NORKOMPLEKS. Some Linguistic Specifications and Applications

Torbjørn Nordgård

Department of Linguistics, University of Trondheim, N-7055 DRAGVOLL, NORWAY

KEYWORDS: lexicography, morphology, syntax

AFFILIATION: University of Trondheim

E-MAIL: torbjorn.nordgard@pan.avh.unit.no
FAX NUMBER: (+47) 73 59 61 19
PHONE NUMBER: (+47) 73 59 63 01

1. Introduction

NORKOMPLEKS is a national lexicon project which aims at providing a computational dictionary to be used in computational linguistics applications in Norway (NORKOMPLEKS is an acronym for NORsk KOMPutasjonelt LEKSikon, a computational lexicon for Norwegian). The project is mainly funded by the Norwegian Research Council, but there is also a substantive amount of funding provided by Telenor. The ancestor of NORKOMPLEKS is NORLEX, and this presentation will describe the results of NORLEX which are to be preserved or pursued in NORKOMPLEKS.

The parts of NORKOMPLEKS to be discussed here are written and implemented in Quintus Prolog, but in the descriptions below the information is given in a more readable form.

2. Linguistic Information in NORKOMPLEKS

NORKOMPLEKS exists in a preliminary version where all Norwegian verbs in the Bokmål standard are described in some detail. We will concentrate on the morphological encoding system and the syntactic distinctions that are made in the system.

2.1 The Morphological Encoding System

Both written standards of Norwegian (i.e. Bokmål and Nynorsk) are special compared to most other written standards since they allow many inflection patterns for one and the same lexeme. Consider an example:

(1)

Lexeme:	suge (English translation: to "suck","absorbe","sap","exploit")
Inflections:	imp:"sug" inf: "suge" p-part: "suga" p-part: "sugd" p-part: "suget" pr-part: "sugende" pres: "suger" pret: "saug" pret: "suga" pret: "sugde" pret: "suget"

Obviously, a computational dictionary for Norwegian must be able to cover all forms which are allowed, i.e. the words shown in example (1). This information can be put into the dictionary in two ways: One rather naive alternative is a full listing of all the possible forms attached to each lexeme in the dictionary. The better solution would be to describe the inflections in (1) as a set of inflectional patterns. The patterns are as follows:

(2) suge (infinitive), sug (imperative), suger (present), suget (past), suget (past participle), sugende (present participle)

(3) suge, sug, suger, sugde, sugd, sugende
(4) suge, sug, suger, saug, sugd, sugende
(5) suge, sug, suger, suga, suga, sugende

In NORKOMPLEKS, each of these patterns has a name. The verb “suge” is described as in (6):

(6)

Word:	suge
Category:	Verb
Inflection Codes:	v1,v8,v121,v11
Syntactic Properties:	Transitive, Intransitive, The Particle "inn"+ NP

In a later section we will consider the syntactic properties in more detail. The implemented Prolog-description of this verb is given in (7):

(7)
```
w(105527,suge,[v1,v8,v121,v11],[[trans],[intrans],
[part([inn]),np]])
```

Multiple inflection patterns associated with one particular form is also found in other languages. The distinction between “British” and “American” English is an example. However, the variation in declensions found in Norwegian is unique not only because there are so many possibilities, as witnessed by the example above, but also because there is no legally established norm as to which combinations of codes are preferred. Quite the opposite; – the majority in “Norsk Språkråd” (Board for the Norwegian Language) will not try to establish a norm which could possibly be interpreted as prescriptive “advice” or demand. This does not mean that such norms are illegal. Newspapers or publishers can freely adhere to their own standards, but the authorities will not try to enforce any such norm. It appears, however, to be a fact that a couple of “styles” or systematic connections between inflectional patterns exist. As an example, the code in (2) co-occurs with the pattern in (8), but not with that in (9) (the verb “stryke” has English translations like “stroke”, “delete”, “brush”, “iron”):

(8) stryke stryker strøk strøket strykende
(9) stryke stryker strauk strøket strykende

The situation is the same for inflectional patterns for nouns. In a “moderate” or “conservative” and “Danish-like” style the masculine form “solen” (“the sun”) is used together with patterns (2) and (8), whereas the more “radical” form “sola” is compatible with (9) and (4). (5) is also a “radical” form, and (3) is somewhat intermediate. Clear-cut distinctions are hard to make, however, and corpus studies should be used as the empirical basis for describing these patterns (cf. the remarks below).

Such facts are familiar to most Norwegians, but they are rarely made explicit so that stylistic consistence can be formalized. A lexicon like NORKOMPLEKS will make it easier to establish norms by inspecting the sets of co-occuring pairs or triples of declination codes in the lexicon. When such pairs or triples are found, they can either be “weighted” or inspected manually, or they can be validated on the basis of tagged corpora. But since the amount of tagged corpora of Norwegian is limited, manual “prescription” is the only option. When tagged corpora are available various norms can be deduced on the basis of selected text material. An algorithm based on the following general principles will be implemented and tested, and some initial results will be presented in the talk (the tests will be applied to a few of the available tagged texts):

i. Compute the set of code pairs, triples and quadruples (quintuples have not been identified yet)
ii. For each tagged word in a corpus, decide whether the stem (dictionary entry) contains a code in the set of code pairs, triples or quadruples.
iii. If so, count the code(s) which is (are) compatible with the surface form.

The code in a pair, triple or quadruple which is observed most frequently will be selected, thereby formally characterizing the morphological stylistic flavour of the corpus. Such results are interesting when it comes to the development of spelling correction systems for a “liberal” language like Norwegian because the spelling corrector will be able to detect inconsistent forms that do not conform to some predefined norm (i.e. a user defined norm or a norm defined by some organization).

2.3 The Syntactic Encoding System

In automatic sentence processing the amount of structural ambiguity during analysis is a well-known and pervasive problem (cf. for instance Hirst 1986), especially when atomic symbols in grammars are replaced by more structured and informative entities, i.e. feature structures (see e.g. Barton et.al.). There are various techniques which can be used in order to reduce these problems, for instance tabular parsing (Early (1970), Kaplan (1973), Wiren (1992)), combinations of data-driven and hypothesis driven algorithms, lookahead buffers (Marcus 1980, Nordgård 1993,1995)), and so on. But it appears to be the case that the amount of ambiguity can most successfully be reduced when lexical items are equipped with detailed information about the local contexts in which they can appear (here we put aside problems related to displaced constituents and empty categories). In this connection another aspect of NORKOMPLEKS is interesting because this lexicon contains quite a lot of syntactic information associated with each verb. An example was given above, cf. (7). The classifications used in NORLEX (and NORKOMPLEKS) are given in (10):

(10)

CODE	EXPLANATION
trans	Transitive verb
intrans	Intransitive verb
trans1	Special intransitive verb
intrans1	Ergative verb
intrans2	Verb with no thematic roles
seg	Reflexive verb
pp[(på)]	PP-complement; head = på
seg,np	Reflexive verb + np
seg,pp[(med)]	Reflexive verb + pp
np,part([bort,vekk])	Particle + np
s([at])	Sentential Complement
s([å])	Infinitival Complement
np,s([å])	np + Infinitival Complement
np,s([at])	np + Sentential Complement
vp	VP-komplement
part[over]	Particle
part([mot]),s([å])	Particle + Infinitival Complement
	np,part([mot]),s([å]) : np + Infinitival + inf.kompl

Observe that the classifications are surface oriented. Their interpretation in some syntactic or semantic theory is a different issue.

A truly novel contribution of the lexicon is the classification of matrix verbs, i.e. verbs which take sentential complements, infinitivals included. This type of information is absent in most electronic dictionaries, but it is very important in natural language processing because it can be used to restrict the number of hypothesized embedded sentences in standard data-driven parsers. Incorrect hypotheses about embedded sentences are expensive because the search space can be dramatically broadened, depending on the exact formulation of the rules, of course.

Since verbs have a fairly rich information about their syntactic properties this lexicon will make it easier to assign the correct syntactic properties to verbs in running text. But on the other hand, some of the descriptions are locally ambiguous, as for the verb “advare” (“warn”). This verb has the following syntactic properties:

(11)

trans	It can be used as an ordinary transitive verb
np,pp(mot)	It can take a nominal object and the particle "mot"
np,part(mot),s(å)	It can take a nominal object, the particle "mot" and an infinitival clause
pp(mot)	It can take a prepositional object, headed by "mot"
part(mot),s(å)	It can take the particle "mot" and an infinitival clause

If a rule-based tagger is expected to choose among these descriptions it will have to inspect the string to the right in order to decide which syntactic version is the correct one. If only one lookahead buffer cell is provided, the tagger will have serious problems if the verb is followed by a noun because all the first three possibilities would be applicable in such a case. A tagger with some general lookahead buffer could be used to handle such cases. Constraint-based tagging systems appear to be of particular interest in this respect because they have flexible means of inspecting the input string backwards and forwards, cf. Voutilainen et.al.

It should be noted that the descriptions in NORKOMPLEKS will be modified in the next couple of years because there is a need to include more syntactically and semantically relevant information, e.g. in the description of control verbs. The statement “np,s([å])” does not say anything about whether we are dealing with subject or object control. Neither is information about thematic roles included.

3. Concluding Remarks

Two important aspects of NORKOMPLEKS have been discussed: The morphological encoding system and the syntactic information associated with each verb. This lexicon is the first large-scale machine-readable dictionary for Norwegian. It has information that goes beyond “standard” information in electronic dictionaries, and it has a morphological coding system which makes it suitable for use in a variaty of language environments, including stylistically marked or specially defined contexts where spelling norms are controversial.

References

Barton, Berwick and Ristad (1987) Computational Complexity and Natural Language. MIT Press, Cambridge, Mass.

Earley, Jay (1970): An Efficient Context-Free Parsing Algorithm. Communications of the ACM, 14, 453-460.

Hirst, Græme (1986). Semantic Interpretation and the Resolution of Ambiguity. Cambridge University Press, Cambridge.

Kaplan, Ronald (1973): A General Syntactic Processor. In Natural Language Processing (Randall Rustin, ed.). Algorithmics Press, New York.

Marcus, Mitchell (1980): A Theory of Syntactic Recognition for Natural Language. MIT Press, Cambridge, Mass.

Nordgård, Torbjørn (1993): A GB-Related Parser for Norwegian. Peter Lang, Berne.

Nordgård, Torbjørn (1995): E-Parser: An Implementation of a Deterministic GB-Related Parsing System. In Computers and the Humanities 28:259-272.

Vuotilainen, Heikkilä and Anttila (1992): Constraint Grammar of English. University of Helsinki, Dept. of General Linguistics.

Wiren, Mats (1992): Studies in Incremental Natural-Language Analysis. Doctoral Dissertation, University of Linköping.