disambiguation, the remaining ambiguity is addressed by the following heuristic rules that are are applied group-wise (i.e. one heuristics level after the other):

<u>Heuristic level 1:</u>

Choose the name reading, if the target word is not placed sentence-initial:

```
SELECT (PROP) (NOT -1 >>>) (NOT 0 ATTR OR HEAD-ORD& OR P& OR N-UDEN-DET) (NOT -1 PRE-N);
```

if the alternatives have either 2 suffixes or 2 prefixes

```
        SELECT (PROP) (0 DERS2 OR DERP2) ;
```

if the alternative is a suffix-analysis with a very short (i.e. uncharacteristic and possibly wrong) suffix

```
        SELECT (PROP) (0 DER-SUFF) (NOT 0 DERS-LONG) ;
```

if the PROP reading is not heuristic, but lexicon-registered

```
        SELECT (PROP) (NOT 0 <HEUR>) ;
```

Discard a proper name reading, if it is competing with a post-attributive reading:

```
        REMOVE (PROP) (0 <post-attr>) ; # se'culo XXI
```

if the PROP reading is heuristic, and the alternative non-PROP reading is not derivated

```
        REMOVE (<HEUR> PROP) (NOT 0 DER) ;
```

if the PROP reading is heuristic, and the alternative non-PROP reading has only one suffix, being long/characteristic

```
REMOVE (<HEUR> PROP) (0 DERS-LONG) (NOT 0 DERS2) ;
```

Heuristic level 2-6:

Choose the PROP reading, if the alternative non-PROP readings are derivated (no matter with how many affixes):

```
SELECT (PROP) (0 DER) ;
```

Prefer the name reading, if the competing analysis is derivational:

```
REMOVE (<DERS) (0 PROP) ;
REMOVE (<DERP) (0 PROP) ;
```

Choose the proper noun reading anyway:

```
SELECT (<*> PROP) (0 NOMINAL) ;
```

Discard a topological name reading if the preceding word is a name itself:

```
REMOVE (<top> PROP) (-1C PROP) ;
```

Appendix: Example sentences

The following sample sentences are typical of Brazilian journalistic texts. All are quotes from 1996 editions of the *Folha de São Paulo* newspaper and the*VEJA* news magazine. The analyses given cover morphology and syntax, while secondary tags for valency potential and semantics have been filtered away (with the exception of the last sample). The purpose of the examples is to provide a coherent and more contextualised picture of the parser's morphosyntactic notation and its differentiation potential, not to demonstrate its statistical performance (which is discussed elsewhere). However, some problems, uncertainties and errors are discussed in accompanying footnotes.

Estudo	[estudo] N M S @SUBJ>
revela	[revelar] V PR 3S IND VFIN @FMV
ação	[ação] N F S @<ACC
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
álcool	[álcool] N M S @P<
sobre	[sobre] PRP **@N<**[266]
memória	[memória] N F S @P<

Pesquisas	[pesquisa] N F P @SUBJ>
norte-americanas	[americano] ADJ F P @N<
mostram	[mostrar] V PR 3P IND VFIN @FMV
que	[que] KS @SUB @#FS-<ACC
três	[três] <card> NUM M/F P @>N
doses	[dose] N F P @SUBJ>
de	[de] PRP @N<
uísque	[uísque] N M S @P<
são	[ser] V PR 3P IND VFIN @FMV
suficientes	[suficiente] <quant> ADJ M/F P @<SC
para	[para] PRP @A<
baixar	[baixar] V INF 0/1/3S @IMV @#ICL-P<
a	[a] <artd> DET F S @>N
atividade	[atividade] N F S @<ACC
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
hipocampo	[hipocampo] N M S @P<
$.	

Essa	[esse] <dem> DET F S @>N
estrutura	[estrutura] N F S @SUBJ>
de	[de] <sam-> PRP @N<
o	[o] <-sam> <artd> DET M S @>N
cérebro	[cérebro] N M S @P<

[266] The postnominal tag @N< is, in fact, ambiguous in this case. It refers to an np-head to the left, but though the parser probably selected the tag in order to match the valency potential of 'ação', the flat notation does not distinguish between attachment to ação or álcool, respectively.

```
é               [ser] V PR 3S IND VFIN  @FMV
a               [a] <artd> DET F S  @>N
responsável     [responsável] ADJ M/F S  @<SC
por             [por] <sam-> PRP  @A<
as              [a] <-sam> <artd> DET F P  @>N
memórias        [memória] N F P  @P<
complexas       [complexo] ADJ F P  @N<
e               [e]    KC @CO
por             [por] <sam-> PRP  @N<
a               [a] <-sam> <artd> DET F S  @>N
orientação      [orientação] N F S  @P<
espacial        [espacial] ADJ M/F S  @N<
e               [e]    KC @CO
temporal        [temporal] ADJ M S  @N<
$.
```

***

```
O               [o]  <artd> DET M S  @>N
presidente      [presidente] N M/F S  @SUBJ>
Fernando=Henrique=Cardoso          [Fernando=Henrique=Cardoso]  PROP M/F S/P  @N<
chega           [chegar] V PR 3S IND VFIN  @FMV
a               [a] <sam-> PRP  @<PIV[267]
a               [a] <-sam> <artd> DET F S  @>N
metade          [metade] N F S  @P<
de              [de] PRP  @N<
seu             [seu] <poss 3S/P> DET M S  @>N
mandato         [mandato] N M S  @P<
em              [em] PRP  @<ADVL
plena           [pleno] ADJ F S  @>N
lua-de-mel      [lua-de-mel] N F S  @P<
com             [com] PRP  @<ADVL
a               [a] <artd> DET F S  @>N
opinião=pública         [opinião=pública] N F S  @P<
$.

Os              [o]  <artd> DET M P  @>N
desafios        [desafio] N M P  @SUBJ>
por             [por] <sam-> PRP  @N<
a               [a] <-sam> <artd> DET F S  @>N
frente          [frente] N F S  @P<
são             [ser] V PR 3P IND VFIN  @FMV
enormes         [enorme] ADJ M/F P  @<SC
$,
o               [o] <artd> DET M S  @>N
que             [que] <rel> SPEC M/F S/P  @ACC>>[268] @#FS-S<
```

[267] An alternativ tag for the valency bound pp would be @<ADV. In fact, ‘chegar’ is lexicon-marked for both <va+LOC>, <a^vp> and <a^xp> valency potentials, but since the pp here doesn’t qualify as “locative”, the @<ADV reading is discarded.

[268] This tag is very rare. It is used where the direct object in question relates to the main verb in a subclause that itself is a dependent of the next main verb to the right. The most common appearance is in ACI constructions of the type: *Ela se*

```
é                [ser] V PR 3S IND VFIN  @FMV
quase            [quase] <quant> <det> ADV  @>A
dispensável      [dispensável] ADJ M/F S  @<SC
dizer            [dizer] V INF 0/1/3S  @IMV @#ICL-<SUBJ
$.

Mas              [mas]    KC @CO
a                [a] <artd> DET F S  @>N
inflação         [inflação] N F S  @SUBJ>
baixa            [baixo] ADJ F S  @N<
funciona         [funcionar] V PR 3S IND VFIN  @FMV
como             [como] <rel> <prp> ADV  @COM @#AS-<ADVL
inestimável      [inestimável] ADJ M/F S  @>N
alavanca         [alavanca] N F S  @AS<
eleitoral        [eleitoral] ADJ M/F S  @N<
$,
o                [o] <artd> DET M S  @>N
que              [que] <rel> SPEC M/F S/P  @SUBJ> @#FS-S<
já               [já] ADV  @ADVL>
se               [se] <refl> PERS M/F 3S/P ACC/DAT  @ACC>-PASS
comprovara       [comprovar] V MQP 1/3S IND VFIN  @FMV
antes            [antes] ADV  @<ADVL
$,
em               [em] <sam-> PRP  @<ADVL
a                [a] <-sam> <artd> DET F S  @>N
Argentina        [Argentina]  PROP F S  @P<
e                [e]    KC @CO
em               [em] <sam-> PRP  @<ADVL
o                [o] <-sam> <artd> DET M S  @>N
Peru             [Peru]  PROP M S  @P<
$,
com              [com] PRP  @<PRED[269]
a                [a] <artd> DET F S  @>N
reeleição        [reeleição] N F S  @P<
de               [de] PRP  @N<
Carlos=Menem     [Carlos=Menem]  PROP M/F S/P  @P<
e                [e]    KC @CO
Alberto=Fujimori       [Alberto=Fujimori]  PROP M/F S/P  @P<
$,
respectivamente  [respectivamente] ADV  @<ADVL
$.

Em               [em] <sam-> PRP  @ADVL>
os               [o] <-sam> <artd> DET M P  @>N
dois             [dois] <card> NUM M/F P  @>N
casos            [caso] N M P  @P<
```

deixou levar. *Ele se fez eleger presidente*. Without a very specific mapping rule, “o=que” in *..., o que é quase dispensável dizer* would be - wrongly - tagged @SUBJ>, as in the following sentence.

[269] The function tag should probably be @<ADVL. Adverbial and predicative clause level adjuncts are usually kept apart by their *form*, the former being adverbs or pp’s, the latter adjectives or adjective phrases. Pp’s headed by ‘com’, however, appear in both functional classes.

$,	
domar	[domar] V INF 0/1/3S @IMV @#ICL-SUBJ>
a	[a] <artd> DET F S @>N
inflação	[inflação] N F S @<ACC
foi	[ser] V PS 3S IND VFIN @FMV
o	[o] <artd> DET M S @>N
principal	[principal] <SUP> ADJ M/F S @>N
fator	[fator] N M S @<SC
que	[que] <rel> SPEC M/F S/P @SUBJ> @#FS-N<
permitiu	[permitir] V PS 3S IND VFIN @FMV
um	[um] <card> NUM M S @>N
segundo	[segundo] <NUM-ord> ADJ M S @>N
mandato	[mandato] N M S @<ACC
$,	
ainda=que	[ainda=que] KS @SUB @#AS-<ADVL
$,	
em	[em] <sam-> PRP @ADVL>
o	[o] <-sam> <artd> DET M S @>N
Peru	[Peru] PROP M S @P<
$,	
o	[o] <artd> DET M S @>N
combate	[combate] N M S @SUBJ>
até=então	[até=então] ADV @ADVL>
aparentemente	[aparente] ADV @ADVL> @>A
bem-sucedido	[bem-sucedido] ADJ M S @N<
a	[a] <sam-> PRP @N<
a	[a] <-sam> <artd> DET F S @>N
luta	[luta] N F S @P<
armada	[armar] V PCP F S @N<
também	[também] ADV @ADVL>
tenha	[ter] V PR 1/3S SUBJ VFIN @FAUX
ajudado	[ajudar] V PCP M S @IMV @#ICL-AUX<
$.	

Há	[haver] V PR 3S IND VFIN @FMV
motivos	[motivo] N M P @<ACC
de=sobra	[de=sobra] PP **@<ADVL**[270]
para	[para] PRP @ADVL>
acreditar	[acreditar] V INF 0/1/3S @IMV @#ICL-P<
que	[que] KS @SUB @#FS-<ACC
$,	
se	[se] KS @SUB
aprovada	[aprovar] V PCP F S @#ICL-ADVL>
a	[a] <artd> DET F S @>N
reeleição	[reeleição] N F S @<SUBJ
$,	
o	[o] <artd> DET M S @>N
presidente	[presidente] N M/F S @SUBJ>

[270] The PP @<ADVL *de=sobra* might also be read as @N<. In the first case the meaning would be something like 'motives exist in abundance', in the second 'plenty of motives exist'.

```
será            [ser] V FUT 3S IND VFIN  @FMV
um              [um] <arti> DET M S  @>N
candidato       [candidato] N M S  @<SC
fortíssimo      [forte] ADJ M S  @N<
$.

Mas             [mas]    KC @CO
a               [a] <artd> DET F S  @>N
comparação      [comparação] N F S  @SUBJ>
com             [com] PRP  @N<
Menem           [Menem]  PROP M/F S/P  @P<
e               [e]    KC @CO
Fujimori        [Fujimori]  PROP M/F S/P  @P<
deve            [dever] V PR 3S IND VFIN  @FAUX
funcionar       [funcionar] V INF 0/1/3S  @IMV @#ICL-AUX<
como            [como] <rel> <prp> ADV  @COM @#AS-<ADVL
sinal           [sinal] N M S  @AS<
amarelo         [amarelo] ADJ M S  @N<
$:
depois          [depois] ADV  @ADVL>
de              [de] PRP  @A<
reeleitos[271]  [eleger] V PCP M P  @P<
$,
ambos           [ambos] <quant> DET M P  @SUBJ>
viram           [ver] V PS/MQP 3P IND VFIN  @FMV
a               [a] <artd> DET F S  @>N
popularidade    [popularidade] N F S  @SUBJ>
despencar       [despencar] V INF 0/1/3S  @IMV @#ICL-<ACC
$,
apesar=de       [apesar=de] PRP  @ADVL
a               [a] <artd> DET F S  @>N
inflação        [inflação] N F S  @SUBJ>
permanecer      [permanecer] V INF 0/1/3S  @IMV @#ICL-P<
baixa           [baixo] ADJ F S  @<SC
$.

O               [o]  <artd> DET M S  @>N
que             [que] <rel> SPEC M/F S/P  @SUBJ> @#FS-SUBJ>
só              [só] ADV  @ADVL>
reforça         [reforçar] V PR 3S IND VFIN  @FMV
a               [a] <artd> DET F S  @>N
análise         [análise] N F S  @<ACC
de              [de] <sam-> PRP  @N<
o               [o] <-sam> <artd> DET M S  @>N
próprio         [próprio] DET M S  @>N
FHC             [FHC]  PROP M/F S/P  @P<
```

[271] One might argue, that *reeleitos* is a contracted clause *(depois de ser reeleitos)*. However, since the governing verb *ser* is missing, only the participle itself remains for tagging, the logical tag being @IMV @#ICL-P<, - which would create the problem of a 1-word-participle clause, not otherwise accepted in the parser's grammar. Postnominal participle clauses, for instance, only are analysed as regular clauses when integrating other constituents, like adverbials or objects, and even this is only partially expressed by means of @A<ARG tags.

```
de              [de] PRP  @<N
que             [que] KS  @SUB @#FS-N<
governar        [governar] V FUT 1/3S SUBJ VFIN  @IMV @#ICL-SUBJ>
não             [não] ADV  @ADVL>
é               [ser] V PR 3S IND VFIN  @FMV
só              [só] ADV  @<ADVL
estabilizar     [estabilizar] V INF 0/1/3S  @IMV @#ICL-<SC
a               [a] <artd> DET F S  @>N
economia        [economia] N F S  @<ACC
$.
```

***

```
Ponderou        [ponderar]  V PS 3S IND VFIN  @FMV
que             [que] KS  @SUB @#FS-<ACC
$,
em              [em] <sam-> PRP  @ADVL>
uma             [um] <-sam> <arti> DET F S  @>N
entrevista      [entrevista] N F S  @P<
concedida       [conceder] V PCP F S  @N<
a               [a] <sam-> PRP  @A<PIV
o               [o] <-sam> <artd> DET M S  @>N
Jornal=Nacional         [Jornal=Nacional]  PROP M/F S/P  @P<
em              [em] <sam-> PRP  @A<ADVL
a               [a] <-sam> <artd> DET F S  @>N
noite           [noite] N F S  @P<
de              [de] PRP  @N<
sexta-feira     [sexta-feira] N F S  @P<
$,
o               [o] <artd> DET M S  @>N
próprio         [próprio] DET M S  @>N
presidente      [presidente] N M/F S  @SUBJ>
de              [de] <sam-> PRP  @N<
o               [o] <-sam> <artd> DET M S  @>N
STF             [STF]  PROP M S  @P<
$,
Sydney=Sanches          [Sydney=Sanches]  PROP M/F S/P  @APP
$,
admitiu         [admitir] V PS 3S IND VFIN  @FMV
que             [que] KS  @SUB @#FS-<ACC
$,
caso            [caso] KS  @SUB @#FS-ADVL>
o               [o] <artd> DET M S  @>N
presidente      [presidente] N M/F S  @SUBJ>
ou              [ou]    KC @CO
seus            [seu] <poss 3S/P> DET M P  @>N
advogados       [advogado] N M P  @SUBJ>
não             [não] ADV  @ADVL>
aparecerem      [aparecer] V FUT 3P SUBJ VFIN  @FMV
em              [em] <sam-> PRP  @<ADVL
o               [o] <-sam> <artd> DET M S  @>N
```

```
julgamento       [julgamento] N M S  @P<
$,
a                [a] <artd> DET F S  @>N
sessão           [sessão] N F S  @SUBJ>
poderia          [poder] V COND 1/3S VFIN  @FAUX
ser              [ser] V INF 0/1/3S  @IAUX @#ICL-AUX<
adiada           [adiar] V PCP F S  @IMV @#ICL-AUX<
$.
Consumada        [consumar]  V PCP F S  @IMV @#ICL-ADVL>
a                [a] <artd> DET F S  @>N
fraude           [fraude] N F S  @<SUBJ
$,
Sydney=Sanches         [Sydney=Sanches]  PROP M/F S/P  @SUBJ> @APP
$,
Ibsen=Pinheiro   [Ibsen=Pinheiro]  PROP M/F S/P  @SUBJ> @<ACC
e                [e]    KC @CO
Mauro=Benevides        [Mauro=Benevides]  PROP M/F S/P  @SUBJ> @<ACC
se               [se] <refl> PERS M/F 3S/P ACC/DAT  @ACC>
reuniram         [reunir] V PS/MQP 3P IND VFIN  @FMV
para             [para] PRP  @<ADVL
marcar           [marcar] V INF 0/1/3S  @IMV @#ICL-P<
o                [o] <artd> DET M S  @>N
novo             [novo] ADJ M S  @>N
julgamento       [julgamento] N M S  @<ACC
$.

Em               [em]  <sam-> PRP  @ADVL>
a                [a] <-sam> <artd> DET F S  @>N
vida             [vida] N F S  @P<
real             [real] ADJ M/F S  @N<
$,
a                [a] <artd> DET F S  @>N
secretária       [secretária] N F S  @SUBJ>
Sandra=Fernandes=de=Oliveira        [Sandra=Fernandes=de=Oliveira]  PROP M/F S/P  @N<
disse            [dizer] V PS 1/3S IND VFIN  @FMV
que              [que] KS  @SUB @#FS-<ACC
o                [o] <artd> DET M S  @>N
contrato         [contrato] N M S  @SUBJ>
de               [de] <sam-> PRP  @N<
o                [o] <-sam> <artd> DET M S  @>N
empréstimo       [empréstimo] N M S  @P<
de               [de] <sam-> PRP  @N<
os               [o] <-sam> <artd> DET M P  @>N
dólares          [dólar] N M P  @P<
foi              [ser] V PS 3S IND VFIN  @FAUX
forjado          [forjar] V PCP M S  @IMV @#ICL-AUX<
em               [em] <sam-> PRP  @<ADVL
o                [o] <-sam> <artd> DET M S  @>N
escritório       [escritório] N M S  @P<
de               [de] PRP  @N<
seu              [seu] <poss 3S/P> DET M S  @>N
```

```
ex-patrão          [patrão] N M S  @P<
Alcides=Diniz      [Alcides=Diniz]  PROP M/F S/P  @N<
$,
o                  [o] <artd> DET M S  @>N
Cidão              [Cidão]  <*1> <*2> PROP M/F S/P  @APP
$,
Ex-pão=de=Açúcar     [Pão=de=Açúcar]  PROP M/F S/P  @APP
$,
que                [que] <rel> SPEC M/F S/P  @SUBJ> @#FS-N<
animou             [animar] V PS 3S IND VFIN  @FMV
o                  [o] <artd> DET M S  @>N
réveillon ALT reveillon          [reveillon] N M S  @<ACC
collorido          [collorido] ADJ M S  @N<
de                 [de] PRP  @N<
1991               [1991] <card> NUM M/F P @P<
em                 [em] PRP  @<ADVL
companhia          [companhia] N F S  @P<
de                 [de] <sam-> PRP  @N<
a                  [a] <-sam> <artd> DET F S  @>N
mulher             [mulher] N F S  @P<
$,
Renata=Scarpa      [Renata=Scarpa]  PROP M/F S/P  @APP
$.
```

***

```
Jesus              [Jesus]  PROP M S  @SUBJ>
e                  [e]    KC @CO
seus               [seu] <poss 3S/P> DET M P  @>N
amigos             [amigo] N M P  @SUBJ>
perambulavam       [perambular] V IMPF 3P IND VFIN  @FMV
por                [por] <sam-> PRP  @<ADVL
a                  [a] <-sam> <artd> DET F S  @>N
Galiléia           [Galiléia]  PROP M/F S/P  @P<
oferecendo         [oferecer] V GER  @IMV @#ICL-<ADVL
milagres           [milagre] N M P  @<ACC
e                  [e]    KC @CO
exigindo           [exigir] V GER  @IMV @#ICL-<ADVL
em=troca           [em=troca] PP  @<ADVL
sentar-            [sentar] V INF 0/1/3S  @IMV @#ICL-<ACC
se                 [se] <refl> PERS M/F 3S/P ACC/DAT  @<ACC
a                  [a] <sam-> PRP  @<ADVL
a                  [a] <-sam> <artd> DET F S  @>N
mesma              [mesmo] <diff> <KOMP> DET F S  @>N
mesa               [mesa] N F S  @P<
que                [que] KS  @COM @#AS-KOMP<
seus               [seu] <poss 3S/P> DET M P  @>N
interlocutores     [interlocutor] N M P  @AS<
$.
```

***

```
Os              [o]  <artd> DET M P  @>N
sabonetes       [sabonete] N M P  @SUBJ>
não             [não] ADV  @ADVL>
são             [ser] V PR 3P IND VFIN  @FAUX
perfumados      [perfumar] V PCP M P  @IMV @#ICL-AUX<
apenas          [apenas] ADV  @>P
para            [para] PRP  @<ADVL
estimular       [estimular] V INF 0/1/3S  @IMV @#ICL-P<
a               [a] <artd> DET F S  @>N
sensação        [sensação] N F S  @<ACC
de              [de] PRP  @N<
bem-estar       [bem-estar] N M S  @P<
$.

Sem             [sem]  PRP  @PRED> @ADVL>
a               [a] <artd> DET F S  @>N
fragrância      [fragrância] N F S  @P<
$,
teriam          [ter] V COND 3P VFIN  @FMV
um              [um] <arti> DET M S  @>N
odor            [odor] N M S  @<ACC
desagradável    [desagradável] ADJ M/F S  @N<
de              [de] PRP  @N<
gordura         [gordura] N F S  @P<
$.

A               [a]  <artd> DET F S  @>N
Givaudan        [Givaudan]  PROP M/F S/P  @SUBJ>
tem             [ter] V PR 3S IND VFIN  @FMV
um              [um] <arti> DET M S  @>N
bom             [bom] ADJ M S  @>N
exemplo         [exemplo] N M S  @<ACC
de              [de] PRP  @N<
até             [até] PRP  @ADV>
onde            [onde] <interr> ADV  @P< @#FS-P<
se              [se] <refl> PERS M/F 3S/P ACC/DAT  @SUBJ>
pode            [poder] V PR 3S IND VFIN  @FAUX
chegar          [chegar] V INF 0/1/3S  @IMV @#ICL-AUX<
$.

Seus            [seu]  <poss 3S/P> DET M P  @>N
técnicos        [técnico] N M P  @SUBJ>
conseguiram     [conseguir] V PS/MQP 3P IND VFIN  @FMV
reproduzir      [reproduzir] V INF 0/1/3S  @IMV @#ICL-<ACC
o               [o] <artd> DET M S  @>N
cheiro          [cheiro] N M S  @<ACC
característico  [característico] ADJ M S  @N<
de              [de] <sam-> PRP  @N<
o               [o] <-sam> <artd> DET M S  @>N
carro           [carro] N M S  @P<
```

novo	[novo] ADJ M S @N<
$.	

Atenção	[atenção] N F S @NPHR
$:	
não	[não] ADV @ADVL>
é	[ser] V PR 3S IND VFIN @FMV
cheiro	[cheiro] N M S @<SC
de	[de] PRP @N<
carro	[carro] N M S @P<
limpo	[limpar] V PCP M S @N<
$,	
é	[ser] V PR 3S IND VFIN @FMV
de	[de] PRP @<SC
automóvel	[automóvel] N M S @P<
zerinho	[zerinho] ADJ M S @N<
mesmo	[mesmo] <quant> ADV @<ADVL
$.	

***

O	[o] <artd> DET M S @>N
melhor	[bom] <KOMP> <SUP> ADJ M/F S @SUBJ>
de	[de] <sam-> PRP @N<
a	[a] <-sam> <artd> DET F S @>N
festa	[festa] N F S @P<
é	[ser] V PR 3S IND VFIN @FMV
que	[que] KS @SUB @#FS-<SC
os	[o] <artd> DET M P @>N
CDs	[CDs] PROP M/F S/P @SUBJ>
não	[não] ADV @ADVL>
apresentam	[apresentar] V PR 3P IND VFIN @FMV
os	[o] <artd> DET M P @>N
ruídos	[ruído] N M P @<ACC
e	[e] KC @CO
chiados	[chiar] V **PCP**[272] M P @<ACC
típicos	[típico] ADJ M P @N<
de	[de] <sam-> PRP @N<
as	[a] <-sam> <artd> DET F P @>N
gravações	[gravação] N F P @P<
antigas	[antigo] ADJ F P @N<
$.	

Em	[em] <sam-> PRP @ADVL>
a	[a] <-sam> <artd> DET F S @>N
maior	[grande] <KOMP> ADJ M/F S @>N

[272] *Chiado* ('creak') is not listes as N in the lexicon which is why ordinary PCP derivation is the only option. The association between participles and adjectives is, however, much more common than that between participle and noun, and words that can be both participle and adjective (in dictionary traditions) therefore keep the primary word class of participle: <ADJ> PCP.

```
parte          [parte] N F S  @P<
de             [de] <sam-> PRP  @N<
o              [o] <-sam> <artd> DET M S  @>N
material       [material] N M S  @P<
$,
de             [de] <sam-> PRP  @ADVL>
o              [o] <-sam> <artd> DET M S  @>N
ponto=de=vista [ponto=de=vista] N M S  @P<
técnico        [técnico] ADJ M S  @N<
$,
é              [ser] V PR 3S IND VFIN  @FMV
como=se        [como=se] KS  @SUB @#FS-<SC
Billie         [Billie]  PROP M/F S/P  @SUBJ>
tivesse        [ter] V IMPF 1/3S SUBJ VFIN  @FAUX
gravado        [gravar] V PCP M S  @IMV @#ICL-AUX<
com            [com] PRP  @<ADVL
os             [o] <artd> DET M P  @>N
recursos       [recurso] N M P  @P<
atuais         [atual] ADJ M/F P  @N<
$.

Coisas         [coisa]  N F P  @NPHR
de             [de] <sam-> PRP  @N<
a              [a] <-sam> <artd> DET F S  @>N
tecnologia     [tecnologia] N F S  @P<
de             [de] <sam-> PRP  @N<
o              [o] <-sam> <artd> DET M S  @>N
CD             [CD]  N M S  @P<
$.
```

***

```
Personagens    [personagem]  N M/F P  @SUBJ>
desaparecem    [desaparecer] V PR 3P IND VFIN  @FMV
sem            [sem] PRP  @<ADVL
deixar         [deixar] V INF 0/1/3S  @IMV @#ICL-P<
vestígio       [vestígio] N M S  @<ACC
$-
nem=mesmo      [nem=mesmo] ADV  @>N
um             [um] <arti> DET M S  @>N
filete         [filete] N M S  @S<[273]
de             [de] PRP  @N<
sangue         [sangue] N M S  @P<
$.
```

[273] The @S< tag ("sentence apposition") may seem less than perfect in this case, since *nem mesmo um filete de sangue* clearly refers to the direct object of the main sentence body, suggesting @APP - or, even better, @PRED function (since the latter would be applicable even in the case of a plural np referring to the *subject*). However, the distinction is very hard to make automatically. Consider, for instance, *Personagens deasaparecem sem deixar vestígio - coisa estranha e perigosa*. Possibly, the intervention of a "prenominal" operator adverb *(nem=mesmo)* could help select the @PRED reading.

```
Outros          [outro]  <diff> <KOMP> DET M P  @SUBJ>
caem            [cair] V PR 3P IND VFIN  @FMV
em              [em] PRP  @<ADV
armadilhas      [armadilha] N F P  @P<
que             [que] <rel> SPEC M/F S/P  @SUBJ> @#FS-N<
não             [não] ADV  @ADVL>
enganariam      [enganar] V COND 3P VFIN  @FMV
uma             [um] <arti> DET F S  @>N
criança         [criança] N F S  @<ACC
de              [de] PRP  @N<
olhos           [olho] N M P  @P<
vendados        [vendar] V PCP M P  @N<
$.

Mas             [mas]    KC @CO
é               [ser] V PR 3S IND VFIN  @FMV
inútil          [inútil] ADJ M/F S  @<SC
sentar-         [sentar] V INF 0/1/3S  @IMV @#ICL-<SUBJ
se              [se] <refl> PERS M/F 3S/P ACC/DAT  @<ACC
diante=de       [diante=de] PRP  @<ADVL
Drácula         [drácula]  N M S  @P<
a               [a] <sam-> PRP  @<ADVL
a               [a] <-sam> <artd> DET F S  @>N
espera          [espera] N F S  @P<
de              [de] <sam-> PRP  @N<
os              [o] <-sam> <artd> DET M P  @>N
fundamentos     [fundamento] N M P  @P<
de              [de] <sam-> PRP  @N<
o               [o] <-sam> <artd> DET M S  @>N
bom             [bom] ADJ M S  @>N
cinema          [cinema] N M S  @P<
$,
de              [de] PRP  @N<
cenas           [cena] N F P  @P<
que             [que] <rel> SPEC M/F S/P  @SUBJ> @#FS-N<
crescem         [crescer] V PR 3P IND VFIN  @FMV
em              [em] <sam-> PRP  @<ADVL
a               [a] <-sam> <artd> DET F S  @>N
imaginação      [imaginação] N F S  @P<
ou              [ou]    KC @CO
de              [de] PRP  @<ADVL[274]
personagens     [personagem] N M/F P  @P<
```

[274] The correct tag is @N<, not @<ADVL, but the parser has difficulties in making the valency link to *a espera de*, which is a rather far context for np-scope, - and even "isolated" by a relative clause. Rather, *de personagens inesquecíveis* is read as an adjunct adverbial *within* the relative clause, in the same way as the other, "true" adverbial, *na imaginação*. Note that the second valency-bound 'de' *(de cenas que ...)* in the three-part co-ordination still receives the correct @N< tag, since the closer head context makes it easier for the CG rules to establish the valency link. The problem *could* be solved by ordinary CG-rules, of course, but the necessary SELECT rule or REMOVE rule set would be fairly long and feature quite complex context conditions, imposing practical limits upon further grammar growth. In other words, the overall recall gain achieved may not justify the man-hours necessary for creating such rules for a multitude of individually rare cases.

```
inesquecíveis   [inesquecível] ADJ M/F P  @N<
$.
```

In the analysis of the last example sentence, finally, valency tags have been instantiated, and semantic tags retained:

```
Ficar           [ficar] <vK> <v-cog> V INF 0/1/3S @IMV @#ICL-SUBJ> ‘[to] become’
sem             [sem] PRP @<SC ‘without’
trabalho        [trabalho] <d> <am> N M S @P< ‘work’
é               [ser] <vK> V PR 3S IND VFIN @FMV ‘is’
ruim            [ruim] <+para> ADJ M/F S @<SC ‘bad’
para            [para] <+hum> <move+> PRP @A< ‘for’
qualquer        [qualquer] <quant2> DET M/F S @>N ‘any’
pessoa          [pessoa] <H> N F S @P< 'person'
,
mas             [mas] KC @CO ‘but’
no=caso=de      [no=caso=de] <c> PRP @ADVL> ‘in the case of’
um              [um] <quant2> <arti> DET M S @>N ‘an’
executivo       [executivo] <prof> <HH> N M S @P< ‘executive’
a               [a] <art> DET F S @>N ‘the’
demissão        [demissão] N F S @SUBJ> ‘lay-off’
vem             [vir] <vi> V PR 3S IND VFIN @FMV ‘comes’
acompanhada     [acompanhar] V PCP F S @<PRED ‘accompanied’
de              [de] PRP @A<ADVL @<ADVL ‘by’
uma             [um] <quant2> <arti> DET F S @>N ‘a’
série           [série] N F S @P< ‘series’
de              [de] PRP @N< ‘of’
mudanças        [mudança] <cP> N F P @P< ‘changes’
que             [que] <rel> SPEC M/F S/P @SUBJ> @#FS-N< ‘that’
muitas=vezes    [muitas=vezes] ADV @ADVL> 'often'
acabam          [acabar] <x+GER> V PR 3P IND VFIN @FAUX ‘end up
comprometendo       [comprometer] <vt> V GER @IMV @#ICL-AUX< ‘compromising’
a               [a] <art> DET F S @>N ‘the’
própria         [próprio] <ident> DET F S @>N ‘as such’
chance          [chance] <ac> <+de+INF> N F S @<ACC ‘chance’
de              [de] PRP @N< ‘of’
conseguir       [conseguir] <vt> V INF 0/1/3S @IMV @#ICL-P< ‘achieving’
uma             [um] <quant2> <arti> DET F S @>N ‘a’
nova            [novo] <ante-attr> ADJ F S @>N 'new'
colocação       [colocação] N F S @<ACC ‘position’

O               [o]  <art> DET M S @>N ‘The’
astro           [astro] N M S @SUBJ> ‘star’
soa             [soar]  <vi> <vK> <rN> V PR 3S IND VFIN @FMV ‘sounds’
,
conforme        [conforme] <rel> <ks> <prp> ADV @COM @#AS-<ADVL ‘in conformity with’
o               [o] <art> DET M S @>N ‘the’
momento         [momento] <dur> <featq> N M S @AS< ‘moment’
,
mais            [muito] <quant> <KOMP> <corr> ADV @>A ‘more’
alta            [alto] <ante-attr> <jn> ADJ F S @<PRED ‘loud’
```

,	
embargada	[embargar] <vt> <vH> <jn> <ADJ> V PCP F S @<PRED ‘handicapped’
,	
ou	[ou] KC @CO ‘or’
apoplética	[apoplético] ADJ F S @<PRED ‘apoplectic’
,	
refletindo	[refletir] <vt> <vH> V GER @IMV @#ICL-<ADVL ‘reflecting on’
a	[a] <art> DET F S @>N ‘the’
terrível	[terrível] ADJ M/F S @>N ‘terrible’
traição	[traição] N F S @<ACC ‘treason’
de	[de] <sam-> PRP @N< ‘of’
a=qual	[o=qual] <-sam> <rel> SPEC F S @P< ‘which’
foi	[ser] <vK> V PS 3S IND VFIN @FMV ‘[she] was’
vítima	[vítima] N F S @<SC ‘[a] victim’

Literature

Almeida, Napoleão Mendes de, *Gramática metódica da língua portuguesa*, São Paulo, 1994

Arndt, Hans, “Towards a syntactic analysis of Danish computer corpora”, in Heltoft, Lars & Haberland, Hartmut (eds.): *Proceedings of the 13th Scandinavian Conference of Linguistics*, Roskilde, 1992

Bache, Carl, "Presentation of a pedagogical sentence analysis system", in *Hermes, Journal of Linguistics*, 17, 1996

Bache, Carl & Davenport, Mike & Dienhart, John & Larsen, Fritz, *An Introduction to English Sentence Analysis*, København, 1993

Bick, Eckhard, *Leksikografiske overvejelser i forbindelse med udarbejdelsen af en portugisisk-dansk ordbog* (unpublished masters thesis), Århus, 1993

Bick, Eckhard, *Portugisisk - Dansk Ordbog*, Mnemo, Århus, 1993, 1995, 1997

Bick, Eckhard, "Automatic Parsing of Portuguese", in Sánchez García, Laura (ed.): *Proceedings of the Second Workshop on Computational Processing of Written Portuguese*, Curitiba, 1996

Bick, Eckhard, "Dependensstrukturer i Constraint Grammar Syntaks for Portugisisk", in: Brøndsted, Tom & Lytje, Inger (eds), *Sprog og Multimedier*, pp. 39-57, Aalborg, 1997

Bick, Eckhard, "Automatisk analyse af portugisisk skriftsprog", in: Jensen, Per Anker & Jørgensen, Stig. W. & Hørning, Anette (eds.), *Danske ph.d.-projekter i datalingvistik, formel lingvistik og sprogteknologi*, pp. 22-20, Kolding, 1997

Bick, Eckhard, "Internet Based Grammar Teaching", in: Christoffersen, Ellen & Music, Bradley (eds.), *Datalingvistisk Forenings Årsmøde 1997 i Kolding, Proceedings*, pp. 86-106, Kolding, 1997

Bick, Eckhard, “Structural Lexical Heuristics in the Automatic Analysis of Portuguese”, in: Maegaard, Bente (ed.): *The 11th Nordic Conference on Computational Linguistics (Nodalida ‘98), Proceedings*, pp. 44-56, Copenhaguen, 1998

Bick, Eckhard, “Tagging Speech Data - Constraint Grammar Analysis of Spoken Portuguese”, in *Proceedings of the 17th Scandinavian Conference of Linguistics*, Odense, 1998 (forthcoming)

Bick, Eckhard, “Portuguese Syntax”, Århus, 1999 (teaching manual, forthcoming)

Brill, Eric, "A Simple Rule-based Part of Speech Tagger", in *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, Italy, 1992

Brill, Eric, "Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging", in *Natural Language Processing Using Very Large Corpora*, Kluwer Academic Press, 1997.

Briscoe, Ted & Carroll, John, "Generalised LR Parsing of Natural Language (Corpora) with Unification-Based Grammars", in *Computational Linguistics*, 19(1): 25-60, 1993.

Castilho, Ataliba Teixeira de (ed.), *Português culto falado no Brasil*, Campinas, 1989

Chanod, Jean-Pierre & Tapanainen, Pasi, "Tagging French - comparing a statistical and a constraint-based method", adapted from: *Statistical and Constraint-based Taggers for French*, Technical report MLTT-016, Rank Xerox Research Centre, Grenoble, 1994

Chomsky, Noam, *Syntactic Structures*, The Hague, 1957

Church, Kenneth, "A stochastic parts program and noun phrase parser for unrestricted text", in *Proceedings of the Second Conference on Applied Natural Language Processing*, Austin, Texas, 1988. (through Cutting, 1992)

Collins, Michael John, "A New Statistical Parser Based on Bigram Lexical Dependencies", in *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, USA, 1996.

Cutting, Doug & Kupiec, Julian & Pedersen, Jan & Sibun, Penelope, "A Practical Part-of-Speech Tagger", in *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, Italy, 1992

Eeg-Olofsson, Mats, *Word Class Tagging, Some computational tools,* University of Göteborg, Department of Computational Linguistics, 1991

Elworthy, D., "Tag set design and inflected languages", in: *Proceedings of the ACL Sigdat Workshop,* pp. 1-10, Dublin, 1995

Francis, W.N. & Kucera, F., *Frequency Analysis of English Usage*, Houghton Mifflin, Boston, 1982. (through Cutting, 1992)

Garside, Roger & Leech, Geoffrey & Sampson, Geoffrey (eds.), *The Computational Analysis of English. A Corpus-Based Approach,* London, 1987

Mateus, Maria Helena Mira & Brito, Ana Maria & Duarte, Inês & Faria, Isabel Hub, *Gramática da Língua Portugesa,* Lisboa, 1989

Järvinen, Timo, "Annotating 200 million words: The Bank of English project", in *Proceedings of The 15th International Conference on Computational Linguistics Coling-94,* Kyoto, Japan, 1994 (quoted from: Pasi Tapanainen, *The Constraint Grammar Parser CG-2,* Publications No. 27, Department of Linguistics, University of Helsinki, 1996)

Järvinen, Timo & Tapanainen, Pasi, *A Dependency Parser for English,* Helsinki, 1997

Karlsson, Fred, "Constraint Grammar as a Framework for Parsing Running Tex*t",* in: Karlgren, Hans (ed.), *COLING-90: Proceedings of the 13th International Conference on Computational Linguistics,* Vol. 3, pp. 168-173

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto (eds.), "Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text, with an application to English", in: *Natural language text retrieval. Workshop notes from the Ninth National Conference on Artificial Intelligence,* Anaheim, CA, American Association for Artificial Intelligence, 1991

Karlsson, Fred, "SWETWOL: A Comprehensive Morphological Analyser for Swedish", in *Nordic Journal of Linguistics* 15, 1992, pp. 1-45

Karlsson, Fred, "Robust parsing of unconstrained text", pp. 97-121, i: Nellike Oostdijk & Pieter de Haan, *Corpus-based research into language,* Amsterdam, 1994

Karlsson, Fred & Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto (eds.), *Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text,* Mouton de Gruyter, Berlin 1995

Koskenniemi, Kimmo, *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production,* Publication No. 11, Department of Linguistics, University of Helsinki, 1983

Koskenniemi, Kimmo & Tapanainen, Pasi & Voutilainen, Atro, "Compiling and Using Finite-state Syntactic Rules", in: *Proceedings of the Fifteenth International Conference on Computational Linguistics, COLING-92,* Vol. I, pp. 156-162, 1992.

Koster, C.H.A.,"Affix Grammars for natural languages", in*: Attribute Grammars, Applications and Systems*, International Summer School SAGA, Prague, Czechoslovakia, June 1991. Lecture Notes in Computer Science, volume 545. Springer-Verlag

Leech, Geoffrey & Garside, Roger & Bryant, Michael, "The large-scale grammatical tagging of text", pp. 47-64, in: Nellike Oostdijk & Pieter de Haan, *Corpus-based research into language,* Amsterdam, 1994

Lezius, Wolfgang & Rapp, Reinhard & Wettler, Manfred, "A Morphology-System and Part-of-Speech Tagger for German", ", in: Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology*, Berlin, 1996

Lindberg, Nikolaj, “Learning Constraint Grammar-style disambiguation rules using Inductive Logic Programming”, in *Proceedings of COLING/ACL ‘98*, volume II, pp. 775-779, Montreal, 1998 and http://www.speech.kth.se/~nikolaj/papers/colingac198/ (15.5.1999)

Magermann, D. & Marcus, Mitchell, "Pearl: A Probabilistic Chart Parser", in *Proceedings of the 1991 European ACL Conference,* Berlin, Germany, 1991.

Marcus, Mitchell, "New trends in natural language processing: Statistical natural language processing", paper presented at the colloquium *Human-Machine Communication by Voice*, organised by Lawrence R. Rabiner, held by the National Academy of Sciences at *The Arnold and Mabel Beckman Center* in Irvine, USA, Feb. 8-9, 1993

Mateus, Maria Helena Mira & Brito, Ana Maria & Duarte, Inês & Faria, Isabel Hub, *Gramática da Língua Portuguesa,* Lisboa, 1989

Monachini, M & Calzolari, N., *Synopsis and comparison of morphosyntactic phenomena encoded in lexicon and corpora,* Technical Report, EAGLES Lexicon Group, 1994. (through Gibbon, 1996)

Müürisep, K., *Eesti keele kitsenduste grammatika süntaksianalüsaator* (Syntactic parser of Estonian Constraint Grammar), Master thesis, University of Tartu, Institute of Computer Science, 1996

de Oliveira, Dercir Pedro, “O Preenchimento, a supressão e a ordem do sujeito e do objeto em sentenças do português do Brasil: um estudo quantitativo”, in: Tarallo, Fernando, *Fotografias sociolingüsticas,* Campinas, 1989

Oostdijk, Nelleke, *Corpus Linguistics and the Automatic Analysis of English,* Amsterdam, 1991.

Padró i Cirera, Lluís, *A Hybrid Environment for Syntactic-Semantic Tagging* (Dissertation at the Universitat Politècnica de Catalunya, Barcelona, 15.12.1997), postscript internet version accessed on 11.2.1998

Perini, Mário A., *Sintaxe Portuguesa, Metodologia e funções*, São Paulo, 1989

Phillips, John D. & Thomsen, Henry S., "A Parser for Generalised Phrase Structure Grammars", in: Haddock, Nicholas & Klein, Ewan & Morrill, Glyn (eds.), *Categorial Grammar, Unification Grammar and Parsing* (Edinburgh Working Papers in Cognitive Science, Vol.1), pp. 115-136 , Edinburgh, Centre for Cognitive Science, University of Edinburgh, 1987 (through Karlsson, 1995)

Puolakainen, T., *Eesti keele morfoloogiline ühestamine kintsenduste grammatika abil* (Morphological disambiguation of Estonian using Constraint Grammar), Master thesis, University of Tartu, Institute of Computer Science , 1996

Radford, Andrew, *Transformational Grammar,* Cambridge, 1988

Samuelsson, Christer & Voutilainen, Atro, “Comparing a Linguistic and a Stochastic Tagger”, in *Proceedings of the 35th Annual Meeting of the Association for Computional Linguistics and the 8th Conference of the European Chapter of ACL,* Madrid, 1999 (forthcoming), and http://www.ling.helsinki.fi/~ãvoutila/cg/doc/e-ac197/e-ac197.html (10.11.99)

Schmid, Helmut, "Probabilistic Part-of-Speech Tagging Using Decision Trees", revised internet version of a paper presented at the *International Conference on New Methods in Language Processing,* Manchester, UK, 1994

Tapanainen, Pasi, *The Constraint Grammar Parser CG-2*, University of Helsinki, Department of Linguistics, Publications no. 27, 1996

Tapanainen, Pasi, *A Dependency Parser for English,* University of Helsinki, Department of Linguistics, Technical Reports, No. TR1, 1997

Togeby, Ole: *PRAXT, pragmatisk tekstteori,* Århus, 1993

Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto, *Constraint Grammar of English, A Performance-Oriented Introduction,* Publication No. 21, Department of General Linguistics, University of Helsinki, 1992

Voutilainen, Atro, *Designing a Parsing Grammar,* Publications No. 22, Department of Linguistics, University of Helsinki, 1994

Wauschkuhn, Oliver, "Ein Werkzeug zur partiellen syntaktischen Analyse deutscher Textkorpora", in: Dafydd Gibbon (ed.): *Natural Language Processing and Speech Technology*, Berlin, 1996

Alphabetical index

Page numbers in bold face indicate main references (chapters, definitions, overviews etc.), italics indicate illustrations or tables. Even where a tag or category is not listed individually in the index, a definition, or examples, can still be found in the appropriate tag set appendix. Also, a relevant index entry may be encountered under a more general heading.

```
ERROR: rangecheck
OFFENDING COMMAND: .installpagedevice

STACK:

-null-
-dictionary-
-savelevel-
```