# Representation, information theory and basic word order

Luke Maurits

*Thesis submitted for the degree of*
*Doctor of Philosophy*
*in*
*Psychology*
*at*
*The University of Adelaide*

School of Psychology


THE UNIVERSITY
*of* ADELAIDE

September 2011

# Contents

**Abstract** **i**

**Signed Statement** **iii**

**Acknowledgements** **v**

**1 Introduction** **1**
1.1 A motivating problem . . . . . 1
1.2 Psychological and adaptationist answers to linguistic questions . 2
1.3 Structure of thesis . . . . . 3

**I The problem: variation in basic word order and how to explain it** **7**

**2 Basic word order** **9**
2.1 Defining basic word order . . . . . 9
2.2 Cross-linguistic distribution of basic word order . . . . . 19
2.3 What we need to explain . . . . . 24
2.4 Summary . . . . . 25

**3 Explanations for the frequencies** **27**
3.1 Explaining language diversity in general . . . . . 27
3.2 Previous functional explanations for word order frequencies . . . 39
3.3 Summary . . . . . 49

**4 A problem with previous explanations and a solution** **51**
4.1 A problem . . . . . 51
4.2 Common directions of basic word order change . . . . . 54
4.3 Assessing the compatibility of standard functional explanations against the diachronic evidence . . . . . 58
4.4 A solution . . . . . 60
4.5 Summary . . . . . 68

**II Initial conditions: majority descent from SOV and mental representation 69**

**5 Evidence and explanations for majority descent from SOV 71**
5.1 Evidence for majority descent and privileged status . . . . . 72
5.2 Explanations for majority descent and privileged status . . . . . 77
5.3 Summary . . . . . 83

**6 Seeking SOV in the mind: experimental results 85**
6.1 What does it *mean* to think in SOV? . . . . . 85
6.2 Experimental investigation . . . . . 92
6.3 Summary . . . . . 115

**7 Discussion of SOV representation 117**
7.1 A proposal for subexplanation E1 . . . . . 117
7.2 Future research . . . . . 118

**III Dynamics: systematic drift away from SOV and UID functionality 123**

**8 Uniform information density and word order functionality 125**
8.1 Introduction . . . . . 125
8.2 Theoretical prerequisites . . . . . 125
8.3 The UID Hypothesis . . . . . 132
8.4 Linking word order and information density . . . . . 138
8.5 Mathematical formalism . . . . . 140
8.6 Summary . . . . . 144

**9 Estimating UID functionality from corpora and an experiment 145**
9.1 Corpus analysis . . . . . 145
9.2 Elicitation of event distribution . . . . . 161
9.3 Discussion . . . . . 167
9.4 Summary . . . . . 172

**10 Discussion of UID functionality 173**
10.1 A proposal for subexplanation E2 . . . . . 173
10.2 Undersanding UID word order functionality . . . . . 174
10.3 Future research . . . . . 177

**IV Conclusion 179**

**11 Extending the explanation beyond basic word order 181**
11.1 A brief review of word order typology . . . . . 182
11.2 Going beyond EIC . . . . . 188

11.3 Summary . . . . . 194

**12 Conclusion** **197**
12.1 Summary of proposed explanation . . . . . 197
12.2 Why are there still so many SOV languages around today? . . . 199
12.3 Answering Tomlin's questions . . . . . 200
12.4 Assessment of proposed explanation . . . . . 201
12.5 Future research . . . . . 203
12.6 Summary . . . . . 208

**Bibliography** **210**

**References** **211**

**Abstract**

Many of the world's languages display a preferred ordering of subject, object and verb, known as that language's *basic word order*. There are six logically possible basic word orders, and while each occurs in at least one known language, not all are found equally frequently. Some are extremely rare, while others are used by almost half the world's languages. This highly non-uniform cross-linguistic distribution of basic orders is a fundamental explanatory target for linguistics.

This thesis tackles this problem from a psychological perspective. It constitutes an advance over previously proposed explanations in that it is compatible not only with the distributions observed today, but with what is known of broad trends in the word order change which happen over hundreds of years. There are two largely independent components of the explanation given in this thesis, which is necessary to be compatible with both synchronic and diachronic evidence.

The first component is focused on the structures which the human mind uses to represent the meanings of sentences. While mental representations of meaning are not inherently serial (hence ordered) like spoken language, we can think of the different components in these representations as being ordered in a different sense, based on some components being more accessible to cognitive processing than others. This thesis develops the idea that the word order used most often in the earliest human languages, which are taken to rely on a direct interface between mental representations and motor control systems, were determined by a "word order of the language of thought".

The second component is focused on the functional adequacy of different word orders for high speed, reliable communication. The driving idea here is that human language represents a rational solution to the problem of communication. The mathematical formalism of information theory is used to determine the gold standard for solutions to this problem, and this is used to derive a ranking of word orders by functionality. This thesis develops a novel perspective on word order functionality in which cross-linguistic preferences are ultimately a reflection of statistical properties of the events which languages describe.

# Signed Statement

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.



SIGNED: ..... DATE: .....

# Acknowledgements

First and foremost, I owe a tremendous debt of gratitude to my supervisors, Dan and Amy. Before beginning this PhD I knew very close to nothing about cognitive science or linguistics, and at the end it of a huge proportion of what I know is either something one of them taught me or something I taught myself after one of them pointed me in the appropriate general direction. Shortly after meeting Dan and having my academic worldview (which I developed as an undergraduate exclusively in maths and physics) violently shifted by learning that there were people who called themselves psychologists *and* used maths (at the same time!), I was offered the opportunity to start a PhD under him, and I am thankful for the leap of faith on his behalf in doing this despite my total lack of relevant background. Looking back at the research proposal I wrote six months into the endeavour, it's embarassing how naively overconfident I was about breaking amazing new ground in psycholinguistics (using Markov chains, no less), and it is a small wonder that when Amy turned up at around this time she took me as seriously as she did. I am very lucky to have had two approachable and helpful supervisors with strong mathematical and computational backgrounds to guide me while I found my feet in a new field.

Natalie May and Tim Ck, my fellow members of Dan and Amy's labs at the University of Adelaide, assisted with recruiting and scheduling participants for the experiments presented in this thesis, which was a tremendous time-saver for me. I am grateful to both of them.

While working on this thesis, I was lucky enough to be able to travel to Vancouver for the Neural Information Processing Systems conference to talk on an embryonic form of the material which appears in Chapter 9 of this thesis. I was able to travel to Canada to attend NIPS thanks to a number of sources of funding beyond that provided by the University of Adelaide, including a grant from the Walter & Dorothy Duncan Trust and a NIPS travel grant which was sponsored by Google.

As part of the trip to NIPS, I was able to visit a number of universities in the northern hemisphere to discuss my work, in many cases with scholars upon whose ideas I was directly buiding. I am very grateful to everyone who allowed me to visit their lab: Nick Chater at the University College, London's Cognitive, Perceptual and Brain Sciences Research Department and University of Warwick's Warwick Business School, Tom Griffiths at the University of California, Berkeley's Computational Cognitive Science Lab, Simon Kirby at the

University of Edinburgh's Language Evolution and Computation research unit and Roger Levy at the University of California, San Diego's Computational Psycholinguistics Lab. Not only did they allow me to speak at their labs, but they and many of their colleagues took particular interest in my ideas on word order and Uniform Information Density and discussed their thoughts on this matter with me at length. Having my ideas taken seriously by "real linguists" was a tremendous confidence boost for someone who very much felt like they had been seriously theorising by the seat of their pants.

I am thankful to the many graduate students at the institutions above who offered their companship during my visits, and particular to those who provided me with transportation or accommodation, who were: Sean Roberts, Marton Soskuthy and Rachael Bailes in Edinburgh, Joseph Austerweil and Karen Schloss in Berkeley and Klinton Bicknell in San Diego. I am also indebted to Matt Hall at UCSD for a very enthusiastic hour of discussing and swapping references on improvised gestural communication.

Merrit Ruhlen of Stanford University, Matthew Dryer from the University at Buffalo and Albert Bickford at the Summer Institute of Linguistics all provided me with copies of papers which I was unable to locate otherwise.

Richard Sproat, from Oregon Health & Science University, whom I met at the HCSNet WinterFest event in Sydney in 2010, introduced me to the World Atlas of Language Structure, which proved to be incredibly useful for speculating about word order typology.

Kayako Enomoto, at the University of Adelaide's Centre for Asian Studies, bent rules against strong Faculty resistance in allowing me to audit the Japanese 1A class, giving me my first practical experience with a language typologically dissimilar to English, which was tremendously helpful in thinking about language, and particularly syntax, in a more general way than I could have otherwise.

Finally, I am grateful to my parents for their encouragement and support throughout the last three and a half years, and to my wife Kirsty for the same, as well as for proofreading and for putting up with a lot of neglect in the final weeks of preparing this thesis!

# Chapter 1

# Introduction

## 1.1 A motivating problem

One of the most important (and, to my mind, the most interesting) problems in the field of linguistics is to answer the general question of why languages are the way they are, and not some other way. This concern lies within the subfield of *linguistic typology*. Typologists endeavour to discover and explain the full extent of linguistic variety in human languages: what are the interesting ways in which two languages can differ? Of all the interestingly different kinds of language (or "types") which could exist in principle, which ones of them do we actually find in practice? Why do we find these types and not others? As part of their effort to chart the full range of human linguistic diversity, typologists often discover what are called *language universals*: statements which are true of all, or almost all, of the known languages.

Like languages themselves, language universals come in a range of types. Universals can be classified as either *absolute* (i.e. statements $P$ such that $P$ is true for every single known language) or *statistical* (i.e. statements $P$ such that $P$ is significantly more likely to be true for any given language than one would expect by chance alone), and also as *implicational* (i.e. statements of the form "if $P$ is true for a language, then $Q$ is also true for that language: $P \rightarrow Q$) and *non-implicational*. Instances of all four of the language universals described by this two-by-two taxonomy have been claimed by typologists, and their scope covers the entire range of linguistic phenomena, from phonetics, through phonology and morphology, to syntax.

However, there is little intellectual satisfaction in merely collecting language universals, as if they were interesting trinkets. What makes universals genuinely interesting is that, at least in the opinion of some, they cry out for an explanation. There appears to be no good *a priori* reason why, for instance, if a language has VSO (verb, subject, object) as its normal word order, it should necessarily use prepositions as opposed to postpositions (e.g "to the city" instead of "the city to"). And yet, so far as we can tell, this is in fact always true (an example of an absolute implicational universal, first stated by

Greenberg (1963)). Surely there must be some good reason *why* we never see postpositional VSO languages?

While the existence of language universals is very well accepted in the linguistics community (although not completely accepted; see, e.g., (Evans & Levinson, 2009)), there is somewhat less agreement over the question of how they should best be explained. One of the main schools of thought on this matter goes by the label of *functionalism*. The general thesis of this school is that languages are the way they are because by being so they do their job - that is, facilitating communication between language users - better than they would if they were not, by virtue of being easier to use, faster to convey information, less prone to ambiguity, etc.

In this thesis I shall consider what seems like a fairly straightforward universal, of the statistical, non-implicational type. The universal relates to the relative frequency of different word orders across the languages of the world (to be more precise, different *basic* word orders, as shall be defined later). The three constituents of a simple declarative utterance are typically taken to be the subject, verb and object, and these can be arranged linearly in only one of six logically possible orderings (for example, VSO, as featured in the example universal above). As we shall see, not all six of these orderings are equally common. Some orders occur very frequently, while others are exceedingly rare. Some fall in between these two extremes. The problem of explaining the particular distribution of word orders has been acknowledged for quite some time, and multiple explanations within the functionalist framework have been put forward. However, I shall argue that these previous explanations fall somewhat short of the mark, and shall develop an alternative.

The new explanation for basic word order frequencies presented in this thesis is more complicated than previous explanations, but this appears to be a necessary price to purchase the required empirical adequacy. I shall show that, when viewed from a historical or diachronic perspective, the problem of explaining word order frequencies factors quite neatly into two logically independent problems. These two problems require us to focus our attention not simply on the distribution of basic word orders observed in languages today, but to consider both the very origin of the human capacity for language, and the dynamic forces which have shaped the diverse landscape of human languages that has developed since that origin.

## 1.2 Psychological and adaptationist answers to linguistic questions

While the problem of explaining the relative frequency of different basic word orders is very squarely a problem in linguistics, my approach to solving the problem shall be just as squarely psychological in nature. Linguistics and psychology are not at all strangers to cross-disciplinarity: the production and

processing of language by the brain has long been studied by the field of psycholinguistics, and a significant part of this thesis shall fall very much within the traditional scope of that field, in that it will deal with the word-by-word incremental processing of utterances by the brain. Another significant part, though, shall be concerned primarily with the mental representation of the the meanings of utterances, something I consider to be extra-linguistic, and something which is more traditionally studied within psychology proper, particularly cognitive psychology. I therefore consider this thesis as a whole to be most appropriately considered a work of cognitive science, a field which I define as the study of the mind as an information processing device. This is a field which subsumes at least parts of both linguistics and psychology (along with neuroscience, artificial intelligence and numerous other fields). Language is considered here not for its own sake (as interesting as it surely is), but as a mirror on the human mind, an approach which has a long and rich tradition. Since the content of this thesis is likely to be of interest to both those who consider themselves psychologists but not linguists and linguists but not psychologists, I shall try to avoid assuming that the reader has knowledge of specialist terminology from either of these fields.

In addition to involving a close interplay of linguistics and psychology, the work in this thesis can also be characterised as being what I shall call "adaptationist" in nature, on multiple levels. As mentioned above, the consideration of the present day relative frequencies of basic word orders shall in fact lead us back to the dawn of the human capacity for language, to consider why the earliest language or languages may have had the form that they apparently did. A central part of these considerations shall be the issue of how a non-linguistic mind becomes adapted to become a mind with language. I shall also consider how languages have changed from this original starting point, adapting slowly over time to better suit the psychological structures and processes underlying human language, and to better facilitate its selected-for purpose of efficient and robust communication. I also shall consider how both the human mind and languages are adapted to fit the physical environment in which they have developed. This common factor to the work herein ultimately traces its intellectual heritage back to the celebrated work of Darwin (1859), but also draws heavily on the traditions of rational modelling in psychology, as formulated by Anderson (1990), and the functionalist school of thought in linguistics, particularly in the more modern and expansive meaning of the term.

## 1.3 Structure of thesis

This thesis is divided into four parts.

Part I sets the stage for the remainder of the thesis. It begins in Chapter 2 by introducing the concept of basic word order and surveying the previous literature on the cross-linguistic distribution of values of this syntactic param-

eter. In so doing, it establishes precisely what this thesis is signing itself up for in aiming to "explain basic word order". Chapter 3 is concerned with what constitutes a viable approach to explaining variation in a linguistic parameter. It begins by considering the general problem of explaining linguistic diversity, of which my goal is a special case, and ends with a survey of previously offered explainations of basic word order distribution which belong to a particular class, known as *functional* explanations. The section is concluded with chapter 4, which identifies a significant shortcoming associated with all previous functional explanations, and outlines the general form that an explanation of basic word order distribution must have in order to avoid these shortcomings. This general form involves a shift in perspective from synchronic to diachronic explanations: basic word order distribution must be explained in terms of both the "initial conditions" present at the origin of human language, and in terms of the dynamics of word order change which have changed the initial conditions into the presently observed data. The remainder of the thesis is then concerned with these two subjects. While these subsequent parts, and especially part III, form the real "meat" of the thesis, the reasoning in Part I which motivates them represents by itself a significant piece of novel work.

Part II is concerned with the question of which basic word order was used by the earliest human language, from which all present languages are descended, and explaining why these word orders and not others should have played this role. Obviously these are difficult questions, but some degree of informed theorising is certainly possible. The emphasis shall be on the idea that all or almost all extant languages are descended from earlier languages with SOV basic word order. Chapter 5 briefly surveys previous work arguing for this common descent from SOV and suggesting explanations for why SOV may have been the ancestral basic word order, including appeals to a linguistic analogue of an effect from population genetics, to linguistic functionalism, and to the structure of humans' mental representations of events. In Chapter 6, I attempt to make the latter explanation above rigorous, and present the results of an experiment, inspired by recent work by Goldin-Meadow and colleagues (Meadow, So, Ozyurek, & Mylander, 2008), suggesting that the SOV word order is the one most compatible with speakers' internal representation of events. Chapter 7 discusses how the results of Chapter 6 fit into the explanatory framework developed in Chapter 4, discusses directions for future research with regard to this particular idea, and briefly considers the question of why humans may have evolved such an internal representation. Any answer to this question must necessarily avoid appeals to linguistic functionalism, suggesting that at least some of the explanation for basic word order distribution must be provided by cognitive science in general, rather than linguistics alone.

Part III is concerned with the dynamics and mechanisms of word order change. The key idea of this part of the thesis is the forging of a link between the basic word order parameter and the Uniform Information Density hypothesis (UID) of Levy and Jaeger (Levy, 2005; Jaeger, 2006). In Chapter 8 I show

that the independently well supported UID hypothesis can be used to derive a novel measure of word order functionality. In Chapter 9 I analyse data from several corpora of spoken and written speech and from an experiment in order to evaluate this functionality measure, and find that it is largely compatible with the available data on word order change, in contrast to previous accounts of word order functionality. Chapter 10 discusses how the results of Chapter 9 fit into the explanatory framework developed in Chapter 4, works toward developing an intuitive theory of why UID word order functionality works the way it does, and discusses directions for future research with regard to this particular idea.

Part IV concludes the thesis. First, Chapter 11 briefly investigates the prospects for extending the basic explanatory strategy used in the thesis to word order typology in general. There are many word order related syntactic parameters, such as the relative order of nouns and adjectives, or the use of prepositions vs postpositions, which have interesting distributions. Both of the major ideas I have applied to basic word order, non-linguistic mental representation of meaning and information density are in principle applicable to these other parameters. Although my treatment of these considerations is necessarily brief and speculative, I suggest that there seem to be no major obstacles to extending the work here to explaining these other facts, and demonstrate some promising interactions between basic word order and two word order parameters. Then, and finally, Chapter 12 summarises the explanation of basic word order distribution developed up to that point, highlighting its various merits and considering future research directions.

# Part I

# The problem: variation in basic word order and how to explain it

# Chapter 2

# Basic word order

## 2.1 Defining basic word order

This thesis is ultimately about the typological property of languages called *basic word order*, and so it seems appropriate to begin by explaining precisely what this property is. However, basic word order is a deceptively simple seeming concept. It is very easy to use a few examples to give a quick and intuitive understanding which will certainly convey most of the flavour of basic word order, but these sorts of definitions quickly turn out to be far too simplistic for serious scholarly work. In fact, somewhat alarmingly, if one insists on finding a definition which meets what might be thought of as the minimum requirements for serious work, such as a being precise, objective and language-general, one will quickly find that there is absolutely no agreement whatsoever in the field of typology on what such a definition might be! If I am to be honest I must admit straightup that basic word order is technically not a well-defined concept. Despite this, as we shall see later in this chapter, there is enough consistency in various cross-linguistic surveys of basic word order that I think we can be confident that, underlying the confusion and uncertainty, there is in fact a "real" concept of interest, and furthermore that we know enough about this concept's distribution to tackle the problem of explaining it.

In this section I shall present a brief overview of the various difficulties involved in defining basic word order. While I have no intention of being able to succeed where others have failed in offering a satisfactory general definition, I hope that in this section I can achieve two things. First, I hope to give the reader enough of an intuitive feel for what basic word order is that they can feel comfortable in following the rest of the material in the thesis. Secondly, I would like to convince the reader that the lack of such a general definition and the various problems which prevent such a definition are not reasons to consider the entire enterprise of explaining basic word order definitions futile.

### 2.1.1 A rough guiding definition

Basic word order is essentially grounded on properties of a certain kind of sentence. The kind of sentence we are interested in is the *declarative sentence*, that is, a sentence which makes a statement about something, rather than being asking a question or issuing a command. Furthermore, it is a declarative sentence which states, very loosely, that something has done something to something else. Examples of the sort of sentence we are interested in, in English, are:

(1) The dog bit the man

(2) John kissed Mary

(3) A fire destroyed the town

Note that we are not interested in declarative sentences involving more than two items (or two noun phrases), such as "John gave Mary the book". An underlying assumption in trying to define basic word order is that all languages contain sentences of the kind described above.

The sentences we are interested can typically be analysed as having three constituents: a subject (S), a verb (V) and an object (O). For instance, in "the dog bit the man", the subject is "the dog", the verb is "bit" and the object is "the man". As a matter of logical necessity, any such sentence must place these three constituents in some linear order, for example SVO in the caes of "the dog bit the man" (or, indeed, all of the three example sentences above). There are six logically possible ways in which these constituents can be ordered, and these are SOV, SVO, VSO, VOS, OVS and OSV.

In most languages, one of these six orders can be considered as being, loosely, the most typical, natural or important. For instance, the majority of English declarative sentences use SVO word order rather than any of the other five, although there are exceptions. On the other hand, Japanese declarative sentences are most often constructed using SOV word order, as the example below shows:

(4) Inu ga otoko o kamimashita
Dog (subj) man (acc) bit
"The dog bit the man"

As a first approximation to a definition, then, we can say that the *basic word order* of a language is the ordering of subject, verb and object which best characterises declarative sentences in that language. In many cases, assigning a basic word order to a language is quite straightforward and this sort of informal definition is arguably entirely adequate. For example, speakers of English should have little trouble in convincing themselves that English has a basic word order of SVO. However, if we wish to assign a single basic word order, with confidence, to every language in the world, we rapidly run into a number of complications and suddenly things are not so straightforward. The following section deals with the major sources of this trouble.

### 2.1.2 Complications

If we wish to turn the rough definition above into a rigorous definition suitable for the purposes of constructing large scale cross-linguistic surveys of basic word order, there are two main complicating issues we have to deal with.

#### Choosing one order out of several

The first is that many languages permit more than one ordering of subject, verb and object. For instance, the Japanese sentence "otoko o inu ga kamimashita" has the same meaning as the sentence "inu ga otoko o kamimashita" that we saw earlier, and is also considered perfectly grammatical, but it has OSV word order. On precisely what grounds should we claim that Japanese is an SOV language and not an OSV one? In some languages, such as Finnish, Hungarian, Portugese and Russian, *any* ordering of the subject, verb and object can be grammatical, because case marking is used to make it clear which role each word is playing. What is the basic word order of a language in which there is more than one grammatical word order? Another problem is that some languages use different word orders for different parts of complicated sentences. For example, in German and Dutch, SVO word order is used in the main clauses of sentences, while SOV word order is used in any embedded clauses. The following German sentences demonstrate this:

(5) Ich trinke Kaffee
I drink coffee
"I drink coffee"

(6) Du weisst dass Ich Kaffee trinke
You know that I coffee drink
"You know that I drink coffee"

The second sentence, "you know I drink coffee" has SVO word order - the subject is "you", the verb is "know" and the object is "I drink coffee". Notice that the object here is itself a sentence, one which is embedded inside a more complex sentence. The embedded sentence has SOV word order, but notice that if we want to say "I drink coffee" by itself, unembedded, SVO word order is used instead. Should German be given a basic word order of SVO or SOV?

With regard to the problem of languages permitting more than one word order, it is not at all difficult to think of ways to decide which one should count as basic, and indeed a variety of criteria have appeared in the literature. Newmeyer (1998) identifies at least the following methods of choosing the basic word order for a language:

- Choose the order with the highest text frequency.
- Choose the order in which S and O are full NPs.
- Choose the order that carries the fewest special presuppositions.

- Choose the order associated with the most basic intonation contour.
- Choose the order associated with the least overall syntactic or morphological elaboration.
- Choose the order at a motivated underlying level of syntactic structure.

Oftentimes it may turn out that most or even all of these criteria will yield exactly the same decision, in which case a basic word order can be assigned without any real controversy. Russian, for instance, is widely accepted to have SVO basic word order (Dryer, 2008), even though all six possible word orders can be used grammatically. In some cases multiple criteria like those above may fail to distinguish any one word order as being basic. These languages are usually said to have *free word order* (often denoted FWO in keeping with the "three capitals" style of notation for basic word order), and no basic word order is assigned. The trickiest case, of course, is when several different criteria such as those listed above all yield a clear decision, but that decision is not consistent across criteria. In this case the only way to assign a basic word order is to make an essentially arbitrary choice of one criteria as "the chosen one" and to use the word order it nominates.

With regard to languages where different word orders are used in main and embedded clauses, this once again comes down to essentially making an arbitrary choice as to which clause class should "count". (Comrie, 1981) notes a tendency for opinion on this matter to be roughly split along what he calls "ideological" lines: linguists identifying themselves as "generativists" tend to take the word order of embedded clauses as basic, while "non-generativists" take the main clause word order. Some linguists will decline to assign a basic word order to languages such as German and Dutch, rather than taking an arbitrary stance on this matter.

Just how much of a problem the need to make arbitrary choices in assigning basic word orders to languages is when it comes to constructing large cross-linguistic surveys depends on just how many languages represent tricky cases where different criteria yield different results. I shall discuss this in more detail later in the chapter, but it seems to me that the problem is not an especially great one. When we take several surveys of reasonable size from a wide variety of authors (many of whom have themselves compiled their surveys by using word orders assigned by multiple other authors), there is almost perfect agreement in how basic word orders should be ranked from most to least frequent.

### Defining subject, verb and object

The second major complication in making the intuitive definition rigorous is that there needs to be some way to objectively decide which of the constituents of a sentence is the subject, which the object and which the verb, and this definition must work for any language we throw at it. This turns out to be a

very difficult task indeed, although the difficulty is mostly confined to distinguishing between subjects and objects. To appreciate just why this is so difficult requires an excursion into the rather discouragingly complicated world of linguistic terminology.

So, what exactly are the "subject" and "object" of a sentence?. For the non-linguist, the intuitively obvious definitions (perhaps the only sensible seeming definitions) are that the subject is the thing which is "doing the verb", and the object is the thing which is "having the verb done to it". From this perspective, the sentence "the dog bit the man" has SVO word order, while its passive voice equivalent "the man was bitten by the dog" has OVS word order. However, this is *not* the sense in which linguists typically use the terms subject and object. Clarifying precisely the appropriate terminology for these ideas requires a brief detour into linguistic terminology.

The term generally used by linguists to refer to the thing "doing" a verb is either *agent* or *actor*. Some authors impose a principled distinction between agent and actor. For instance, for Jackendoff (Jackendoff, 2007), anything doing a verb is an actor, and the term agent is reserved for things which initiate or cause verbs, thus requiring some degree of volition or animacy on behalf of the agent. Thus, "cat" is both an actor and agent in "the cat ate the fish", but "car" is only an actor in "the car hit the tree". Presumably other authors use other conventions. In this thesis, I shall consistently use only the term agent to refer to the thing which is doing a verb, and will never use the term actor, since I have no need of the distinction between things which do and do not have volition or animacy, or any other distinction. In the example sentences below, I have italicised what I shall be calling the agent.

(7) The *boy* chased the duck.

(8) The *fire* destroyed the house.

(9) The charity was funded by a *millionaire*.

The linguist's term used for the thing that a verb is done to seems to be, fairly consistently, *patient*. Agent and patient are both examples of what are called, variously, *thematic relations, thematic roles, semantic relations*, or *semantic roles*[1]. To avoid confusion, I shall always use the term *semantic role* in this thesis to describe the concepts of agent and patient.

What, then, of subject and object? To a linguist, subject and object are examples of what are called, variously *syntactic relations, syntactic roles, grammatical relations* or *grammatical roles*. To avoid confusion, I shall always use the term *syntactic role* (so as to make maximally clear both the relation to and contrast with semantic role). Syntactic roles, as the name suggests, have nothing to do whatsoever with the semantics of what a sentence is describing, and are instead defined entirely syntactically. For example, in English, the subject

[1]The term thematic role is not to be confused with the term *theta role* used in some flavours of generative grammar, although the concepts are not unrelated.

of a sentence governs, among other things, the phenomenon of verb agreement. In the four examples below, I consider the singular and plural cases of the two noun phrases in a group of closely (semantically) related sentences:

(10) The cat is sitting on the mat

(11) The cat is sitting on the mats

(12) The cats are sitting on the mat

(13) The cats are sitting on the mats

Note that the choice of auxiliary verb (either "is" or "are") is constrained by the plurality of the first noun phrase ("the cat" or "the cats") but is entirely unaffected by the plurality of the second noun phrase ("the mat" or "the mats"). This is one example of a purely syntactic phenomenon in English which distinguishes one noun phrase, "the cat(s)", from another, "the mat(s)". The privileged noun phrase is called the subject, while the other is called the object. Sentences like the following:

(14) * The cats is sitting on the mat

are ungrammatical in standard English because the subject and verb do not "agree". Verb agreement in English is not limited to plurality considerations. Consider the first, second and third person singular sentences below:

(15) I run

(16) He runs

(17) They are running

Here the verb "run" changes its form in each sentence to agree with the subject, even though in each case the subject describes a single person (though in the third case it could just as well describe a group of people).

In the examples above, the subject of each sentence is also the agent, so that these two roles, one syntactic and one semantic, overlap. However, consider the passive voice equivalents of the sentences above:

(18) The mat is being sat upon by the cat

(19) The mats are being sat upon by the cat

(20) The mat is being sat upon by the cats

(21) The mats are being sat upon by the cats

Notice that here the noun phrase "the mat(s)" is the subject of the sentence, as determined by verb agreement, while the agent is still, in every case, "the cat(s)". In switching from the active to passive voice, the semantic roles have not changed, but the syntactic roles have. This illustrates that subject and agent are, in fact, quite distinct phenomena, and also illustrates that the English passive voice construction still has SVO word order, not OVS word order as one might intuitively expect.

In languages other than English, the subject may be distinguished from other noun phrases in a sentence by various means, including verb agreement (on the basis of plurality, person, gender or other properties of noun phrases), case marking, and others, but the important point for current purposes is that whatever it is that identifies a noun phrase as being a subject, it is syntactic in nature and not semantic.

A similar discussion holds for the concept of "object": object is not the same thing as patient, although in many cases the two may overlap. Objects, like subjects, must be defined on purely syntactic terms. In English, they are defined mostly in contrast to the subject, i.e. the object is the noun phrase of a sentence which verbs are *not* required to agree with. In other languages, of course, objects can be defined more substantively, although always in syntactic terms.

So far so good, but now for the twist: the answer to the entirely reasonable question "are there single, rigorous definition of "subject" and "object", in purely syntactic terms, which hold across all languages and which a mjaority of linguists are in agreement over?" is a resounding "no". There are difficult cases in which there is no consensus on how to identify the subject or object of a sentence, and there is even uncertainty as to whether there *are* single coherent concepts of subject and object which can be applied across all languages. For more detailed discussion of these issues, see, e.g. (Comrie, 1981). This whole situation seems to leave the definition of basic word order grounded on very little indeed: if linguists are uncertain that subject and object are well defined cross-linguistically, it seems to follow as an immediate logical consequence that any concept defined in terms of the relative order of these concepts is necessarily also not well defined.

### 2.1.3 How big is the problem really?

At this point I may well have done a very good job of convincing the reader that the goal of this thesis, to provide a psychological explanation for the cross-linguistic distribution of basic word orders, is an entirely futile endeavour. The definition of just what basic word order *is* appears to be ultimately an arbitrary choice among multiple competing definitions. One might wonder how well we can even claim to know what the cross-linguistic distribution is, and how one can go about constructing an explanation for something when one isn't even sure exactly what it is. Furthermore, I have declared my intentions to base my explanation in psychology, but the abstract syntactic nature of subject and object must seem like poor things to base psychological explanations on, at least compared to agent and patient. One feels relatively confident in assuming that at least in the majority of cases, any two people can agree on which is the agent and which the patient in some given situation, regardless of what their native language may be, and that these semantic roles are the kind of thing which may plausibly feature in some characterisation of the universal nature of human

thought. The same is plainly not true of subject and object. One also feels that the physical environment of reality contains universal properties which can be stated cleanly in terms of agent and patient. For example, if a sentence involves two noun phrases, one meaning "man" and the other meaning "stone", it is overwhelmingly more likely that the man is the agent and the stone the patient than vice versa, and this has been true everywhere on Earth for as long as humans have existed. As such, it seems at least plausible that human cognition in general and perhaps language in particular may have adapted in some way which reflects this and other similar universal properties. However, this statement in terms of agents and patients does not necessarily translate cleanly into a corresponding statement about subjects and objects. In this section I shall do my best to argue that things are not as bleak as they may seem.

To start with, I consider the question of how well we actually know what the cross-linguistic distribution of basic word orders is, given the fact that assignments of basic word orders to individual languages often comes down to a matter of arbitrary choice. Essentially, our certainty of the cross-linguistic distribution is related to the number of languages for which there is not wide-ranging agreement on a basic word order. The only systematic consideration of this question which I am aware is due to Newmeyer (1998), who takes a single sample of 174 languages as a starting point, and compares the basic word order assignments of the languages in this sample to assignments given to those same languages in a variety of other samples found in the literature. He concludes "the disconcerting fact, then, is that there is open controversy about the basic word order assignment of 18% of the languages", and adds that he suspects "a literature search will lead to this figure being elevated to well over 20%". This is a disconcerting fact indeed, although we should not lose sight of the fact that there is good agreement of the basic word order assignment of a strong majority of languages in the sample. In order to assess just how bad the 20% disagreement rate is, we need to consider just what sort of cross-linguistic distributions surveys are producing. This is the subject of a later section in this chapter, but for now it will suffice for me to say that all surveys which I am aware of, and certainly all of the most heavily cited surveys in the literature, suggest that SOV and SVO are overwhelmingly the most common basic word orders, together accounting for something like 85% of languages, split almost equally between these two orders. There are also some word orders which show up in surveys very rarely indeed, never accounting for more than 1 or 2% of the languages surveyed. Now, the majority of the disagreements which Newmeyer mentions deal with whether or not languages should be assigned a basic word order of SVO or SOV. There seems to be essentially no chance that, say, many of the languages which have been assigned one of these basic word orders should, in fact, have been assigned one of the extremely rare word orders. In the very worst case scenario, where absolutely all of the 20% of the languages under dispute turn out to be either SOV or SVO, the net effect of this will be to

make one of these two orders clearly more frequent than the other, but this is all. The fact that almost all languages are subject-initial will be unchanged, and the three extremely rare word orders would remain extremely rare. As such, it seems to me that there are interesting aspects of the cross-linguistic distribution of basic word order which we can be quite certain of, and as such there exists a clear explanatory target for this thesis to aim for.

Even if we accept the fact that there is a clear structure to the cross-linguistic distribution which is in want of an explanation, one could be forgiven for feeling uneasy about the fact that we don't know exactly what it is that we seek to explain the distribution of. Do we want to explain why some word orders are more likely to be the most frequently used word order in a language, or do we want to explain why some word orders are more likely to be associated with the least overall syntactic or morphological elaboration? Surely we need to know this before we can begin theorising? I think that, perhaps unexpectedly, we genuinely do not need to know. The fact that there exists such clear and significant structure in the cross-linguistic distribution of basic word order, no matter which of the multiple different criteria we use to define it, suggests that one of these criteria or some other criteria like them are identifying some concept of linguistic interest. Precisely what that concept is can in fact only be determined by an explanation provided for the structure! Suppose, for example, that it could be shown that a ranking of the basic word orders from least to most frequent precisely matched a ranking of the possible word orders according to how much syntactic or morphological elaboration is necessary on average to avoid ambiguity in declarative sentences, and we wished to claim this as an explanation of the cross-linguistic distribution of basic word order. In this case it would clearly be most appropriate to define basic word order with reference to the degree of syntactic or morphological elaboration. A different kind of explanation might suggest a different definition of basic word order as the most appropriate. The important thing to realise is that not only is there no reason to consider it problematic to begin searching for an explanation of basic word order distribution before deciding upon a precise definition of basic word order, there is actually no other principled way to proceed. A definition chosen in advance of an explanation would necessarily be arbitrary.

With regard to the apparent unsuitability of syntactic roles for use in psychological explanations, while semantic and syntactic roles are certainly not identifiable as one and the same, and are logically independent, it turns out that the mappings between them in the world's languages are not as arbitrary as they could in principle be. Agent and patient are significantly correlated with subject and object: that is to say, the subject of a sentence is significantly more likely to also be the agent of that sentence than it is to be the patient, and the object is significantly more likely to be the patient than the agent (see, e.g., (Tomlin, 1986)). Comrie (1981) suggests a significant link between subjects and agents, stating that "many facets of subjecthood can be understood by regarding the prototype of subject as the intersection of agent and topic".

As such, for most of the world's languages, any argument one may be able to formulate which suggests that putting the agent before the patient is somehow "better" than vice versa can double perfectly well as an argument that putting the subject before the verb is also "better" than vice versa. This is precisely the strategy that I shall use in this thesis, and this sort of strategy has been used before multiple times in the literature, apparently without drawing any criticism.

In fact, I would like to emphasise that, rather than considering an explanation for basic word order based on semantic roles as being a sort of "cheating", in that it deals with close correlates of the true subjects of interest, I am in fact in principle free to redefine the problem so that agent and patient are precisely what we are interested in. Suppose I define the "basic semantic role order" of a language to be the linear ordering of agent, patient and action which best characterises declarative sentences in that language, according to some criteria like those which have been used for basic word order; say, the order of agent, patient and action which is used most frequently. I can then construct large cross-linguistic surveys of basic semantic role order and look for interesting patterns in the resulting distribution. This would actually be a substantially easier task than constructing a cross-linguistic survey of basic word order, since semantic roles are much easier to identify across different languages than syntactic roles. I could explain any interesting patterns in the cross-linguistic distribution of basic semantic role order without making any references to subjects or objects at all, and this would be entirely appropriate. The patterns I would be explaining would, of course, be essentially the same patterns that have been found and studied for decades in basic word order distribution, under appropriate substitutions of semantic and syntactic roles, but who is to say that the semantic perspective is any less valid a way of looking at the data than the syntactic perspective? This is not something which should or can be decided *a priori*, but rather depends on the relative quality of the explanations which can be produced under the different perspectives. If the best explanation which can be found approaching the problem from the syntactic perspective is considerably more convincing than the best explanation which can be found approaching the problem from the semantic perspective, then we should consider basic word order to be the true parameter of interest, with the similar patterns found in the cross-linguistic distribution of basic semantic role order simply being a reflection of the "true" patterns in basic word order, mediated by the correlation between syntactic and semantic roles. And, of course, the opposite is also true. I am actually of the opinion that far more convincing explanations can be found from the semantic perspective, but for the sake of convention I shall refer throughout this thesis to basic word order and use the standard abbreviations of SVO, SOV, VSO, etc. I consider this to be a case of awkward terminology being fixed by historical accident, not entirely unlike, say, the distinction between conventional current and electron flow in physics and engineering.

## 2.2 Cross-linguistic distribution of basic word order

I shall now summarise several major cross-linguistic surveys of basic word order, focused on those surveys which have featured in attempts at *explaining* the observed frequencies. The different surveys considered differ in the number of languages considered, the particular languages considered, the manner in which those particular languages were sampled, and the method by which each language's basic word order was determined. My motivation for presenting data from such a wide range of surveys is to demonstrate that despite the many methodological difficulties involved in attempting to measure the cross-linguistic distribution of basic word order, there is a clear structure to distribution which can be considered universally agreed upon. Summary figures of all the surveys are presented in Table 2.2.1.

### 2.2.1 On language sampling

The task of sampling languages in order to construct a large language survey for the purpose of investigating the cross-linguistic distribution of some language feature, such as basic word order, is far from straightforward. By no means is it simply a matter of "the more languages included the better". In addition to the genuine universal preferences that such a survey aims to uncover, the distribution of language features can also be substantially influenced by two other factors, namely genetic and areal relatedness amongst languages.

It is widely accepted that different languages are derived from different changes to common ancestral languages (I shall discuss this in much more detail in Chapter 3), so that it makes sense to speak of languages belonging to "families", in direct analogy to biological families. Two languages are said to be *genetically related* if both have a common ancestor. Unsurprisingly, genetically related languages tend to display greater similarity to one another than to unrelated languages. For example, English, German and Dutch are all quite closely genetically related (they belong to the Germanic language family), and as such there are distinct similarities between them, for instance at the levels of the lexicon (e.g. book/Buch/boek, mother/Mutter/moeder, cat/Katze/kat) and syntax (e.g. all three languages are prepositional and in all three languages adjectives preceed nouns (Dryer, 2008)). If we were to randomly sample 10 languages from the Germanic family and attempt to infer from this sample any universals of language, we would draw several spurious conclusions, concluding that features which are very common to Germanic languages are in fact common to all languages, when this may not be the case (for example, it is significantly more common cross-linguistically for nouns to preceed ajdectives, but the opposite is normal in the Germanic family). In order to accurately measure cross-linguistic distributions of features, one should strive to sample languages so that each family is represented in proportion to its actual size.

Only about 1.5% of the world's languages belong to the Germanic family, and as such in a sample of, say, 100 languages, only one or two Germanic languages ought to be included.

In addition to languages bearing significant similarities due to genetic relatedness, languages can also be similar to one another due to *areal relatedness*. Two languages which are spoken in geographically nearby areas may come to share similarities even if they are genetically unrelated, or only distantly related, because some language features are transferred from one language to another (or "borrowed" by one language) by means of language contact. A *linguistic area* is a region in which it is generally accepted that significant quantities of borrowing have taken place between the languages in that area. The probability of borrowing between two languages generally increases as geographic distance decreases, although geographical considerations such as mountain ranges or rivers, as well as cultural or political considerations can act against areal transfer of language features. In the same way that a language survey for the purposes of uncovering universals should attempt to balance its sampling across language families, such a survey should also attempt to balance its sampling across language areas.

In this section I shall consider language surveys which have been built using a range of sampling techniques, which I group into three categories: convenience sampling, diversity sampling, and proportional sampling. Convenience samples, as the name suggests, include simply those languages which are convenient for the relevant author to employ at the time, usually for reasons of personal familiarity with the languages or the easy availability of reliable data on the languages in the literature. Convenience samples tend to be unduly dominated by languages from the Indo-European family (which account for only 6 or 7% of languages), since this family is the most widely studied by academic linguistics. These samples therefore do a relatively poor job of either capturing the overall diversity of language on any particular dimension, or in balancing languages across family groups and thus accurately reflecting statistical biases in language variation. Diversity samples are constructed primarily to capture the extent of language diversity on some number of dimensions, and thus tend to include a number of relatively exotic languages where these are required to demonstrate that some rare set of features is actually attested. As such, diversity samples, like convenience samples, tend to be unbalanced with respect to genetic and areal relatedness. Finally, proportional samples attempt to ensure that the number of languages in the sample belonging to a particular language family or language area is roughly proportional to the actual number of languages in that family or area. Clearly, proportional samples are those on which claims of linguistic universals should ideally be based. However, due to the significantly increased time and effort required for their construction compared to convenience or diversity samples, this is not often the case in practice. Note that efforts to survey *all* of the world's languages will, as the sample size grows very large, tend toward producing samples which could classify as both

Table 2.2.1: Relative frequency of basic word orders according to prominent surveys

| Survey | Size | % SOV | % SVO | % VSO | % VOS | % OVS | % OSV |
|---|---|---|---|---|---|---|---|
| (Greenberg, 1963) | 142 | 45.1 % | 36.6 % | 16.9 % | 0.0 % | 0.0 % | 0.0 % |
| (Ruhlen, 1975) | 435 | 51.0 % | 35.6 % | 10.8 % | 1.8 % | 0.5 % | 0.2 % |
| (J. Hawkins, 1983) | 336 | 51.8 % | 32.4 % | 13.4 % | 2.4 % | 0.0 % | 0.0 % |
| (Tomlin, 1986) | 402 | 44.8 % | 41.7 % | 9.2 % | 3.0 % | 1.2 % | 0.0 % |
| (Dryer, 2008) | 1057 | 47.0 % | 41.3 % | 8.0 % | 2.5 % | 0.9 % | 0.4 % |

diversity and proportional samples.

### 2.2.2 Greenberg

Modern word order typology is generally considered to have begun with the seminal work of Greenberg (1963), which is the starting point of most recent interest in word order universals. Many of the universals proposed in this paper were inferred from a very small survey of 30 languages, but the paper's appendix actually contains a larger (though still quite small as far as surveys go) sample of 142 languages, and basic word order is provided for each of them. Greenberg states that the languages in the smaller survey were chosen for convenience. No discussion is provided as to how the additional languages for the larger survey were collected, although it seems best to consider a larger convenience sample by default, since the fact that only half of the logically possible basic word orders are attested means it cannot qualify as a diversity sample, and does not appear to have been deliberately constructed to be proportional. It is worth noting, though, that for a convenience survey, especially considering its small size, the genetic and areal range are fairly extensive. SOV is by far the most frequent word order in the sample, at 45%, followed by SVO at 37% and then VSO at just 17%. The VOS, OVS and OSV word orders are completely unattested in Greenberg's survey, which is entirely typical for surveys of the time.

Greenberg's survey distribution: SOV (45%) > SVO (37%) > VSO (17%) > VOS (0%) = OVS (0%) = OSV (0%)

### 2.2.3 Ruhlen

Ruhlen (1975) presents a survey of approximately 700 languages, with basic word order data for 435 of them (I am using Manning and Parker (1989)'s breakdown of Ruhlen's survey). This survey is noteworthy for the fact that it contains at least one language for all six of the logically possible basic word orders, which at its time of publication was very rare (the next two surveys which I will consider, from 8 and 11 years later, do not have this property). No discussion is provided on how the languages used were selected. However

since the survey is presented as part of a "guide to the languages of the world", and since it contains representatives of all six of the logically possible basic word orders, it seems appropriate to consider it as closer to the ideal of a diversity survey than either a convenience or proportional survey. While VOS and the object-initial languages are all attested in Ruhlen's sample, SOV, SVO and VSO still dominate the sample, and in the same order as in Greenberg's survey, at 51%, 36% and 10% respectively.

Ruhlen's survey distribution: SOV (51%) > SVO (36%) > VSO (10%) > VOS (2%) = OVS (0%) = OSV (0%)

### 2.2.4 Hawkins

J. Hawkins (1983) constructs a survey by taking (Greenberg, 1963)'s 142 language survey and then extending this to contain 336 languages in total. Hawkins explains that the survey is "a convenience sample, with the choices of languages reflecting the interests and expertise of the contributors, the availability of sources of information and the original Greenberg sample", but clarifies that "contributers were encouraged not to work on the same language family or group wherever possible, in order to maximize genetic diversity". The lack of object-initial languages in the survey is interesting, considering these languages being documented at the time of publication (in, e.g., Ruhlen's sample). There are, however, VOS languages in the sample, so it does capture more variety than Greenberg's original sample. The relative frequencies of the basic word orders in Hawkins' survey match those in Ruhlen's quite closely, with SOV, SVO and VSO dominating heavily.

Hawkins' survey distribution: SOV (52%) > SVO (32%) > VSO (13%) > VOS (2%) = OVS (0%) = OSV (0%)

### 2.2.5 Tomlin

Tomlin (1986) presents a carefully constructed proportional survey, featuring 402 languages. These 402 languages were sampled from a larger database of basic word order on 1063 languages which Tomlin constructed from a variety of smaller samples. The particular 402 language sample was produced by repeatedly drawing 402 languages in an independent, random manner from the larger database, and testing each 402 language sample for significant genetic or areal biases. This process was repeated until a sample without any such biases was produced. Testing for biases was conducted using the Kolmogorov goodness-of-fit test, based on genetic and areal information provided by (Voegelin & Voegelin, 1977). As such, Tomlin's survey represents a very careful use of proportional sampling, making it a significantly more suitable survey for present purposes than Greenberg's, Ruhlen's or Hawkins'.

Tomlin's survey features languages attesting all six of the possible basic word orders except for OSV. Similarly to earlier surveys, the subject-initial

orders dominate strongly, with VSO being by far the next most frequent and VOS and OVS only slightly attested. Interestingly, and unlike any other survey author I am familiar with, Tomlin tests the differences between the number of languages belonging to each word order for statistical significance. On the basis of this, he concludes that the often claimed relative rankings of SOV > SVO and VOS > OVS are unwarranted: while the absolute numbers in his survey are compatible with these rankings, the differences in number are inadequate to reject the null hypothesis that SOV = SVO and VOS = OVS. Tomlin thus proposes the ranking SOV = SVO > VSO > VOS = OVS > OSV. Dryer (1989) suggests that areal bias still exists in Tomlin's sample, due to the existence of very large linguistic areas (areas spanning entire continents), and that this bias has caused SOV languages to be underpreresented and SVO languages to be overrepresented in the sample. He thus argues that the SOV > SVO ranking should be considered genuine.

Tomlin's survey distribution: SOV (45%) = SVO (42%) > VSO (9%) > VOS (3%) = OVS (1%) > OSV (0%)

### 2.2.6 Dryer

The final survey I shall consider, (Dryer, 2008), is a large diversity sample consisting of 1057 languages. Since this survey is part of an attempt at cataloging all languages (the World Atlas of Language Structures), we can assume it to be relatively close to proportional, at least closer than the samples of Greenberg, Ruhlen and Hawkins. Dryer's survey is the only one of those I have considered here in which all six of the logically possible orders are attested, with OSV languages appearing for the first time, although in extremely low numbers. As per all the previous surveys three word orders, SOV, SVO and VSO dominate strongly, together accounting for 97% of languages, with SOV and SVO dominating at 88%.

Dryer's survey distribution: SOV (47%) > SVO (41%) > VSO (8%) > VOS (2%) > OVS (1%) > OSV (0%)

### 2.2.7 Summary

Although the exact relative frequencies of different basic word orders can be seen to vary across surveys, in some cases by up to as much as 10%, it is clear that when it comes to *ranking* the different basic word orders by frequency, there is wide agreement. Every survey included above is consistent with the ranking SOV > SVO > VSO > VOS ≥ (OVS, OSV). The relative ranking of OVS and OSV is the only place that there is any disagreement. The most commonly claimed relative ranking is OVS > OSV, and indeed the ranking SOV > SVO > VSO > VOS > OVS > OSV has generally been taken as the "correct" complete ranking in the literature. Accordingly, I shall take this ranking as correct in this thesis, but I will not push the OVS > OSV part of the

ranking very hard at all. I remain entirely open to the possibility that in fact OSV > OVS, and I don't believe that any subsequent discovery of evidence in support of this will negate any explanations for the cross-linguistic distribution of basic word order which I will develop in this thesis.

In addition to this very strong qualitative agreement across surveys, there is also a considerable degree of quantitative consistency. In every case, the frequencies of the subject-first word orders, SOV and SVO, are many times greater than the frequencies of any other word orders, and the frequency of VSO is many times greater than the frequency of any of the word orders which occur less frequently than it. So, using the symbol $\gg$ to denote "is much more frequent than", we can write SOV > SVO $\gg$ VSO $\gg$ VOS > OVS > OSV.

There seems to be a degree of "internal consistency" within this ranking which exceeds what one might expect if selecting a ranking at random. Note, for instance, that subject-first word orders strictly outnumber verb-first word orders, which strictly outnumber object-first word orders. Also, word orders in which subject precedes object strictly outnumber word orders in which object precedes subject. Regularities like these lead to an intuition that the highly non-uniform distribution of basic word orders is not merely some sort of historical accident, but that in fact there are genuinely interesting, and perhaps deeply informative, principles underlying the distribution.

## 2.3 What we need to explain

We have seen in the previous section that, despite the inherent difficulties in defining a basic word order for every language and the methodological challenges involved in accurately sampling the world's languages, there is in fact good agreement amongst linguists that the following is an accurate ranking of basic word orders by frequency: SOV > SVO $\gg$ VSO $\gg$ VOS > OVS > OSV. It is this ranking which I seek to explain in this thesis. But what exactly would it mean to explain this?

Of all the researchers to have examined the problem of basic word order frequency previously, Tomlin appears to be the only one to have given this question any consideration. He establishes the following as the "minimum requirements on an explanation of the distribution of basic constitutent order types":

1. Identify those constituent orders which occur frequently.

2. Identify those constituent orders which occur infrequently.

In other words, if absolutely nothing else, any worthwhile explanation of the cross-linguistic distribution of basic word orders must be able to explain what it is about SOV, SVO and VSO word orders which differentiates them from VOS, OVS and OSV word orders with regard to frequency of occurence.

Obviously, there is rather more that we would like an explanation to say than this. Tomlin offers the following as "additional requirements on the explanation":

1. Why do subject-initial languages outnumber verb-initial languages?
2. Why do object-initial languages occur so much less frequently than do subject-initial or verb-initial languages?
3. Why do VSO languages outnumber VOS?
4. Why do VOS and OVS languages outnumber OSV?
5. Why do SOV languages outnumber VSO?
6. Why do SVO languages outnumber VSO?
7. Why are VOS and OVS languages of approximately equal frequency?
8. Why are SOV and SVO languages of approximately equal frequency?

In addition to meeting the minimum requirements outlined above, an explanation should meet as many of these additional requirements as possible, and the more the better. In Chapter 12 I shall revisit each of these questions and provide answers to as many of them as I can based on the overall explanation I shall develop throughout the thesis.

## 2.4 Summary

The concept of basic word order is substantially more difficult to give a satisfactory definition for than one might first imagine. In particular, there is no agreed upon general definition of subject or object which holds across language, and there are languages in which different clauses of an utterance use different word orders, with no agreement on which clause's word order should be considered basic. Despite this scope for disagreement on which basic word order to assign various languages, we have seen that, across language surveys, there is strong agreement on the ranking of basic word orders by frequency, the consensus ranking being SOV > SVO $\gg$ VSO $\gg$ VOS > OVS > OSV. So, with regard to basic word order, there is an agreed-upon and interesting question which needs answering: why this ranking rather than any other? While the syntactic roles of subject and object are not perfectly well-defined, and are of an abstract and purely syntactic nature, they correlate reliably with the semantic roles of agent and patient, which are well-defined, independent of language and objectively the same for all observers. Thus, it seems at least plausible that we may be able to base an answer to the question under consideration on psychology.

# Chapter 3

# Explanations for the different frequencies

## 3.1 Explaining language diversity in general

Before I begin working toward an explanation of basic word order distribution, I want to take some time to consider exactly what constitutes a valid explanation for language universals. Furthermore, although this thesis is concerned only with explaining the cross-linguistic frequencies of basic word order, I want to begin this chapter with a discussion of how one might go about explaining linguistic variation *in general*. The problem of interest for this thesis is simply a special case of this general problem. My motivation for first considering the more general problem is to facilitate careful consideration of exactly what kinds of explanations should be considered valid candidates for the particular problem later in this work.

Estimates of the total number of languages spoken in the world today vary. It is highly unlikely that a single authoritative count will ever be established, for various reasons, including the fact that there is no universally accepted definition of the difference between languages and dialects. Nevertheless, we can get some idea of the scale of linguistic diversity on Earth. Ruhlen (1975) states that he "think[s] most linguists today would agree that the number of distinct human languges spoken on the earth at the present time [i.e. in 1975] is probably on the order of 4,000 to 8,000". The most extensive language survey that I am aware of, the Summer Insitute of Linguistiscs' *Ethnologue*, contains 6,912 languages in its latest edition (Lewis, 2009). Krauss (2007) considers this count "on the high side", and suggests instead 6,000, a figure right in the middle of Ruhlen's broad estimate. So around 6,000 languages are spoken today[1]. These thousands of languages display great diversity in

[1]I note in passing that, whatever the true number, it is expected to decrease dramatically in the coming century and even decades: Krauss (2007) estimates that only around 300 languages (5% of the global total at best) can be considered quite probably safe from going extinct by 2100. In addition to the obvious socio-cultural implications of this situation, it

all of their characteristics. The idea of people in different parts of the world speaking different languages is so familiar and obvious to most of us that it is quite easy to go an entire lifetime without considering the question of *why* this should be the case. Despite its total familiarity, the situation is in fact a surprising one. After all, most people would agree that children basically learn the same language as the adults which surround them early in life, and that each person then continues to use basically that same language unchanged for the remainder of their life. However, the logically expected result of this state of affairs is a perfectly static linguistic landscape. Everyone alive today should be speaking one of the finite (and presumably fewer than 6,000) number of languages which were spoken by the very first members of the species to have acquired the capacity for language. Very clearly this is not the case, and so the field of linguistics is obliged to provide some explanation of why languages change over time and why that change leads to a wide diversity of languages, rather than to, say, the coalescence of many languages into a few.

Discussing this problem in generality allows me to follow the excellent account given in (Nettle, 1999). My discussion of linguistic variation is necessarily much more limited than that given by Nettle, and the interested reader is directed to that reference for more detail.

### 3.1.1 Nettle's framework

The problem for linguistics of explaining the wide diversity of languages spoken across the world is not entirely unlike the problem for biology of explaining the vast diversity of species of organism on Earth. In that second case there is now a very widely accepted explanation for the diversity, namely the celebrated theory of evolution (Darwin, 1859).

The essence of this explanation is that while organisms can inherit certain physiological and behavioural traits from their parents, the inheritance is not perfect. Random variation in traits occurs naturally, so that occasionally organisms are born which differ from their parent organism in some trait(s). The primary mechanisms for this random variation are the chance occurence of random mutations in an organism's genome and the random shuffling together of two genomes which occurs during sexual reproduction. Random variation in traits is not by itself sufficient to explain the observed diversity of species: alone it leads only to universal heterogeneity between organisms, rather than the observed clustering into species. Speciation requires some mechanism by which variation may be either amplified or suppressed.

This amplification and suppression is provided primarily by the process of *natural selection*, wherein variation which causes an increase or decrease in

is also certainly troubling for language typology. As we saw in the previous chapter, the largest basic word order surveys today cover around 1000 languages, or 17% of the current estimated total. The window of time in which these samples could be improved to include data on the majority of languages is closing rapidly.

the probable reproductive success of an organism (either directly or indirectly by aiding survival) is correspondingly more or less likely to be inherited by future generations. Because there are very many different environmental and ecological niches which organisms may inhabit, and because traits which are beneficial for survival in one may be deleterious in another, natural selection will favour different variations in different situations. The tendency of the process is thus not to push all lifeforms toward a single optimal form, but rather to differentiate life into a vast collection of specialised species, each optimised for its particular niche. Biological variation is also amplified and suppressed, to some degree, by *sexual selection*, wherein variation which makes an organism more or less attractive to prospective mates is correspondingly more or less likely to be inherited.

So we see that the diversity of life is well explained by two essential mechanisms: *variation* in the traits of organisms during reproduction, and forces of *selection* acting on this variation. It is generally accepted that a very similar explanation holds for language diversity. That is, random variation in languages occurs naturally as languages are transferred from one generation of speakers to the next, and this variation is either amplified or suppressed by some mechanism.

There are two possible sources of variation in language, which play roles analogous to genetic mutation in biology. They are *imperfect production* and *imperfect learning*. Adult speakers will occasionally make mistakes in using their language, and these mistakes become part of the data from which members of the next generation learn the language (the so-called *primary lingusitic data*, or PLD). Similarly, language learning children will occasionally make mistakes in inferring the rules of their language on the basis of the PLD produced by the previous generation, and thus will come to themselves produce different input for the next generation. As per the biological case, this random variation is not by itself sufficient to explain the proliferation of distinct languges, as opposed to widespread heterogeneity: we need some selective mechanism or mechanisms to amplify and suppress different variations.

One plausible candidate for such a mechanism is what Nettle terms *functional selection*, which is entirely analagous to natural selection in biology. Functional selection holds that those linguistic variations which result in such things as a language being easier to learn, to speak, or to understand will be passed to future generations of speakers preferentially over those variations which make a language harder to use. In other words, languages are selected differentially according to how well they fulfill their presumed function of enabling efficient, reliable and expressive communication of thoughts between speakers. The presumed bases for any asymmetries in ease of acquisition, production and comprehension across linguistic variants are generally taken to be either physiological or cognitive in nature. For example, how easy a language is to learn may depend on how systematic it is, or how in line it is with any learning biases which may exist in whatever inductive machinery is used to

acquire language[2]. How easy a language is to speak may depend on how much energy and/or coordination is required to operate the vocal cords and tongue when speaking it. How easy a language is to understand may depend on such things as how much working memory is required to process a sentence, or how easy it is to make an utterance unambiguous.

Linguistic variation may also be directed by what Nettle terms *social selection*, which is entirely analogous to sexual selection in biology. Social selection holds that those linguistic variations which are characteristic of members of a language community who are held in some special social esteem will be passed to future generations of speakers more readily than other variations, due to language learner's preferential imitation of esteemed speakers.

At this point of our exposition, having identified mechanisms for lingustic variation and for selection on that variation, we have still not completely provided an explanation for lingustic diversity. Recall that in the biological case, an essential part of the explanation was the ecological diversity which organisms inhabit. This fact explains why natural selection favours the diversification of life into different species filling different niches, rather than convergence of all of life into a single universally optimal organism. At first glance, a corresponding situation seems to be missing in the linguistic situation. The very same cognitive, articulative and perceptory machinery is available to all humans, so it appears that functional selection should push all languages in the same direction, toward a single optimal language. This is obviously not what happens, and so we must account for the difference. The solution proposed by Nettle involves consideration of the fact that linguistic variation happens across many dimensions: changes to a language may, for instance, be phonological, morphological or syntactic, and there are many possible changes for each of these options, for instance a syntactic change may involve altering the basic word order of the language or it may involve changing the relative order of nouns and adjectives which modify them. Whether or not functional selection favours a particular change along one dimension may depend crucially upon the current state of the language along the *other* dimensions at the time that the variation is first introduced. Thus, the multidimensional nature of the problem provides a sort of "virtual ecosystem", so that functional selection may favour a given variation in some languages but not in others[3]. As long as the functionality landscape contains multiple local optima, this state of affairs is compatible with the diversification of a single or small number of starting language(s) into a wide range of languages, each occupying the area surrounding a different optimum.

[2]The precise nature of this machinery is perhaps the most hotly debated subject in all of linguistics. The account of language diversity presented here is, however, entirely agnostic on the matter.

[3]The same is true, of course, in the biological case, although it is not strictly necessary there to explain biological diversity, as the actual ecosystem suffices.

### 3.1.2 Justifying a focus on functional explanations for basic word order

We have seen how linguistic variation due to imperfect production and learning, amplification of that variation by functional and social selection, and the existence of a virtual ecosystem caused by simultaneous variation along many dimensions, can together account for linguistic diversity in general and thus, in principle, for the cross-linguistic distribution of basic word order in particular. For the remainder of this thesis, I shall focus my attention on *functional* explanations for basic word order frequencies, giving no further consideration to social selection, and before proceeding I want to first justify this decision.

Why look to functional and not social selection to explain basic word order? I do not deny either the existence nor the significance of social selection in amplifying language variation and hence contributing to language change. In fact, I am quite open to the possibility that social selection may generally exert *more* influence on language change than functional selection (this seems especially likely to me with regard to lexical change). However, I am skeptical that social selection can explain *systematic* or *universal* effects in language change. If some language community should contain a person or group of people who are esteemed in the community for their strength, wisdom or beauty, and if that person or group should have a tendency to use some word order, say SVO, more often than is typical for the language, it is entirely plausible that the prominence of SVO word order may increase in the community due to social selection. However, it seems to me entirely implausible that, across languages and across cultures, it should turn out that strong, wise or beautiful people should consistently prefer, say, SVO over VSO word order, more so than any other speaker of that language. Rather, word order preference should vary fairly randomly with social status, so that the overall effect of social selection, cross-linguistically, should be to impose a certain degree of "random noise" on top of any overall structure. Where there *is* a systematic preference for some word order over another, across languages and across cultures, then the underlying cause must surely be functional and not social in natures.

I have justified focusing on functional selection over social selection. Does this represent an exhaustive consideration of explanatory options? Almost. There remains one other logical possibility, and that is that the cross-linguistic word order distribution is to be explained on what are commonly called formalist grounds. The word "formalist" here is derived from the word "form", as in structure, not the word "formal". A formalist explanation for some phenomena explains the phenomena in terms of more general linguistic principles, stated entirely in syntactic terms without any reference to communicative function, processing efficiency, etc. A certain caricature of the formalist-functionalist distinction is unfortunately widespread in linguistics. This view holds the two as fundamentally incompatible, diametrically opposed opposites: formalists supposedly believing that language-external considerations, including functional

considerations, are completely unable to influence the form of language, which is dictated by biologically innate principles of an autonomous faculty of syntax; functionalists supposedly denying the existence of this faculty outright and insist that functional principles alone are fully able to explain all facts about language form. Caricature functionalists espouse the idea that only functionalist explanations are "true" explanations, with formal explanations being entirely vacuous. Unfortunately, a great many *actual* linguists do not fall particularly far from these characterisations. Newmeyer (1998) does an excellent job of thoroughly dispelling the myriad misconceptions embodied by this falsely diachotomous view, showing in particular that the autonomy of syntax is not incompatible with functionalist explanation, that it does not logically require biological innateness, and that both formal and functional explanations can be "true" explanations. Along similar lines, Kirby has demonstrated that functional explanations and innateness explanations can not only co-exist but can be ultimately dependent upon one another (Kirby & Hurford, 1997; Kirby, 1999, 2000).

I cannot and do not claim to rule out *a priori* the possibility that basic word order frequencies are best explained formally. However, I am equally unable to make any kind of case for assuming *a priori* that the phenomenon must *necessarily* be formally explained. To me, the most prudent approach is to assume the existence of a functional motivation as a null hypothesis. If an earnest search unveils no promising candidates, only then does it make sense to posit an arbitrary formal explanation for any given phenomena. This approach ensures that, where functional motivations for the form of language *do* exist (and certainly their existence is possible in principle, whatever you believe about the nature of the language faculty), there is some hope of uncovering them. The alternative approach, where all phenomena are taken by default to be explained on purely formal grounds, will necessarily fail to find them. A similar sentiment has been expressed by J. Hawkins (1983): "innatenes is a residue, to be invoked when other, more readily observable, explanatory principles fail", and also by Newmeyer (1998): "I certainly have no objection to appeals to innateness in the abstract, nor do i feel that there is something in principle suspect about hypothesising that a particular feature of language exists because it is innate. If the evidence leads in that direction, well and good. But often there are plausible explanations for a typological pattern that do not involve appeal to an innate UG (universal grammar) principle. In such cases, harm *is* done by assuming innateness. What we would have are two contrasting explanans: one that says that the pattern results from such-and-such motivated principle or force, the other that says it is merely a genetic quirk. All other things being equal, we should choose the former".

In short, I believe that social selection is unable to explain broad trends in language, and I believe that arbitrary formal explanations should be used only as a last resort. As such, in this thesis I restrict my attention to functional explanations for basic word order frequencies.

### 3.1.3 Clarifying what constitutes functionality

The definition of "functionality" which emerges from Nettle's characterisation of language change is by no means a standard or even common definition of the term. Indeed, the concept of functionality in linguistics dates back at least to the so-called "Prague school" of the 1930s, predating Nettle's framework by decades. However, so far as I can tell, there *is* no standard or common definition of functionality which is both precise enough to serve as a respectable definition and general enough to apply to all of the ideas which have made their way into the literature under the banner of functionality. I am fond of Nettle's formulation mainly because it explicitly embeds functionalism into an evolutionary model of language change, and thus provides some kind of causal mechanism by which functional pressures come to actually shape language diversity (something which seems to be often taken for granted in the functionality literature). However, somewhat unsatisfyingly, it is very much an *implicit* definition: functionality is anything non-social which facilitates either the amplification or the suppression of a linguistic variation. I have no qualms about using Nettle's definition throughout this thesis, due to the lack of any servicable alternative. However, because the issue of functionality is such an essential concept in this work, and because the term can mean quite different things to different people, I feel some obligation to make more explicit precisely what I will consider to be "functional".

As I understand things, functionalism can essentially be factored into two more or less distinct traditions, one older than the other, which both claim to study how the form of language facilitates language use, but go about this in quite different ways: or at least, ways which seem quite different at first glance. An analogy due to Hurford (1990) will be useful here: consider a spade. There are multiple aspects of a spade's design which are obviously the way they are for the sake of the spade's usability. Some of these aspects have to do with the essential nature of the task of digging holes, for example the fact that the end of the spade is pointed, the fact that it is thin, the fact that it is concave, and the fact that it is made of a durable material such as steel, etc. Other aspects have less to do with the essential nature of digging and more to do with the practical problems raised by the fact that the spade is designed to be used by humans, for example the fact that the shaft of the spade is roughly as long as the distance that an average sized human's hands are above the ground, the fact that the diameter of the shaft is such that it can be gripped tightly in an average sized human's hands, etc. Both of these sorts of considerations are of essential importance in determining how well the spade functions overall as a tool to help humans dig holes, but are not essential for the task of digging holes itself. Modern automated machines for digging holes, for instance, still feature pointed, thin, concave steel components, and it is hard to imagine a digging machine without these, but in general they do not have human-sized shafts attached to them. Along similar lines, functional approaches to explaining

linguistic universals come in two kinds: those relating to what is taken to be the essential nature of the task of communicating information between agents, and those relating to the fact that languages are systems for communication which are implemented by the particular neural and physiological hardware that humans have at their disposal.

The first of these two schools, that associated with the fundamental task of communication, is the older of the two, so that this could be claimed to be the "traditional" meaning of functionality. The second school is associated essentially with the implementation of language in the brain rather than in the abstract, and so I shall call it "cognitive" as opposed to traditional functionalism. Cognitive functionalism is more recent and somewhat less accepted than traditional functionalism. In the sections that follow I shall discuss what have been some of the most important considerations in these two schools, and ultimately argue that the schools as they have been characterised thus far are actually not as distinct as may seem at first. Specifically, I shall argue that most of the first school rests on very flimsy ontological grounds, and is in fact best understood by reinterpreting it in such a way that it actually belongs quite squarely to the second school. This is not, however, to deny the distinction between the two kinds of functionalism. I believe this to be genuine, and indeed in Part III of the thesis I shall develop an account of word order functionality which belongs quite clearly to the older tradition, in that it focuses on what is required to facilitate communication over a serial channel in the abstract, with no consideration of the specifics of embedding this communication within the human mind.

### Traditional functionalism

The traditional school of functionalism seems to have had two major focuses, these being *iconicity functionalism* and *discourse functionalism*. I provide a very brief overview of these focuses below.

The following discussion of iconicity is based on Newmeyer (1992), to which I direct interested readers for a more thorough account. Strictly speaking, the examples I shall mention here are examples of what Newmeyer calls *structure-concept iconicity*: I believe that this sub-class of iconicity has been the most influential historically. Within this school of thought, an item of language, such as a word, phrase or sentence is considered to be iconic to the extent that there is some non-arbitrary relationship between its structure and the conceptual structure of what it represents, i.e. the structure of its meaning. Structure-concept iconicity comes in many varieties. Briefly, Newmeyer identifies *iconicity of distance*, whereby "the linguistic distance between expressions corresponds to the conceptual distance between them"; *iconicity of independence*, whereby "the linguistic separateness of an expression corresponds to the conceptual independence of the object or event which it represents"; *iconicity of order*, whereby "'the order of elements in language parallels that in phys-

ical experience or the order of knowledge"; *iconicity of complexity*, whereby "linguistic complexity reflects conceptual complexity"; and *iconicity of categorisation*, whereby "concepts that fall into the same grammatical category tend to be cognitively similar". In short, iconicity functionalism holds that the surface form of language is such that it reflects the underlying conceptual form of what the language represents.

Whereas iconicity functionalism focuses on the structure of what is being talked about using language, discourse functionalism focuses instead on the structure of the communicative process itself. Discourse functionalism revolves around an extensive core of specialist terminology which it is well beyond my present scope to summarise. I shall, however, do my best to give a general flavour of the overall approach. Discourse functionalism is concerned with such facts as when language used for conversation, there exists some thing or things which the conversation is "about": things which all participants in the conversation share an awareness of as common knowledge and are relating the content of the conversation to; these things are called, variously, *topics* or *themes*. Similarly, there are parts of a conversation which are not themselves the topic or theme of conversation but are rather statements *about* the topic of conversation; these things are called, variously, *comments* or *rhemes*. Furthemore, the various constituents of a sentence can be divided into "new information" and "old information", depending on whether or not they refer to concepts which have occurred previously in the conversation, and so are part of the assumed context which all participants are aware of. The new information in a sentence is sometimes called that sentence's *focus*. Topics, themes, comments, rhemes and focii are examples of *pragmatic roles*. Central to discourse functionalsm are claims that certain orderings of different pragmatic roles are more natural than others, or more conducive to communication. For instance, it is typically taken that it is best for topics (themes) to precede comments (rhemes), and that old information should precede new information.

So far as I can tell, iconicity and discourse functionalism make up the substantial bulk of the traditional school of functionalism. Returning to Hurford's spade analogy from before, this is the school of functionalism which, applied to spades, would focus on the shape and physical properties of the blade which make it well-suited to digging, rather than the properties of the shaft and handle which make it a convenient tool for humans to dig with.

### Cognitive functionalism

More recently, the meaning of functionalism has expanded to include considerations of factors which, while not fundamentally *about* communication in the same way that iconicity or discourse considerations are considered to be, are still relevant to communication between humans using the human language faculty. This newer construal of functionalism includes concerns of a more "psycho-mechanical" nature, such as minimising working memory re-

quirements, information processing effort or muscular effort in articulation. I take it that these sorts of concerns are rather more intuitive to grasp, especially for non-linguists, than those of iconicity and discourse functionalism, so I shall not take the time here to present more detailed examples. Suffice it to say that cognitive functionalism is concerned with "implementation details" that result from embedding language within the particular environment of human cognition, rather than analysis of the abstract requirements of communication in general. In our ongoing spade analogy, this is the school which would focus on the shape and size of the shaft and handle.

It is worth noting that cognitive functionalism has not always been (and perhaps still is not) accepted by all linguists as "real functionalism". Writing about a workshop on the topic of "explanations for language universals", Hyman (1984) says:

> While everyone would agree that explanations in terms of communication and the nature of discourse are functional, it became apparent...that explanations in terms of cognition, the nature of the brain, etc. are considered functional by some but not by other linguists. The distinction appears to be that cognitive or psycholinguistic explanations involve formal operations that the human mind can vs. cannot accommodate or 'likes' vs. 'does not like', etc., while pragmatic or sociolinguistic explanations involve operations that a human society or individual within a society can vs. cannot accommodate or likes vs. does not like. Since psycholinguistic properties are believed to reflect the genetic makeup of man, while sociolinguistic properties reflect man's interaction with the external world, some definitions hold that only the latter truly has to do with 'function'. In the other view, any explanation which relates grammar to anything other than grammar is 'functional'.

### Uniting the two schools

In this section I shall argue that the two schools introduced above are not as distinct as they have been made out to be and, in fact, while it is the newer of the two schools of functionalism, cognitive functionalism is much closer to being the "one true functionalism" than traditional functionalism is. Some terminology developed by Hauser, Chomsky, and Fitch (2002) (hereafter HCF) has become popular in the literature and will be useful for this purpose[4]. HCF distinguish between the faculty of language in a narrow sense, or FLN, which consists of "the abstract linguistic computational system alone", and the faculty of language in a broad sense, or FLB, which is the union of the FLN's abstract computational system with other systems which are not specific to language

[4]I shall also use this terminology occasionally throughout the rest of the thesis. I wish to stress that, by adopting HCF's terminology, I am *not* accepting their hypothesis from the same paper that the human FLN consists purely of a facility for recursion.

but which are nonetheless a necessary part of the overall language faculty. The FLB includes but is not necessarily limited to a sensory-motor system (SMS), consisting of those physiological and neurological capabilities involved in hearing and speaking language, and a conceptual-intentional system (CIS), which is used to represent the meaning of utterances. Note that aspects of the SMS and CIS are permitted to be - and presumably are - utilised for non-linguistic purposes as well. For instance, the CIS is presumably also used to support reasoning about the sorts of things which can be the meanings of utterances.

What I consider to be one of the greatest problems with the traditional school of functionalism is that it has been presented (to be fair, some authors are less guilty of this than others, or entirely innocent) as being quite separate from cognition and the human mind and/or brain. This strikes me to be entirely inappropriate. It seems to me that the fundamental concepts in traditional functionalism - the conceptual structure of language referents and the flow of information in discourse - are entirely irrelevant to human language if they are considered as Platonic ideals, and can be made components of a convincing functional account of language only if they are recast in cognitive terminology.

Take, to begin with, the conceptual structure of meaning, which iconicity functionality holds should be reflected in the structure of language. Suppose that a linguist assigns the meaning of a given sentence some particular conceptual structure, breaking it down into objects and actions, properties, states, events, roles or whatever else. Where does this structure reside? Is it, in fact, some fundamental truth about that meaning, independent of human cognition, which was true before the science of linguistics was born and will be true after the last linguist is dead? Or is it the particular structure which the human mind imposes upon its sensory input as part of larger tasks, such as building models of the world to guide future action? I do not intend here to deny the existence of Platonic ideals of meaning, but I do wish to raise the question of their relevance to language. Even if these forms do exist, the human mind cannot be taken, under any materialist account of cognition, to have direct access to them - it only has access to the structures that it constructs. In light of this, why should human language users care if the structure of their language reflects the Platonic conceptual structure of its referents? If the idea is that this reflection makes language comprehension faster or easier or more reliable, then clearly the only conceptual structure which it makes sense to speak of reflecting is that imposed by the mind. In other words, structure-concept iconicity as a tennet of language functionality only makes sense if we think of it in terms of the interface between the human CIS, which is responsible for representing the meanings of utterances, and the FLN, which performs the necessary computations to translate conceptual structures into syntactic structures. The more similar these two structures are, the better we should expect language use to proceed, since the less information processing has to be performed to translate one into the other. However, under this view iconicity functionalism

is in no way external to the human mind, but instead intimately tied up with it. The appropriate conceptual structures to use in functionalist theorising can be inferred only from psychological experimentation, rather than from pure philosophising. It is, of course, possible in principle that the conceptual structure the human mind imposes on the world is a very close or even exact match to the true Platonic structure, but (i) it is still a property of the mind which must be invoked to explain language structure, and (ii) we are then forced to account for this perfect correspondence: the most likely candidate for such an explanation seems to be to invoke natural selection, so that we are still working very much in the realm of the biological and cognitive.

A relatively similar train of thought can be applied to discourse functionality. On what grounds can our ideas about topics and comments, themes and rhemes etc. be considered necessary properties of discourse in general? Does it even make sense to talk about discourse in general? Human language is literally the only thing which is remotely like itself that humanity has discovered thus far. Even if we are willing to grant the existence of Platonic discourse, we once again have to ask by what means the structure of this discourse can have any causal effect on human language processing. If humans are able to understand conversations easier when topics precede their comments, then it can only be the case that the cognitive machinery used for language comprehension is structured such that this is the case. A different kind of language faculty in a different kind of mind could exist in principle which was built so as to expect comments before topics. So what language is really doing according to discourse functionalist accounts is not adapting to some abstract requirements of communication, but is in fact adapting to the way that the human mind is built to process conversations. So we see that, once again, a major component of traditional functionalism is not in fact distinct from the human mind, but is intimately connected to it.

In light of the above considerations, it seems to me that the apparent dichotomy between, to use Hyman's words above, functionalism based on communication and discourse and functionalism based on cognition or the nature of the brain, is entirely false *in so far as* it applies to most of what has taken place in functionalist linguistics so far. Conceptual meaning and discourse as Platonic abstractions may or may not exist, but even if they do, only meaning and discourse as instantiated in the human mind can have any causal effect of language use. Thus, I propose that most of functionalism in the linguistics literature really is cognitive functionalism, whether the relevant authors have called it this or not. I still think it still makes sense to think of two distinct kinds of functionality here, even though both are inherently tied up in the human mind. To see this, we can take Hurford's example of the two kinds of functionality-driven features of a spade, and replace the spade - a tool made to facilitate a certain kind of interaction between humans and the external world - with, say, a toothbrush. Some design features of a toothbrush are in aid of the fundamental task of brushing teeth, such as the size and shape of

the bristled region, the length of the bristles and the material the bristles are made of, and some design features are in aid of making a toothbrush easy for humans to use, such as the size, shape and texture of the handle. However, *all* of the design features of a toothbrush are inherently tied up with human anatomy - some with the anatomy of the mouth, and others with the anatomy of the hand. I propose that in the same way, almost all of the design features of human language are inherently tied up with human cognition, including many which may seem at first blush and which have traditionally been characterised as concerning communication in the abstract. Because of this, functional linguistics - properly construed - should be of tremendous interest to cognitive science. I will stop short of insisting that all functionalism is necessarily cognitive functionalism, because I do believe that there are aspects of functionality which genuinely do apply to serial communication in the abstract rather than the specific instantiation of the human capacity for language. However, as far as I can see, this sort of genuinely abstract functionality has not had a substantial presence in the linguistics literature, despite what some functionalists may think. In Chapter 8 I shall formulate an account of word order functionality which is a genuine example of functionalism in the traditional sense.

## 3.2 Previous functional explanations for word order frequencies

In this section I present what I believe to be an exhaustive review of previously offered functional accounts of word order. The accounts appear in chronological order of publication. All of the accounts attempt to provide an explanation for basic word order frequencies in the same way: by showing that the frequency with which a given word order is the basic word order of some language is proportional to how functional that word order is. In other words, they attempt to show that SOV > SVO > VSO > VOS > OVS > OSV is a correct ranking of the word orders not only by frequency but also by functionality. I shall term explanations of this kind *standard functional explanations*. I have included in the survey functional accounts which produce a close rather than perfect match between frequency and functionality rankings, for reasons which will become clear later. Before continuing, I wish to emphasise two points.

Firstly, I will consider only those explanations which actually purport to explain the frequencis of basic word order, rather than taking explanations for more general word order universals and unfairly limiting their scope to basic word order. As an example of this sort of unfair dealing, consider (Manning & Parker, 1989)'s criticism of Greenberg's Cross Category Harmony (CCH) hypothesis. The intended purpose of CCH is to explain the widely agreed upon fact that there are interesting correlations between the basic word order of a language and other logically independent word ordering parameters such as the use of prepositions or adpositions, or the relative order of nouns and

adjectives or genetives which modify them. Despite this general scope, CCH can be used to produce a ranking of basic word orders alone from most to least harmonic, when all else is equal. While CCH fares quite well at explaining the correlated nature of the observed settings of several logically independent parameters, when artificially restricted in this way it does not agree with the data quite so well, and Manning and Parker use this to dismiss the theory outright. In this section I will consider only explanations which have been developed strictly to explain the relative frequency of the six basic word orders, since that is the problem actually under discussion. As such there shall be no mention of Greenberg's CCH, nor any of the many influential ideas of John Hawkins, which have a similar intended scope to CCH (however, there shall be some brief discussion of these theories and their explanada in Chapter 11).

Secondly, for some of the explanations I review below, the connection to linguistic functionality will not be at all clear, despite the explanation having received the label of "functional" (both in this thesis and in the linguistics literature). This reflects a tendency in linguistics to declare that a particular linguistic property is functional on the basis of quite indirect evidence. For example, if a property is observed to be consistently preferred cross-linguistically to an alternative, and the property is defined in terms of concepts such as agent and patient or topic and comment which are not purely syntactic in nature, then that property may be granted the label "functional" even if the precise manner in which it *is* functional is unknown. Obviously, those explanations in which the functionality *is* explicitly identified should be preferred over those in which it is merely assumed due to a lack of plausible alternatives, but this is, unfortunately, relatively rare, and in this review I have erred on the side of completeness.

For convenience, after discussing each explanation I present on a new line the ranking by functionality which each account yields.

### 3.2.1 Diehl

Diehl (1975) presents a very lengthy and complicated functional account of word order, with the ambitious goal of constraining not only the relative ordering of subject, verb and object but also the relative ordering of these constituents and indirect objects[5] (I) and "non-terms"[6] (N). I will preface my discussion of Diehl's work by pointing out that the work is very poorly written, by Diehl's own admission ("because of limitations of space, this paper is

[5] I did not discuss the notion of direct vs indirect objects in Chapter 2. Comrie (1981) argues that, in fact, the distinction between direct and indirect objects, at least in English, is not a genuine syntactic distinction, and that instead the terms as they are commonly used reflect a purely semantic distinction. I shall not discuss the matter in detail here since, in considering Diehl's work I am interested only in the relative ordering of S, O and V.

[6] "Non-term" is a term defined in the formalism of relational grammar. For my present purposes it can simply be thought of as meaning "anything which is not a subject, verb or direct or indirect object.

fragmentary, i.e., this extraction omits pages of important material. Because of limitations of time, this work is still preliminary"), and that my discussion is based largely on the understanding gleaned from the much clearer (though shorter and less detailed) explanation given by Tomlin (1986).

Diehl essentially makes use of four independent functional principles in accounting for basic word order, three of which are clearly recognisable as examples of iconicity functionalism and the last of which as discourse functionalism, to use the terms which I introduced earlier in this chapter. These principles are: Egodeictic Iconicity (EI), which holds that "outward order (e.g. from head of construction in syntax...) reflects sequence of acquisition which itself reflects sequence in cognitive development", Temporal Iconicity (TI), which holds that "sequence in lingusitic representation refelcts sequence in nonlinguistic cognitive experience", Linkage Iconicity (LI), which holds that "an element relating or "linking" two terms is shown in strings to do so by its being ordered between the two terms", and Discourse Structure, which holds that "the degree to which a given natural language shows free variation in word order is nothing more or less than its degree of sensitivity to either topical structure or information structure or both". These four principles are essentially distillations of all the important ideas in the traditional school of functionalism: the surface form of language is determined by the various structures of the meanings of sentences and by the structure of the discourse process in terms of topics and new or old information.

From his four principles, Diehl derives a number of "sub-principles" which dictate the relative ordering of S, O, V, I and N). The precise manner in which the sub-principles are derived from the main principles is often explained in very theory-specific specialist terms which it is beyond my present scope to explain here, so I shall resort to simply stating the sub-principles and their implications for word order without further justification. I shall state only the sub-principles which can be stated straightforwardly in terms of the ordering of S, O and V. These sub-principles are:

1. Nuclear integrity (derived from EI): this sub-principle essentially states that the verb and direct object of a sentence should be adjacent, i.e. O and V should not be separated by S. The word orders VSO and OSV are not compliant with this sub-principle.

2. Actor before patient / Agent before object (derived from TI): this sub-principle simply requires that the subject of a sentence should precede the object, as in SOV, SVO and VSO.

3. Noun-Verb-Noun (derived from LI): this sub-principle states that verbs should be positioned between the terms they relate, as is the case in SVO and OVS.

Curiously, Diehl does not seem to derive implications for the relative order of S, O and V from his principle of discourse structure by claiming a correlation

between topic and subject, which is something we shall see in later accounts of word order functionality. The three sub-principles above (and other subprinciples Diehl states but which I have not mentioned here due to their lack of straightforward connection to our matters of interest) eliminate many of the possible orderings of S, O, V, I and N, but do not constrain them entirely, so that some amount of arbitrary choice is left. Diehl deals with this by proposing an ordered set of 4 yes-or-no questions, some of which basically boil down to choosing a priority of some functional principles over other, which can be used to completely define an ordering of the 5 constituents.

Applying Diehl's sub-principles and his system of 4 ordered questions, only six orders can be derived, these being NSIOV, VSOIN, SVOIN, VOSIN, NSIVO and NSVOI. If we ignore the I and N constituents, we are left with the basic word orders SOV, SVO, VSO and VOS. In other words, Diehl's functional principles completely exclude the possibility of object-initial basic word orders. We can rank the four permitted orders in terms of how many of the independent principles they comply with, in their various "full" (i.e. including I and N) forms. By this criteria, SVO is the most functional word order, followed by SOV, then VSO and finally VOS. Thus, overall, Diehl's functionality ranking is SVO > SOV > VSO > VOS > OVS = OSV.

Diehl's functionality ranking: SVO > SOV > VSO > VOS > OVS = OSV.

### 3.2.2 Mallinson and Blake

Mallinson and Blake (1981) endeavour to explain word order in terms of three general principles. These are:

1. TOPIC-TO-THE-LEFT PRINCIPLE: "More topical material tends to come nearer to the beginning of the clause (to the left) than non-topical material"

2. HEAVY-TO-THE-RIGHT PRINCIPLE: "Heavy material tends to come nearer to the end of the clause (to the right) than light material"

3. "Constituents tend to assume a fixed position in the clause according to their grammatical or semantic relation or category status (noun, verb, prepositional phrase, etc.)"

The third of these principles may appear odd, and some clarification is required: these principles are intended to apply more generally than to the basic word orders of those languages which have them. They are supposed to apply also to the utterance-by-utterance choices made by speakers of languages with free word order, and also to those cases where speakes of languages *with* a basic word order choose to use alternative orderings. The third principle, then, is presented to highlight the fact that in many languages, word order is not totally free to vary on an utterance-by-utterance basis in accordance

with the first two principles. Mallinson and Blake offer Latin (which has free word order) as an example of a language which is particularly responsive to the first two principles and does not abide by the third principle, and Cambodian (also known as Khmer) as an example of a language which follows only the third principle ((Dryer, 2008) assigns Khmer SVO basic word order). Most languages presumably occupy a position between these extremes, having a basic word order as per the third principle, but permitting alternate orders, which are used in a manner guided by the first two principles. Mallinson and Blake do, however, make the claim that we should expect the fixed orders covered by the third principle to be orders which are generally in line with the first two principles, and as such these principles are relevant to the present interest in basic word order.

The TOPIC-TO-THE-LEFT PRINCIPLE uses the term topic "in the sense of what is being talked about" and the term comment "in the sense of what is being said about the topic". It is argued that (at least in accusative languages) subjects are more often topics than comments, and as such this principle essentially boils down to the claim that S should precede O. Thus, the principle "can be used to explain, at least in a weak sense, the preponderance of SO orders in language". Much to their credit, Mallinson and Blake are quick to point out that this "is a satisfactory explanation...only if we can explain independently why a topic should precede a comment". They stop short of supplying a fully detailed explanation of this (by no means a simple task), but observe that "topics do precede comments in mediums of communication other than language", citing the examples of mime, dance and television commercials. Presumably, these appeals to intuition are hinting at an explanation something along the lines of topic preceding comment being in some sense a natural property of the CIS.

The HEAVY-TO-THE-RIGHT PRINCIPLE uses the term "heavy material" to mean "internally complex material". As examples, they state that an NP consisting of two coordinated NPs is heavier than a "simple" NP, i.e. "the boy and the girl" is heavier than "the children". Further, NPs with phrasal or clausal complements ("the girl on the magazine cover" or "the girl who was featured in the centerfold of the Financial Review") are heavier than simple NPs ("the girl"). This principle is much better motivated than the previous principle, by the claim that increasingly severe violations of this principle place increasing demand on short-term memory. Any partial information or expectations derived from linguistic input before an especially heavy phrase must be maintained in memory for the duration of the heavy phrase before they be combined with or satisfied by any input which comes after the heavy phrase. Thus, placing the heavy material at the very end of the utterance reduces the maximum amount of time which these expectations must be kept in memory. With regard to the relative ordering of S and O, it is pointed out that topics (which are more often S) tend to be either pronouns or simple noun phrases, their nature being well understood by context. In contrast, "complements to

heads of noun phrases typically occur with material that is part of the comment, part of what is being presented as new". Thus, subjects (which are more often topics) will tend to be less heavy than objects (which are more often comments), so that this principle also boils down to S before O.

When it comes to ranking the six possible word orders according to this account of functionality, it is not precisely clear how to proceed. The principles essentially claim that S should precede O, so that immediately we have the ranking (SOV, SVO, VSO) > (VOS, VOS, OSV). This ranking meets Tomlin's minimum requirements for an explanation of basic word order, in that it separates the frequently occuring word orders from the infrequently occuring word orders, but it would be nice to go further. It seems to me that SVO abides by the principles better than SOV or VSO, since it places the subject at the very beginning and object right at the very end of the utterance. Simlarly, OVS violates both principles more severely than VOS or OSV. This leads to the ranking SVO > SOV = VSO > VOS = OSV > OVS. I can think of no principled way to impose a more fine-grained ranking than this: SOV better adheres to the TOPIC-TO-THE-LEFT PRINCIPLE than VSO, by virtue of being subject-initial, but VSO better adheres to the HEAVY-TO-THE-RIGHT PRINCIPLE than SVO, by virtue of being object-final. If both principles are considered of equal importance, the two orders must be considered equally functional. A similar argument holds for VOS and OSV.

Mallinson and Blake's functionality ranking: SVO > SOV = VSO > VOS = OSV > OVS

### 3.2.3 Krupa

Krupa (1982) offers an account of word order functionality based on three functional principles, ranking the six word orders by how well they obey each of the principles. Each of the word orders is taken to be either completely, partially or not at all compliant with each of the principles, and is assigned a score of 2, 1 or 0 for each principle respectively.

Krupa's first principle relates to minimising what he calls SENTENCE DEPTH. This is a cognitive functional principle, in that it is motivated by minimising working memory demands when parsing sentences. Following earlier work by Yngve, Krupa defines the sentence depth of a word order to be the number of right-hand constituents which must be kept in short term memory while left-hand constituents are generated, assuming that the generation of constituents proceeds as per their phrase structures. This assigns a sentence depth of 1 to SOV and SVO, and 2 to all other word orders. Krupa thus assigns the subject-first orders scores of 2 (full compliance) for this principle, and the others scores of 0 (no compliance).

The second principle that Krupa asserts is what he calls "cognitive nature". This is essentially an appeal to the traditional functionalist concept of iconicity, but Krupa also cites the fact that anthropocentrism in human language moti-

vates placing noun phrases which refer to humans before noun phrases which do not. Krupa takes SVO to be the only word order which is in complete agreement with the cognitive nature of the meanings it expresses, and thus awards it a score of 2. SOV, VSO and OSV conform to the principle partially, earning a score of 1, and OVS does not conform to it at all, and is scored at 0.

The third and last of Krupa's principles concerns Relative Structural Independence, the idea being that those constituents which are "most central and independent" ought to occur earliest in sentences. Krupa claims that the verb of a transitive sentence is the most structurally independent, followed by the subject and then the object. Thus, VSO word order is in full compliance with this principle, its mirror image OSV is in the least compliance with this principle, and the remaining orders SOV, SVO, VOS and OVS are in intermediate compliance. Krupa thus assigns these three sets of orders scores of 2, 1 and 0 respectively.

Summing the scores assigned to each word order over the three principles, we obtain the ranking SVO (5) > SOV (4) > VSO (3) > VOS (2) > OVS (1) = OSV (1), which is in quite good overall agreement with the basic word order frequencies. The only discrepancies are that Krupa's account ranks SVO above SOV, and it ranks OVS and OSV equally, rather than OVS ranking higher. This latter discrepancy can be considered quite minor since OVS and OSV are both so rare. However, note that Krupa's overall ranking depends significantly on the fact that each word order receives a score of either 0 or 2 for compliance with the sentence depth principle. This scoring seems somewhat poorly motivated to me, and it is primarily a result of Krupa's having to force only two possible levels of functionality (1 and 2 are the only sentence depths possible) onto a three-point scale. While assigning scores of 0 and 2 is perhaps the most obvious way to do this, it isn't clear that the difference in functionality between the two sentence depths is necessarily as great as that between, say, VSO and OSV with regards to relative structural independence. If sentence depths of 1 and 2 are assigned scores of 2 and 1 or 1 and 0 respectively, then the ranking changes to SVO > SOV = VSO > VOS > OVS = OSV, which is not in as good agreement with the observed word order frequencies as Krupa's published ranking.

Krupa's functionality ranking: SVO > SOV > VSO > VOS > OVS = OSV.

### 3.2.4 Tomlin and Song

Tomlin (1986), like Krupa, presents a set of three functional principles for word order and uses them to derive a ranking of word orders. The overall approach is quite similar to Krupa's, although each word is either compatible or incompatible with each of Tomlin's principles: there is no middle ground. Tomlin shows that frequency of each word order is proportional to the number of principles the order is compatible with: SOV and SVO are compatible with all three principles, VSO is compatible with two of them, VOS and OVS with

only one of them and none of the principles are realised by OSV word order.

The first of Tomlin's principles is the THEME-FIRST PRINCIPLE (TFP), which states that: "in clauses information that is relatively more thematic precedes information that is less so". Tomlin defines the notion of "thematic" as follows: "information in an expression is thematic to the extent the speaker assumes the hearer attends to the referent of the expression". Tomlin demonstrates a statistical correlation between a noun phrase being thematic and being the subject of an utterance, so that the TFP essentially states that subjects ought to precede objects. Thus, SOV, SVO and VSO are all compatible with the TFP, while VOS, OVS and OSV are not. Note the obvious similarity between this principle and Mallinson and Blake's topic-to-the-left principle.

The second principle that Tomlin presents is VERB-OBJECT BONDING (VOB), which states that: "the object of a transitive verb is more tightly bonded to the verb than is its subject". The claim is that "a transitive verb and its object form a more cohesive, unified syntactic and semantic whole" and that, as a consequence, it is "more difficult to interfere with the...unity, by attempting syntactic insertions, movements and so on". Thus, word orders in which V and O are adjacent, such as SOV, SVO, VOS and OVS are compatible with VOB, whereas the orders VSO and OSV, in which the cohesive whole of V and O are interrupted by the subject, are incompatible with VOB.

Tomlin's final principle is the ANIMATED-FIRST PRINCIPLE (AFP), which states that "in simple basic transitive clauses, the NP which is most animated will preced NPs which are less animated. Similar to his treatment of the TFP, Tomlin shows that the subject of a sentence is significantly more likely to also be the most animated NP in the sentence than the object, so that, like the AFP, the TFP essentially states that subjects ought to preced objeccts. The pattern of compatibility and incompatibility of the six word orders with the AFP therefore matches the TFP.

Ranking the basic word orders by the number of Tomlin's principles they permit realisation of yields SOV (3) = SVO (3) > VSO (2) > VOS (1) = OVS (1) > OSV (0). This ranking is largely in agreement with the observed frequency ranking. Typically, SOV is taken to be more frequent than SVO and VOS to be more frequent than OVS, but Tomlin's language survey, discussed in the previous chapter, is part of the same work as his functional explanation, and recall that in that survey he concluded that differences in frequency between these pairs of word orders was statistically insignificant, so that this ranking is in full agreement with his conclusions. I mentioned that subsequent work by Dryer has questioned our inability to assert that SOV > SVO, but nevertheless, the two word orders are quite close in frequency and so this shortcoming of Tomlin's account can be considered minor.

(Song, 1991) suggests a slight modification to this ranking scheme, wherein VSO is held to realise the TFP and AFP less well than SOV and SVO, insofar as the thematic/animate subject comes after the verb. This modification maintains the same overall ranking, but increases the degree to which SOV

and SVO are more functional than VSO, in accordance with the fact that the subject-first orders are substantially more frequent than VSO.

Tomlin's principles lead to a ranking of word orders which matches the data quite well, however, it is important to consider the question of how well motivated the principles are. If a principle has no independent justification beyond its ability to explain the observations then it is in essence nothing more than a useful recharacterization of the data, and so to offer it as an *explanation* of patterns in that data is circular. Tomlin acknowledges the need for this deeper justification, stating that "it is to be expected that one will be able to demonstrate that each of the functional principles proposed here should find its ultimate motivation in the processes and limitations of human information processing abilties". Tomlin only provides independent motivation for the TFP, by citing earlier work (Tomlin & Kellog, 1986). This work describes a behavioural experiment in which participants are required to place letters in coloured boxes in response to spoken instructions. Before the task begins, participants are either briefed that their task is to involve the manipulation of letters, with no mention of boxes, or vice versa. This is assumed to establish either letters or boxes but not the other as the primary target of the participant's attention. Spoken instructions are given to the participants in two forms, either with the letter or the box to be manipulated mentioned first (e.g. "put the E in the blue box" or "in the blue box, put the E"). The results of the experiment show participants completing individual sorting tasks faster on average when the focus of their attention (either the box or the letter) is mentioned first. Tomlin concludes that "to the extent that thematic information does reflect attention allocation in discourse production and comprehension, this experiment provides direct evidence that earlier ordering of thematic information facilitates comprehension while later ordering inhibits it". The VOB and AFP remain lacking in independent motivation, so far as I know.

Tomlin's functionality ranking: SOV = SVO > VSO > VOS = OVS > OSV.

### 3.2.5 Manning and Parker

A very unusual functionalist account which manages to perfectly match the desired SOV > SVO > VSO > VOS > OVS > OSV ranking is presented by Manning and Parker (Manning & Parker, 1989). The core of their proposal is that the various syntactic orderings of subject, verb and object are influenced by the application of a figure/ground visual perception process to a psychologically real "semantic diagram", which visually represents the inherent semantic relationships between the constituents. In the same way that the famous "Rubin's vase" illusion can be interpreted as an image of either a vase or of two inward-pointing faces, depending on which of the black and white parts of the image are interpreted as the figure and ground, respectively, Manning and Parker suggest that the diagram shown in Figure 3.2.1 can be interpreted in

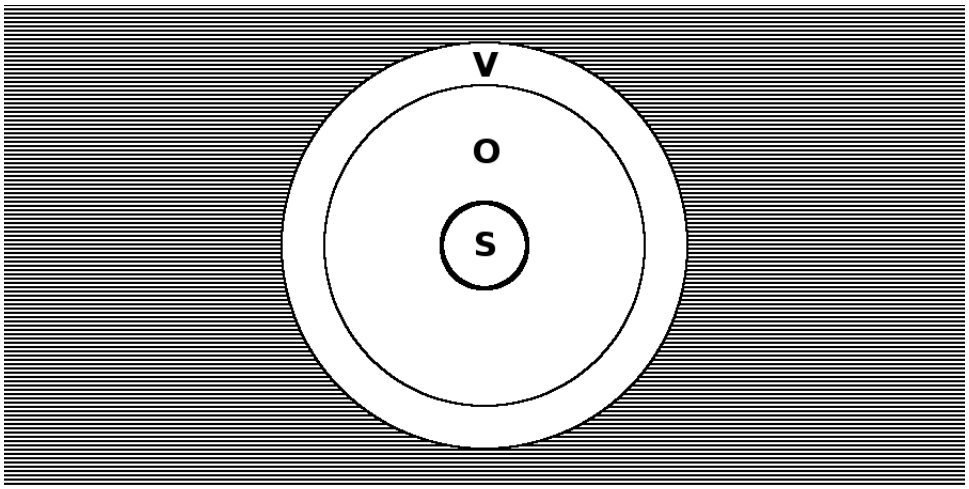


Figure 3.2.1: Figure/ground interpretation image underlying Manning and Parker's explanation for basic word order frequencies

one of six possible ways. The six possible interpretations correspond to the six possible ways of placing the circles labelled S, O and V in front of or behind one another, each of which can be taken to correspond to a word order. For instance, the word order SOV corresponds to the perception of the image in which the S circle is in front of the O circle, which is in turn in front of the V circle.

Manning and Parker use two general principles of figure/ground perception to establish an ordering on these possible interpretations, from the most to least prefered by the visual system. The two principles are "if one area encloses another, the enclosed area is likely to be seen as the figure. If a figure is divided into two areas, the smaller of the areas is favored as the figure", and "the smaller the area of the enclosing region with respect to the enclosed region, the easier it is to interpret the enclosing region as the figure". According to these two principles, the possible interpretations of the diagram can be ranked according to preference as SOV > SVO > VSO > VOS > OVS > OSV, exactly matching the observed frequencies of the corresponding basic word orders.

In order for this explanation to stick, two additional things are required: a justification for why the circles of the diagram should be labelled S, O and V in the way that they are, rather than by one of the six possible alternative labellings, and a justification for why word order acquisition should be at all influenced by any processes derived from visual perception.

On the first of these points, Manning and Parker offer four independent motivations. These are: an appeal to the similarity between their diagram and Venn diagrams in classical predicate logic, the "relationship between S, O and V in terms of generality of reference", the morphological behaviour and relative independence (as appealed to by Krupa) of S, O and V, and some similarities between the diagram and the behaviour of cases in ergative

languages. Personally, I find these arguments fairly unconvincing. Song (1991) argues convincingly that the argument from morphology is in fact incorrect, and also that the three language-based arguments are in fact circular: the diagram is both being used to explain and justified by linguistic phenomena.

On the second point, even less motivation is presented. This is troubling, as it is this point which arguably requires the most justification: why on Earth should word order have anything to do whatsoever with visual perception of a diagram? What does it actually mean for the diagram in Figure 3.2.1 to be psychologically real and "subject to the same interpretive process as visual form"? Certainly nobody actually *sees* this diagram during the process of language perception. In motivating their overall approach, Manning and Parker seem content simply to quote bare assertions by Peirce. The explanation is thus in the unusual position of being the only published explanation which offers a perfect match to the data, but also the explanation which feels by far the least convincing.

It is worth noting that Manning and Parker themselves do not consider their explanation to be functionalist, explicitly stating as much in their paper. This position stems from their unusual position of considering functionalism as claiming that language external semantic and functional considerations directly determine language form, without any sort of underlying causal mechanism. In the paper, functionalism is repeatedly compared to the discredited Lamarckian theory of evolution. To my knowledge, no other authors have considered this to be an essential aspect of functionalism. As we shall see later, some functionalist researchers have concerned themselves quite explicitly with the causal mechanism by which asymmetries in functionality give rise to differences in frequency, though admittedly this practice was not widespread at the time Manning and Parker wrote their account. Despite this issue often being overlooked, I strongly suspect that the overwhelming majority of functionalist researchers have always believed that some sort of naturalistic mediating mechanism is at work when languages adapt in response to functionalist pressures, rather than some kind of "spooky" direct causal relation. Insofar as it links basic word order to issues of compatibility between linguistic form and a general semantic representation in the mind, and achieves this link by appeal to the preferences of a computational process, Manning and Parker's account is clearly within the scope of cognitive functionalism.

Manning and Parker's functionality ranking: SOV > SVO > VSO > VOS > OVS > OSV.

## 3.3 Summary

The most plausible framework within which to seek an explanation for linguistic diversity, including the special case of cross-linguistic distribution of basic word orders, proceeds in analogy to the theory of evolution from biology. Ran-

dom variations in language, due to imperfect production and acquisition, are amplified or suppressed by various selective pressures. Of particular interest for my purposes is *functional selection*. Functional selection acts to amplify those aspects of random variation which serve to make a language more functional, that is, better able to achieve its purpose of facilitating communication. A number of functionalist explanations for the cross-linguistic distribution of basic word order have been proposed in the literature over the previous decades. Each of these accounts have approached the problem in the same way: by striving to show a direct relationship between functionality and frequency, i.e. to show that the $n$-th most frequent basic word order is the $n$-th most functional. A number of explanations have succeeded or come very close to succeeding in finding such a relationship.

At this point one may very well get the impression that the non-uniform distribution of basic word orders is well on its way to being explained in functional terms, if it is not already there. The wide variety of explanations considered above seems closer to an "embarassment of riches" than evidence of a hard problem. Explanations such as Krupa's, Tomlin's and Manning's and Parker's yield functionality rankings relatively close (or in Manning and Parker's case, perfectly matching) the ranking by frequency, so that some minor tweaking and/or combination of these accounts of functionality can provide multiple perfect matches. In the following chapter, I will do my best to argue that the problem only *appears* solved from a strictly synchronic perspective: that is, a perspective which takes into account only the present state of affairs and does not attempt to incorporate what we know about historical, or diachronic, change in basic word order. I will argue that when we adopt a diachronic perspective, the standard functional explanations all fail, and more complicated ideas are required to explain both the distribution observed today and what we can reasonably infer about how the distribution must have changed over time.

# Chapter 4

# A problem with previous explanations and a solution

In the previous chapter I justified my decision to consider only functional explanations for the cross-linguistic distribution of basic word orders, and surveyed a range of previously proposed functional explanations. All of these explanations were of the same general kind, which I termed standard functional explanations. Standard functional explanations are characterised by the idea that the ranking of word orders by frequency should match precisely their ranking by functionality. In this chapter I identify a significant shortcoming of all standard functional explanations: their contradiction of what we know about common directions of basic word order change.

## 4.1 A problem

### 4.1.1 Implicit theories of word order change in standard functional explanations

None of the standard functional explanations I reviewed in the previous chapter dealt explicitly with the issue of word order change. All that was claimed was that presently observed distribution of basic word orders mirrors the ranking of those word orders by functionality. But recall earlier in that chapter that functional selection can be embedded into a model of language change, wherein functional selection acts to amplify or suppress random variations. Indeed, functional explanation of language universals is not well-founded without such an embedding (in fact, even more work is generally required to establish a functional explantion for a universal, as I shall discuss briefly in Chapter 11). As such, the standard functional explanations make implicit claims about historical word order change: word order changes which correspond to movement up the functionality ranking, such as SVO $\rightarrow$ SOV, VSO $\rightarrow$ SVO and VSO $\rightarrow$

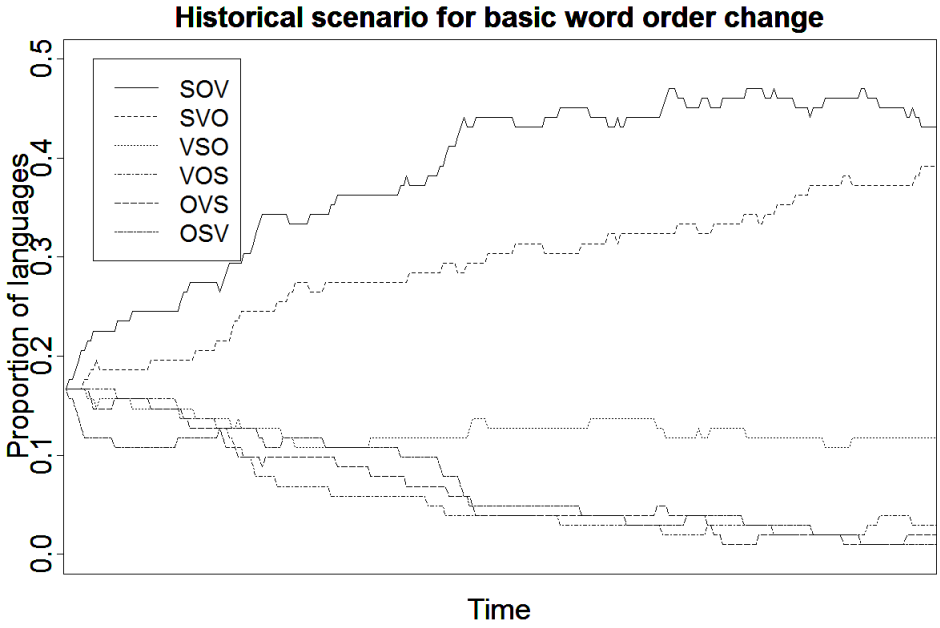


Figure 4.1.1: Implicit historical scenario for standard functionalist explanations of basic word order frequency.

SOV[1], should be expected to occur much more frequently than changes which move down the ranking, such as SOV → SVO, SOV → VSO or SVO → VSO. The implicit historical scenario is something like the following, as illustrated in Figure 4.1.1:

**S1** In the distant past, the frequencies of the six different basic word orders were roughly equal. Over long extents of time, the word order of languages changed, generally, from less to more functional word orders. As such, the proportion of languages with highly functional word orders increased while the proportion of languages with poorly functional word orders decreased. This process has reached something approximating equilibrium, so that the present day frequency of each word order is proportional to its functionality.

**S1** requires some clarification and justification, since it is very much an implicit scenario and I am in effect putting words into the mouths of several scholars. First of all, the equal or nearly equal frequency of the bssic word orders at the "dawn of linguistic time" is not actually strictly necessary. If change from less functional to more functional word orders is dominant, then given sufficient time we should expect there to be no residual trace of the

[1]The > character is traditionally used to denote direction of word order change, e.g. a change from SOV to SVO would be written SOV > SVO. Since I have already used the > character in this thesis to denote that one word order is a more frequent basic order than another, I shall use → to denote direction of word order change to avoid confusion

initial frequencies and instead just a distribution which reflects functionality. However, an equal starting point seems like the most sensible null hypothesis to ascribe to the standard functional word orders. A strongly non-uniform initial distribution would firstly require some additional explanation, and to the best of my knowledge none of the standard functional word orders suggest either a non-uniform initial distribution or a possible explanation for one. Furthemore, strongly non-uniform initial distributions can necessitate claims about either the age of human language or the average speed of word order change. As an extreme example, if the initial distribution was heavily skewed toward the least functional word orders, with the most functional word orders being extremely rare, then a claim that the present day frequencies directly reflect functionality implies either that word order change happens "very fast" or that the initial conditions existed "a very long time ago", where the phrases in quotation marks should be interpereted as relative to what we must assume given unbiased initial conditions. So the near-equal initial frequencies component of **S1** simply amounts to attributing the bare minimum amount of implicit assumptions to other authors, which seems proper.

I should also justify my claim that standard functional explanations are in fact implicitly assuming a predominance of change from less to more functional word orders. Earlier I alluded to the definition of functionalism I provided in the previous chapter, which embedded functionality in a model of language change and assigned it the role of a selective force amplifying or supressing random variation. However, I did note in that chapter that I was somewhat arbitrarily choosing one characterisation of functionality, in the lack of any clear standard or consensus. Perhaps the authors of standard functional explanations have chosen for themselves ideas of functionalism which are not inherrently involved in language change? It is not hard to see that this approach cannot really work out. There are two alternatives to the idea of word order change strongly prefering increases in functionality over decreases in functionality. First is the reverse scenario, where word order change is typically in the direction of *decreasing* functionality. Of course, if this kind of change happens over a long period of time then the result is that the most functional word orders should be the *least* frequent, and so claiming that SOV is the most functional word order is squarely at odds with it being the most frequent word order and the standard functional explanation completely fails to account for the data. The other alternative is that word order change is indifferent to functionality, with word orders changing from less to more and more to less functional with roughly equal frequency. Of course, if this kind of change happens over a long period of time then the result is a close to uniform distribution over word orders. Under this view of word order change, the standard functional explanation can be made to work only if we assume that the initial frequencies of word orders very closely matched the functionality of word orders (and some explanation would certainly need to be provided for this situation!) and also that word order change happens "very slowly" (again relative to what we must otherwise

assume), such that the present day distribution still closely matches word order functionality only because there has not been enough time to allow the inevitable diffusion to uniformity. Obviously, the idea that word order change most often proceeds from less to more functional orders is the most sensible to ascribe to the standard functional explanations.

So, if we attribute the fewest implicit assumptions possible, the standard functional explanations make two simultaneous sets of predictions which must be tested against empirical data: the *synchronic predictions*, namely that the presently observed frequency ranking of basic word orders should mirror the functionality ranking, and the *diachronic predictions*, namely that we should find more evidence for historical word order changes moving up the functionality ranking than for changes moving down it. In order for the standard functional explanations to be truly satisfactory, both of these predictions should survive empirical testing. In the case of the synchronic predictions, most of the standard explanations obviously fare relatively well, as they were explicitly constructed to survive just this kind of scrutiny. However, as far as I know, none of the authors whose explanations were considered in the previous account paid any attention to their explanation's implicit predictions about word order change and how compatible they may be with the relevant data. In the following section, I review what is known about trends in diachronic word order change so as to be able to make this assessment.

## 4.2 Common directions of basic word order change

### 4.2.1 Inferring word order change

The reader without a background in historical linguistics may wonder how it is that changes in word order which are taken to happen over long periods of time can be reliably inferred. The process of inferring word order change is complicated and also somewhat unreliable, certainly having a strong dash of art to go along with the science. A detailed description of word order change inference is beyond the scope of this thesis, but I shall attempt to provide a taste of some of the most basic methods - certainly there are more sophisticated options available than those I mention here, but I hope in this paragraph to show that inferring changes of word order in the undocumented past is certainly not impossible. The most straightforward case, although unfortunately one rarely applicable in practice, concerns those relatively few languages (such as English and Chinese) where there are written records in that language which extend back for many centuries, so that we can directly observe word order change happening. More frequently, word order changes must be inferred in a less direct manner. For instance, it is relatively straightforward to estimate the degree to which two languages are likely to be genetically related through careful com-

parison of their lexica, in a process known as lexicostatistics[2]. This allows us to group languages into families and subfamilies with some degree of confidence, provided we do not attempt to infer relationships from too distant a time. If we look at groups of closely related languages determined in this manner and observe that most of the languages in the group have one word order but one or two languages have a different word order, the most parsimonious explanation is to attribute the most common word order in the group to a common ancestor language and posit word order changes in the languages which are "odd ones out" today. Alternatively, as I shall discuss briefly in Chapter 11, there are a number of word order parameters other than basic word order which are more or lesss reliably correlated with basic word order. If we observe a language which has, say, SVO basic word order but whose other word order parameters all exhibit settings which are known to be better correlated with SOV basic word order, then we might infer a change from SOV to SVO to have taken place in this language. Finally, if the grammar of a language requires the use of different word orders in different situations, such as the mix of SOV and SVO seen in German, one of these word orders can sometimes be inferred to have originally been the only order if we have independent evidence that word order in some parts of a sentence are more or less prone to change than word order in other parts. For a more detailed discussion on reconstructing past word order, including more sophisticated methods relying on regular correspondences between the syntax of languages analogous to the correspondences between lexica used to infer genetic relatedness of languages, see (Harris & Campbell, 1995).

In the remainder of this section I shall review all instances that I am aware of in the literature where scholars have made a case for any overall directional trends in word order change. Such claims are relatively rare, and I consider only four cases below. In some respects, some of the claims below contradict each other quite flatly, but despite this there are certain directional trends which are acknowledged in every case. It is only these consensus facts which I shall use when it comes to assessing the compatibility of standard functional explanations for basic word order frequencies with diachronic data.

### 4.2.2 Vennemann

Vennemann (1973) presents a fairly complex schema of permitted basic word order changes. His schema includes as a distinct type of language free word order (FWO). The permitted changes are FWO → SOV → SVO → VSO; SVO → FWO; VSO → SVO; VSO → FWO. The schema is shown diagrammatically in Figure 4.2.1. Essentially, the very long term behaviour of this schema is of

[2]Lexicostatistics can hint strongly at the relatedness of two languages, but the standard of evidence required in historical linguistics to demonstrate the relatedness of two languages is more than simply similarity between lexicons. Typically, it is required that the lexica can be shown to exhibit a regular correspondence, i.e. that words in one lexicon can be derived from cognate words in the other though the application of fixed transformation rules.

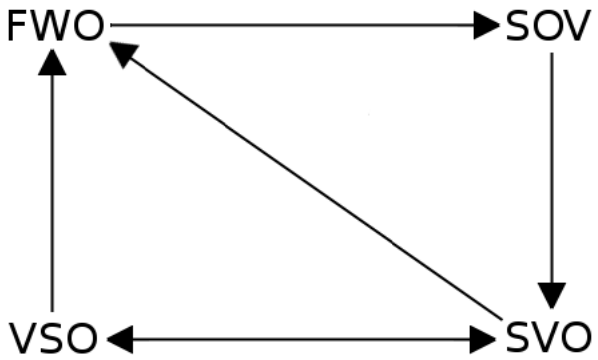


Figure 4.2.1: Permitted directions of word order change in Vennemann's schema. FWO is free word order.

languages constantly being attracted back to free word order from particular basic word orders. "Escape" from FWO is always made via an SOV stage, but languages always change from SOV to either SVO or VSO before ultimately returning back to FWO to repeat the cycle again. Flip-flopping between SVO and VSO is permitted. Of course, Vennemann should not be construed as claiming that languages necessarily actually engage in infinite loops through this system: I have described the system above as if it were, say, a Markov chain process, just to give an intuitive feel for the overal dynamics. While presumably Vennemann has in mind several reconstructed word order changes which are consistent with this schema, the schema seems largely motivated not by word order change data, but by Vennemann's ideas on the causes and mechanisms of word order change.

### 4.2.3 Li

Li (1977) makes the claim: "with the exception of the Chinese case...the only documented types of word order changes that are not due to language contact are SOV to (VSO) to SVO". The "Chinese case" he mentions as an exception is a controversial claim that Chinese is in the process of an SVO → SOV change. This claim has been called into question by, e.g. (Light, 1979) and (Sun & Givón, 1985).

### 4.2.4 Givón

(Givón, 1979) says: "the most common natural drive in word-order change seems to be SOV → VSO → SVO, with a much more restricted drift of SOV → VSO → VOS also attested in a number of cases. The drift *toward* SOV from any other typology is relatively rare". Givón doesn't offer any specific evidence for this claim in the same publication, and presumably it is based on

his knowledge of many reconstructed word order changes.

### 4.2.5 Gell-mann and Ruhlen

Gell-Mann and Ruhlen (In press) claim that the most common non-contact induced word order changes are SOV → SVO → VSO/VOS, with occasional VSO/VOS → SVO reversion. Their claim is notable for being based less on a database of reconstructed changes and more on a family-based analysis of current word order frequency data. Their approach is generally along the lines of: a particular language family has predominantly SOV basic word order with than SVO languages, so the more likely situation is that the family was originally SOV and some SOV → SVO change has occurred, rather than the family having originally been SVO and very extensive SVO → SOV change having occurred. Their resulting word order change trends are largely compatible with Vennemann's much earlier schema, although it is motivated quite differently (from data on word order distribution rather than a model of functional word order change). Gell-Mann and Ruhlen notably exclude FWO as an independent state, in the way that Vennemann uses it, claiming that FWO is more of an apparent than real state of affairs which occurs when a language is changing from one basic word order to another via a process of gradual decrease in the frequency of some word order(s) with corresponding gradual increase in the frequency of some other(s).

It is important to note that the particular large-scale language families used in this account are not uncontroversial. They are based upon a technique for inferring relatedness between languages called "multilateral comparison"[3], which many historical linguists reject as invalid (often quite strongly). This technique was pioneered by Greenberg and today is most vigorously championed by and typically associated with Ruhlen. The question of whether or not multilateral comparison is a valid tool for building linguistic taxonomies is well beyond the scope of this thesis. For a defence of the technique see (Ruhlen, 1994), and for criticisms see (Campbell, 2001).

### 4.2.6 Summary

As mentioned at the beginning of this section, these claims, taken as a whole, feature some degree of consensus and some degree of contradiction. For instance, while Li and Givón have VSO languages typically changing to SVO languages rather than vice versa, Gell-Mann and Ruhlen argue for just the opposite: that SVO languages more commonly change into VSO languages. Vennemann sits squarely in between these two extremes, permitting changes in both directions between VSO and SVO. Suffice it to say, on the question of

[3]The term "mass lexical comparison" is also sometimes used, primarily by detractors of the method.

which is the most common direction of change between SVO and VSO, there is no clear agreement.

In contrast to this, note that Li, Givón and Gell-Mann and Ruhlen all propose trends of change in which the overall direction of change is *away* from SOV. Givón states explicitly that drift toward SOV is rare. Vennemann permits SVO and VSO languages to change to SOV, but only via a FWO intermediary stage: he does not permit direct SVO → SOV or VSO → SOV changes, and here he is in agreement with all the other authors. I think it can therefore be fairly said that there is general agreement that word order change away from SOV is substantially more common than word order change toward SOV. I shall come to push this point quite hard.

## 4.3 Assessing the compatibility of standard functional explanations against the diachronic evidence

In this section I shall consider the question of how the implicit word order change predictions of standard functional explanations of basic word order frequency correspond to the consensus trends in word order change reviewed in the previous section.

### 4.3.1 Assessing overall functionality rankings

I shall begin by considering what is the most glaring incompatibility between the predictions and data, and this involves the SOV word order. This is the most frequently attested basic word order today and so, according to standard functional explanations, it is the most functional. It thus follows that we should see an overall drift *toward* SOV from other, less functional word orders. This is the implicit diachronic process which underlies standard functional explanations. However, as noted above, there is general agreement that this is in fact rare, and the data suggests an overall drift *away* from SOV. This incompatibility seems, to me, to be unresolvable. If SOV languages have a tendency to change into VSO languages (whether via an SVO intermediary or as an intermediary to an eventual change to SVO), but VSO languages rarely change into SOV languages, any claim that SOV word order is significantly more functional than VSO word order immediately looks quite suspect.

### 4.3.2 Assessing some individual functional principles against the diachronic evidence

Let us now consider some individual functional principles and how they fare with regard to the diachronic data. The ideal situation here is that common

directions of change in word order should correspond to the changes from word orders which are not consistent with these principles to word orders which are consistent with them, or at least are more consistent with them. If it should turn out that the overall diachronic trend is toward a certain principle being motivated, this can be seen as casting doubt on the significance of this principle for explaining basic word order distribution.

Diehl's principle of nuclear integrity and Tomlin's principle of verb object binding are essential identical, and are realised by SOV and SVO word orders but not by VSO (since in VSO the verb and object are separated by the subject). Under the account of Gell-Mann and Ruhlen, where SOV ultimately drifts toward VSO, there is an overall drift toward the *violation* of this principle. Under the accounts of Li and Givón, this principle is sometimes first broken by the change from SOV to VSO and then recovered in the change SVO via VSO. Under the account of Vennemann, this principle can be constantly broken and then recovered by the allowed bidirectional change between SVO and VSO. Under none of the accounts of trends in word order change is there any overall drift toward establishing and then maintaining this principle.

Krupa's three principles do not fare particularly well when subject to this kind of analysis. None of the proposed diachronic trends is consistent with a preference for maintaining minimum sentence depth. With regard to the cognitive nature principle, Gell-Mann and Ruhlen's story favours compliance whereas Li and Givón's and Vennemann's stories suggest indifference. Gell-Mann and Ruhlen's word order changes, however, suggest that Krupa's relative structural principle can be freely broken and recovered, whereas Li and Givón posit an overall trend in favour of this principle. Ultimately, none of the feasible diachronic scenarios have all of Krupa's principles being consistently respected.

All of this is, of course, not to say that these principles are not genuine functional principles. It may in fact be the case, for instance, that there is some functional advantage to having verb and object adjacent to one another, as Diehl and Tomlin suggest, but that this advantage is relatively small compared to the functional advantage which can be gained in other ways. As such, adherence to this principle may be "sacrificed" by a change which still leads to an increase in functionality overall. There is nothing fundamentally difficult about this, although these sorts of relative rankings of the principles by importance are something which proponents of the principles are obliged to justify. But if a principle supposedly offers such a small functional advantage relative to other principles then it seems unwise to rely upon that principle to play a strong role in explaining basic word order frequencies. However, for example, in Tomlin's account the VOB does just that: the fact that VSO word order violates VOB is the only thing which sets VSO apart from SOV and SVO in Tomlin's theory (all three orders obey the theme-first and animate-first principles), and the difference in frequency between VSO and the subject-initial word orders is very large. If the functional advantage of VOB were relatively small, we should expect VSO to be only a little less common than SOV or SVO.

It should be noted that functional principles which motivate ordering subjects before objects (and there are many of these: Diehl, Mallinson and Blake and Tomlin all propose at least one) are not violated at any stage under any of the accounts of word order change I have considered (except perhaps Vennemann's, in that it explicitly includes a FWO state in which any principle can be violated), and thus there is no reason stemming from diachronic considerations to question the importance of these principles to determining basic word order frequency. Of course, these principles themselves are not adequate for a full explanation of basic word order frequency: they do explain the rarity of VOS, OVS and OSV word order, but they make no distinction between SOV, SVO and VSO, despite the latter being far less frequent than the former.

Ultimately, none of the functional accounts of basic word order which have appeared in the literature so far have managed to meet the ideal of consisting entirely of functional principles which are all preferentially attained and then maintained during long term changes in word order.

### 4.3.3 Summary

Even though our present understanding of the historical changes of word order which have brought about the different frequencies of basic word orders we see today is far from perfect, it seems to me that we can fairly state that the implicit diachronic predictions made by standard functional explanations for basic word order frequencies conflict very badly with the diachronic knowledge we do have. This is no small problem. If functionalism is embedded within a framework based on language change such as Nettle's (and it is only such an embedding that makes it a valid explanatory tool), it is absolutely necessary that languages tend, on the whole, to change from less functional word orders to more functional word orders. Since the most well established property of word order change - that it is typically away from and not toward SOV - is both the primary cause of the failure of most standard explanations to agree with diachronic evidence and the one fact about historical word order change of which we are most certain, and since ranking SOV as the most functional word order is an essential component of any standard explanation, it seems to me that an entirely new approach to functional explanation of basic word order frequencies is required.

## 4.4 A solution

In order to avoid the incompatibility with diachronic data which the standard explanations suffer from, I argue that it will ultimately be necessary to reject the historical scenario S1 which implicitly underlies those explanations. A new historical scenario will be required, and an account of word order functionality to match it, whose synchronic and diachronic predictions both match the data. In particular, I will motivate the rejection of the implicit "fair start" hypothesis

of S1: that in the distant past, all word orders were roughly equally frequent as basic orders, so that only the functionality considerations which have driven the linguistic landscape away from that starting point have had a hand in shaping the present day distribution of basic word orders. This hypothesis is attractive primarily in that it feels like a natural null hypothesis, and as such does not require much in the way of explanation itself. If we hypothesise, as I shall, that the initial distribution of basic word orders was *not* close to uniform, then that assumption itself also requries explanation, quite independent of the functional explanation of the changes which have occurred since that initial distribution. In other words, the problem of explaining basic word order frequency factors quite neatly into two independent problems.

As an illustrative example of the situation we are now facing, consider the case of a naturalist who, upon arriving on a previously unexplored island, observes two subtly different varieties of a species of lizard, varieties A and B. Variety A lizards represent about 80% of the island's population, with variety B lizards accounting for only 20%. What is the relative fitness of the two varieties? On the basis of this data alone, we cannot be sure. One possibility is the A lizards are substantially better adapted to the island's environment than the B lizards, and hence are generally more successful at reproduction than the B lizards, explaining their greater numbers. However, another possibility, equally consistent with the population data, is that the B lizards represent a newly arisen variety of lizard who are in fact better adapted to the island than the A lizards, and are in the very early stages of displacing their less fit ancestors. Unless the difference between the two varieties has obvious implications for fitness, the only way to settle the matter is to perform a second survey of the island's lizard population at a later date. If a second survey 50 years in the future finds A lizards now account for 90% of the island's population and B lizards have dropped to 10%, clearly the A lizards are the fitter variety. However, if the new data finds a 70%:30% split between A and B lizards, the only conclusion is that B lizards are in fact the fitter variety, and the greater number of A lizards on the island is due solely to the fact that they are the original variety. The naturalist needs answers to two questions to fully explain the current state of the island: why did the first lizards on the island resemble type A lizards more closely than type B lizards (such that type A could be expected to appear on the island first), and what makes type B lizards fitter than type A lizards? Had the naturalist began searching, after her first visit to the island, for reasons as to why type A lizards were fitter than type B lizards, she would in fact have been looking in entirely the wrong direction. Despite this, of course, she may well have been able to come up with multiple plausible reasons for why type A lizards were the fitter variety. Ultimately, however, these reasons would have been misleading, as the most important factors concerning fitness favour the type B lizards.

I intend to argue that a similar story to the lizard island story above represents the best explanation for currently observed frequencies of basic word

orders, in the sense that it allows us to fit both the synchronic and diachronic data that we have, and that standard functional explanations for the observed frequencies may be misleading in the same way that the plausible arguments for lizard type A's superior fitness were. I will then outline what this story implies is necessary for an ultimate explanation of the observed frequncies. As per the story above, there are two questions which need answering, one of which will require a novel account of word order functionality.

### 4.4.1 Deriving a new historical scenario and corresponding functionality ranking

In this section I shall derive a historical scenario for basic word order frequencies which differs from the scenario **S1** which is implicit in standard explanations, as well as a ranking of word orders by functionality which is compatible with the scenario. This is a necessary first step toward creating an account of basic word order frequencies which is compatible with both the synchronic and diachronic evidence available. Since the explanatory framework developed here underpins the entire rest of the thesis, I shall proceed very carefully and formally.

I shall make use of the following five facts, which I believe to be relatively uncontroversial:

**F1** The only basic word orders which appear with significant frequency today are SOV, SVO and VSO.

**F2** Of these three word orders, SOV is currently the most frequent.

**F3** Changes of basic word order from SOV to SVO or VSO (either directly or through the other as an intermediary) are generally believed to occur with significant frequency.

**F4** There is disagreement on what is the most frequent direction of change between SVO and VSO is.

**F5** Changes of basic word order from SVO or VSO to SOV are generally believed not to occur with significant frequency.

I shall also make use of the following two hypotheses:

**H1** Any broad trends in basic word order change must be ascribed to functional selection.

**H2** The overall dynamics of word order change do not alter over time, i.e. whatever broad trends may be observed to exist today or inferred to have existed in the recent past have *always* existed.

I stated and justified **H1** earlier in Chapter 3, but I repeat it here so that all of the reasoning which motivates the rest of the thesis is clearly and explicitly

present in one place. **H2** is arguably an immediate corollary of **H1**, under the conventional assumption that the innate physiological and psychological endowments of human beings, which together define the cognitive functionality a language, have not changed significantly since the evolutionary origin of the language faculty. However, once again, I want all of my assumptions to be explicit.

From the facts **F2**, **F3** and **F5** and the hypothesis **H2** above, I draw the following conclusion:

**C1** In the distant past, the frequency of SOV as a basic word order must have been significantly higher than it is now.

No other state of affairs is consistent with both the high frequency of SOV today *and* the tendency for SOV languages to change to non-SOV languages *and* for non-SOV languages to rarely change to SOV languages. **C1** is by no means novel; almost exactly the same steps of deductive reasoning appear in (Newmeyer, 2000), at least. However, firstly, relatively little progress has been made in explaining *why* **C1** should be true, and secondly, the implications of this conclusion for functional explanations of basic word order frequency appear to have gone curiously unnoticed. This takes me to my second conclusion.

From the facts **F1**, **F3**, **F4** and **F5** and the hypothesis **H2** above, I draw the following conclusion:

**C2** The ranking of word orders by functionality is not SOV > SVO > VSO > VOS > OVS > OSV, but rather must be of the form (SVO, VSO) > SOV > (VOS, OVS, OSV), or possibly (SVO, VSO) > VOS > SOV > (OVS, OSV)

No ranking by functionality which does not fit this form is consistent with observed tendencies in basic word order change *and* the sparsity of VOS, OVS and OSV basic word order languages today. To my knowledge, this conclusion has never previously been explicitly stated in the literature.

The basic story told by **C1** and **C2**, as shown in Figure 4.4.1, is as follows:

**S2** SOV is not the most frequently observed basic word order today because it is the most functional. Rather, the high frequency of SOV languages is because modern languages are descended from an initial stock of languages where a strong majority had SOV basic word order. SOV's current high frequency is an echo of its dominance in the distant past. SOV word order is in fact *less* functional than SVO and VSO word order, so that there is a tendency for SOV languages to change over time into SVO and VSO languages. This is the reason why SVO and VSO are the other two well-attested word orders today.

This story is somewhat more complicated than the story told by standard functional explanations, but it seems to be the only story which is consistent

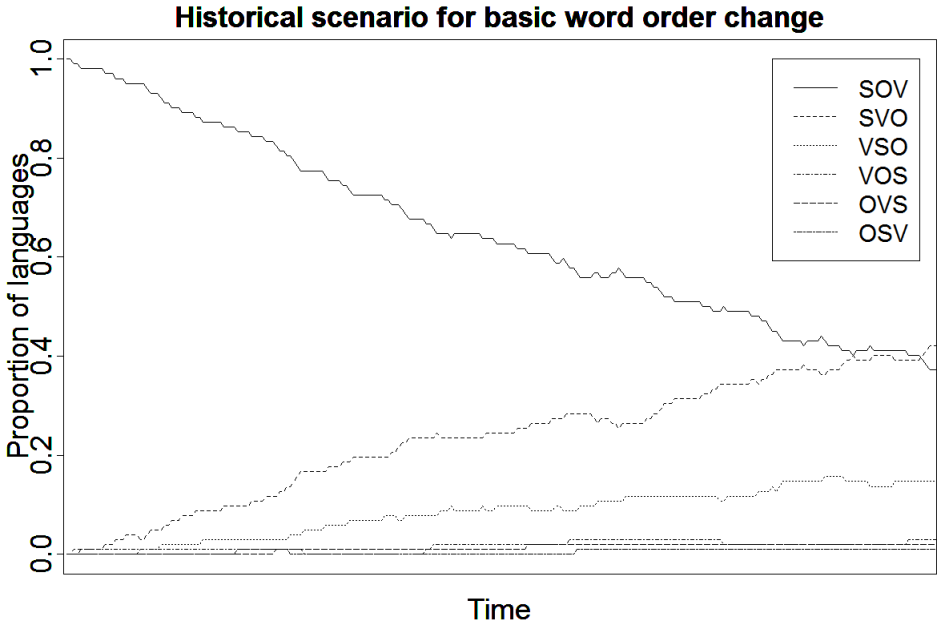


Figure 4.4.1: Alternative historical scenario which is compatible with both diachronic and synchronic evidence.

with both the current synchronic picture and what we know about diachronic trends.

The reasoning which has led to this new scenario makes it clear that a totally satisfactory, ultimate explanation of basic word order frequencies must be composed of two logically independent "sub-explanations":

**E1** An explanation for why the initial distribution of basic word order in human language was heavily skewed toward SOV.

**E2** An account of word order functionality which places SOV lower in the functionality ranking than SVO or VSO, but above the other three orders, except perhaps VOS.

### 4.4.2 Traces in the literature

For the most part, there appears to be nothing in the literature which comes close to providing a satisfying candidate for **E1** or **E2** above. However, there is not a *complete* lack of work in these directions, as I shall discuss briefly below.

With regard to **E1** in particular: there has been a sudden increase in attention to this subject in recent years, motivated primarily by the results of experiments involving gesture-based communication (I will discuss these experiments in detail in Chapter 5). Most of this work amounts to a growing consensus that there is clearly *something* special about SOV word order. The evidence for this concensus is psycholinguistic in nature, suggesting that under

certain circumstances there is either a clear preference to use SOV word order when conveying information gesturally, or that gestural communication is easiest to understand when SOV word order is used. This work is promising, in that it singles out as special precisely the same word order singled out by the quite different historical linguistics approach outlined above. However, there has been only the beginnings of speculation as to *what* is so special about SOV.

With regard to **E2**: as we saw in the previous chapter, all previous functional accounts of word order were intended to explain the basic word order frequencies from a purely synchronic perspective, and as such did not even aim to conform to the requirements outlined above. However, recall that I included in my survey accounts of word order functionality which did not yield precisely the "correct" ranking by functionality, according to the standard perspective. My motivation for this was so that I could now ask: can any of these old accounts, which would previously have been considered flawed, enjoy a new life by virtue of being correct according to our new diachronic perspective? Unfortunately, this is not the case. None of those accounts which do not quite meet the standard criteria meet these new criteria either. Krupa's account seems to come closest, in that it ranks SVO more highly than SOV, however it still ranks VSO below both of the subject-first word orders. Clearly, a totally new account is required. Following the observation above that some previously offered functional principles, such as Tomlin's verb-object binding principle, do not appear to be respected by trends in word order change, we might refine **E2** above to state that the account of functionality should be based on one or more functional principle(s) which are preferentially acquired and then maintained by trends in word order change.

I am aware of one partial account of word order functionality which is compatible with the requirements for **E2** above, which was not included in the previous chapter due to it not providing a complete ranking of all six word orders by functionality. Vennemann (1973) has argued that word order change from SOV to SVO may be functionally motivated in situations where an SOV language has (for whatever reason) lost its case marking. The functional motivation is the avoidance of ambiguity during so-called "object fronting". In object fronting, the object of a sentence or clause is moved to the beginning of that sentence or clause, regardless of its original position, usually for pragmatic reasons such as to emphasis. An example in English would be "I hate dogs, but cats I like", in which "cats" has been moved to the beginning of the phrase "cats I like", instead of being left at the end as in the more normal phrase "I like cats". Vennemann notes that, in an SOV language without case marking, object fronting does not alter the phrase structure: SOV and OSV utterances are both of the form NP NP VP, resulting in ambiguity. In SVO languages, by contrast, object fronting changes the phrase structure from NP VP NP to NP NP VP, which is disambiguous even without case marking. While it is clear that SVO has a functional advantage over SOV in this situation[4], this principle

[4]Note that this is a genuine example of language functionality as it was conceived of, but

alone cannot explain why SOV languages without case marking should change into SVO rather than VSO, which would also solve the ambiguity problem, and also cannot explain change between SVO and VSO word order in either direction. Thus Vennemann's ambiguity ideas can be considered, at best, to potentially be one part of a multi-part functional explanation for basic word order frequency.

A curious piece of work is due to Lupyan (2002). In this paper, which explores the influences of word order and case marking on language learnability, the authors train a certain class of artificial neural network, known as Simple Recurrent Networks, to incrementally assign syntactic roles such as subject, object and verb to words in artificial languages with varying word orders. They find that, in the absence of case marking, the networks are able to identify the correct syntactic roles with a success rate of 100% for SVO and VSO languages, 90% for SOV languages, 85% for VOS, 80% for OVS and 74% for OSV. Thus, when the word orders are ranked by the learnability by the networks, we get (SVO, VSO) > SVO > VSO > OVS > OSV which is perfectly compatible with the requirements for the new account of word order functionality which I have argued for. However, Lupyan and Christiansen are essentially unable to explain this ranking. Like all but the simplest of connectionist models, the SRNs used in this paper are "black boxes", and it is far from clear why any particular word order should be relatively easier or harder for them to learn. The authors speculate that "the likely reason SOV-no case grammars did not achieve perfect accuracy is because they contained two unmarked nouns prior to the verb. Since the networks learn to map different types of verbs to different argument constructions, verb-final grammars are at a disadvantage - in these grammars the grammatical role that provides the most information about that is to come, is received last", but it is not clear that this translates into a functional principle which is clearly applicable to human language learners. Furthermore, this explains only one part of the overall ranking. As tantalising as it is to see just the word order ranking which the diachronic data demand mentioned previously in the literature, there does not actually appear to be much to be learned from Lyupan and Christiansen's work for my purposes.

The only previous work I am aware of which comes even close to providing a possible instance of *both* **E1** and **E2** is (Langus & Nespor, 2010), which suggests that the high frequency of SOV and SVO basic word orders results from a "struggle" between different cognitive systems for which different word orders are better suited. Langus and Nespor suggest that SOV word order is preferred for direct interaction between the CIS and SMS of the broad language faculty, whereas SVO word order (and, more generally, any orders in which V preceeds O) is preferred by the abstract comptuational system of the narrow

rarely actually was, in the traditional school of functionalism: the need to avoid ambiguity is an essential part of communication in general, and in no way specific to the cognitive details of implementing a communication system within the framework of the human mind and body.

language faculty. I think that this description is most probably true, and is broadly compatible with the explanation I shall develop in this thesis, but it is very much a *description*, rather than an explanation. We will not have the complete picture until we understand *why* different subsystems of the language faculty prefer different word orders.

### 4.4.3 Looking ahead

Most of the remainder of this thesis is an attempt to provide each of the sub-explanations called for above, or to at least provide substantial pointers in the direction of such sub-explanations. Part II investigates the possibility of basing **E1** on a "language of thought" word order, and Part III provides **E2** via an account of word order functionality based on the Uniform Information Density hypothesis from psycholinguistics, a hypothesis which has enjoyed considerable explanatory success in recent years but which has not previously been applied to considerations of basic word order. I wish to stress three points before continuing to these sections.

First of all, the contents of Parts II and III are entirely independent. If future research should show the contents of one part to be hopelessly incorrect, this will have no repercusions whatsoever for the contents of the other part. Any explanation of early SOV dominance combined with any account of word order functionality which yields the appropriate ranking can together constitute a satisfactory explanation of basic word order frequencies.

Second of all, the claim that explaining basic word order frequencies requires a two-part approach, focused on the initial distribution of word orders and a functional explanation for subsequent change, is entirely independent of my particular solutions to these two parts. Even if the contents of Parts II and III are *both* subsequently shown to be incorrect, the necessity of a two-part approach will not change. What will be required then is simply two new explanations for the two independent questions. To discredit the overall idea of a two-part approach would require either substantial new data showing that one of the five facts **F1**-**F5** I have relied upon is false, a strong argument against my two hypotheses, **H1** and **H2**, or the demonstration of a logical fallacy in deriving my two conclusions, **C1** and **C2**, from the facts and hypotheses.

Finally, the two explanatory components offered by this thesis do not entirely supplant all of the functionalist explanations reviewed in the previous chapter. The reasoning above does not necessarily show that all of these previous accounts are wrong. The overall functionality of any language trait can, and presumably often is, a consequence of multiple competing motivations; indeed, many of the accounts I considered were composed of several such motivations, such as the three interacting principles proposed by Mallinson and Blake, Krupa or Tomlin). The fact that none of the previous accounts can explain word order change from SOV to SVO or VSO can only be strictly interpreted as meaning that there must exist at least one additional motivation

which, under certain conditions, out-competes the others to a sufficient degree that it tips the ranking so that these changes, but not the reverse changes, are possible. These additional motivations are of considerable importance, in that they alone allow us to account for the trends in word order change, but they do not necessarily invalidate the previous motivations. For instance, the previous motivations all place the object-first word orders very low in the rankings, which is perfectly compatible with the diachronic perspective I have developed here, and there is no reason why these motivations cannot assist the new motivations in enforcing this requirement. Furthermore, it is entirely possible that there are scenarios other than basic word order change (such as pragmatic considerations with certain kinds of sentences), where some of the previous motivations play a decisive role in determining outcomes. In short, the new ideas in this thesis augment previous ideas, and augment them in an important way, but do not completely discredit or displace them.

## 4.5 Summary

The general form of the functionalist explanations of cross-linguistic distribution of basic word order reviewed in the previous chapter – that is, explanations in which the $n$-th most frequent basic word order is posited to be the $n$-th most functional – has a very significant shortcoming: it is incompatibile with the observed patterns of basic word order change. SOV is the most frequent basic word order, and so purportedly the most functional in standard explanations, yet historical linguists appear unanimous in their belief that changes away from SOV (primarily to either SVO or VSO, or occasionally VOS) basic word order are significantly more frequent than changes to SOV from any other word order. By the very definition of functional selection I am using, basic word order change should be from less to more functional word orders, not vice versa, suggesting that SOV is not the most functional word order. I have argued, not originally, that in order to resolve this conflict we must conclude that the currently observed predominance of SOV basic word order is a remnant of an even stronger dominance in the past, rather than a consequence superior functionality. I have also argued, originally as far as I know, that this conclusion establishes definite new requirements for functional accounts of word order, which previous accounts all fail to meet. This motivates the search for two new pieces of theory: an explanation for why SOV word order should have dominated early language, and a new account word order functionality which is consistent with the new requirements derived from taking a diachronic perspective on the problem.

# Part II

# Initial conditions: majority descent from SOV and mental representation

# Chapter 5

# Evidence and explanations for majority descent from SOV

In the previous chapter, I argued for a two-part explanation of the observed data on basic word order which featured as a fundamental part the idea that the proportion of the world's languages having SOV basic word order was significantly higher in the past than it is today (i.e. around 0.45). While I maintain that such a scenario is the only one logically consistent with the uncontroversial fact that SOV is the most common word order today and that word order changes away from SOV are significantly more frequent than word order changes toward SOV, I grant that the claim is in fact a fairly strong one, and that it certainly seems to have been argued for on the basis of fairly little. Fortunately, I am by no means the first person to argue for this scenario, nor are the facts I considered in the previous chapter the only things to support the scenario. In this chapter I shall, to begin with, attempt to put the argument on more convincing footing by reviewing other instances of the same argument.

Before I continue, I wish to attend to a few minor matters of terminology. First of all, the scenario I am considering in this thesis is a weaker version of a much stronger scenario, which looks like this:

**S1** All of the world's languages are descended from a single common ancestor language ("proto-human" or "proto-world"), which had SOV basic word order.

I single out this scenario, known as "linguistic monogenesis", because it has been proposed in the literature previously and, while not uncontroversial, has some vocal adherents, most notably Merrit Ruhlen. However, I wish to emphasise that the work in this thesis is also compatible with similar but weaker polygenetic scenarios, including:

**S1** The world's languages are descended from multiple distant ancestors, and each of those ancestral languages had SOV basic word order.

**S2** The world's languages are descended from multiple distant ancestors, and these ancestral languages featured a range of basic word orders, but SOV was substantially more common than others.

While the question of which is these scenarios best represents the truth is surely fascinating, it is beyond the scope of this thesis and I shall remain entirely agnostic on the matter. For further discussion of the relative merits of monogenesis and polygenesis, see (Wang & Minnet, 2005). **S1** above is often referred to as "common descent from SOV". Throughout this thesis I shall use instead the term "majority descent from SOV" to encompass all of the language genesis scenarios above without preference. Firm believers in common descent should be able to simply substitute these terms without any loss of coherence.

Second of all, in this chapter I shall consider the results of several experiments which are not concerned directly with the phylogenetic origins of language in humans, but which show that SOV word order is in some sense preferred over others in various situations. These experiments are interesting and relevant here because it seems natural to consider the possibility that a common cause underlies majority descent from SOV and these present-day effects. In general I shall talk about the "privileged status" of SOV word order when I wish to refer to these effects.

## 5.1 Evidence for majority descent and privileged status

### 5.1.1 Evidence from typology and historical linguistics

I am aware of three instances in the literature where scholars have argued for majority descent from SOV (indeed, in all three cases the argument is for common descent, according to **S1** above) on the basis of either the presently observed distribution of basic word orders or this distribution *and* the apparent rarity of word order change toward SOV, compared to word order change away from it.

Givón (1979) and Newmeyer (2000) both argue for common descent from SOV on the basis of the current high frequency of SOV languages and the fact that OV $\rightarrow$ VO change appears more frequent than VO $\rightarrow$ OV. Newmeyer sets his argument out in a very similar manner to mine in Chapter 4. Gell-Mann and Ruhlen (In press) also argue for common descent, this time purely on the basis of current basic word order distribution. Note that Givon and Gell-Mann and Ruhlen disagree on the dominant direction of change between SVO and VSO (Givon favouring VSO $\rightarrow$ SVO and Gell-Mann and Ruhlen SVO $\rightarrow$ SVO), but are in explicit agreement on the issue of common descent from SOV.

One other way one might infer evidence for common descent on the basis of presently observed data is to consider the case of so-called "language isolates". These are languages which lack any known genetic relationship to any other

currently spoken language. Common examples of language isolates are Basque and Korean. It seems sensible to me to make the assumption that language isolates are relatively old, and as such considering the distribution of basic word orders amongst isolates is a very rough way of "looking back in linguistic time". Using data from (Dryer, 2008), the proportion of languages with SOV word order jumps from approximately 0.45 when considering all languages to approximately 0.70 when considering only language isolates, a result consistent with the idea of majority descent from SOV.

### 5.1.2 Evidence from signed languages

I am aware of two data points on the subject of majority descent from SOV which come from the consideration of signed languages. The first is the discovery by Goldin-Meadow and Mylander of cross-cultural word order preferences in the improvised systems of gestural communication which are spontaneously created by deaf children who have speaking parents and who are not exposed to conventional sign languages. The other is Sandler et al's investigation of a recently created sign language used in a Bedouin village in Israel with a high frequency of congenital deafness, which has acquired strict word order very early on in its evolution.

Susan Goldin-Meadow and her colleagues have dedicated a tremendous amount of time and effort to the study of improvised gestural communication in both deaf and hearing people. Among their most important discoveries are that children who are born deaf to hearing parents, and who are not exposed to a traditional sign language such as American Sign Language (ASL), will nevertheless develop gesture systems for communication. More importantly for our current considerations, it has been observed that these children use consistent ordering of gestures. (Meadow & Mylander, 1998) considers eight such children, 4 in the US and 4 in Taiwan, and observed that all of these children used the equivalent of SV order when gesturing an intransitive event, and the equivalent of OV order when gesturing a transitive event (omitting the subject due to it typically being clear from context). This overall patterning is consistent with SOV basic word order. Moreover, when the hearing parents of these children use improvised gestures to communicate with their children, they do *not* use a consistent order (Meadow & Mylander, 1983). This suggests that the SOV order exhibited by the children has not been learned from input, but is something they have imposed upon their utterances by themselves. This suggests that the SOV word order is in some sense "natural".

The second data point is from Sandler, Meir, Padden, and Aronoff (2005), who describe the Al-Sayyid Bedouin Sign Language (ABSL). Al-Sayyid is a Bedouin village founded around 200 years ago in the Negev region in Israel, by a single husband and wife pair. The village today has a population of around 3,500, with the youngest generation being the 7th in the village's history. Two of the founding couples' sons were born with a congenital condition leaving

them deaf at all frequencies from a very young (pre-linguistic) age. An unusually high level of intra-village marriage, due to the village's geographical and social isolation, has led to this condition becoming quite prevalent within the village today, with around 150 currently living deaf members. This high level of congenital deafness has lead to the development of the ABSL, which is at least three generations old. Interestingly, the deaf members of the group do not bear any stigma on account of their condition, and interactions between deaf and hearing members of the village occur on a daily basis. Because of this, ABSL is used not only by the deaf members of the group, but also by many of the hearing members. With the help of hearing Al-Sayyid residents who spoke a dialect of Arabic and also understood ABSL, Sandler and his colleagues were able to analyse the grammatical structure of a corpus of 158 ABSL gestures, produced by 8 signers who belong to the 2nd of the 3 identified generations of signers.

The results of this analysis are strongly indicative of SOV basic word order. For gestures involving only a subject and verb, SV is the only attested order, with none of the 8 signers using VS. For gestures involving only a verb and an object, the order OV outnumbers the alternative VO by a factor of roughly 3.5. For transitive gestures involving subject, verb and object, the order SOV outnumbers the order SVO by a factor of roughly 3, and no other orders are attested. Overall, 136 of the 158 gestures were verb-final, with only 22 utterances placing the verb before the end of the utterance.

This finding is particularly striking for two reasons. First is the remarkable speed with which such a strong word order preference has evolved in the language. Second of all is the fact that the bias toward SOV word order in particular cannot be attributed to any external influence. Sandler et. al. identify four languages present in the area around Al-Sayiid. These are: the local dialect of Arabic, with SVO basic word order; classical Arabic with VSO; Hebrew with SVO, and another signed language, Israeli Sign Language, which permits either SVO, OVS or SOV, but with SOV being quite rare compared to the others. The fact that ASBL has developed SOV basic word order so quickly and with a near-total lack of external influence in this direction suggests that this is a "natural" development, the same conclusion that we might draw from Goldin-Meadow et. al.'s findings on home sign languages.

### 5.1.3 Experiments showing privileged status

Beyond the evidence from typology, change reconstruction and signed languages, which bear on language directly, I am aware of two experiments which point to what I have called the privileged status of SOV. The experiments are at least partially concerned with improvised gestural communication, and as such it seems to me entirely plausible that the results of these experiments and the findings above on home signing languages and ABSL have a common cause, and that this cause may also be able to explain the hypothesised SOV word

order of the first spoken language(s).

The first experiment is reported by Meadow et al. (2008). The participants in this experiment were 80 speakers of 4 different languages, with 10 speakers of each language in each of 2 experimental conditions. The four languages were English, Spanish, Turkish and Chinese. All of these languages use SV order for intransitive utterances[1], but differ in the word order used for transitive sentences. English and Spanish use SVO order regardless of whether the action being undertaken happens in place or while moving through space, and Turkish uses SOV in both of these conditions. Interestingly, Chinese uses SVO for actions performed in place and SOV for actions performed while in motion. All participants in both conditions were shown, in random order, a total of 36 vignettes, 20 depicting events usually described with intransitive sentences and 16 corresponding to transitive sentences, some of each kind ocurring in place and some while in motion. In one condition, participants were asked to convey the meaning of each vignette to another participant using only hand gestures, without speaking. This task is non-verbal, but still communicative. In the other condition, participants were asked to recreate the meaning of each vignette by vertically stacking transparencies, each of which had drawn on it one element of the vignette's meaning (either a drawing of the subject or object, or an arrow indicating the verb). This task was both non-verbal and non-communicative. In both conditions, participants were given no indication that the order in which they gestured or stacked was of any relevance.

The surprising result of this experiment is that none of the participants showed a significant tendency to use the ordering of their native spoken language when either communicating with gestures or stacking transparencies. Based on the orderings of the different languages involved, one may have expected some participants to use SOV order exclusively, some to use SVO exclusively and some to use either SOV or SVO depending on the vignette. In actual fact, across all four language groups, participants used SOV (or the compatible SV order for intransitive caess) more than 80% of the time.

A similar but more extensive set of experiments are reported by Langus and Nespor (2010), with results consistent with those of the Goldin-Meadow et al study. Langus and Nespor report the results of 4 experiments, each using native speakers of Italian (SVO) and Turkish (SOV). Their first experiment essentially reproduces the gestural component of Goldin-Meadow et al experiment above (using a different set of 32 vignettes, all showing events best described with transitive sentences), showing that Italian and Turkish speakers both have a strong tendency to use SOV order when communicating gesturally. Even though the events in the vignettes all required three separate gestures to convey the full event, it was found that participants sometimes used only two gestures,

[1]To their credit, Goldin-Meadow et al are careful to use the notation Ar, A and P for actor, action and patient rather than S, V and O throughout their paper, acknowledging the distinction between semantic and syntactic roles. Here I state their results using the S, O, V terminology for consistency

typically dropping the subject but sometimes the object or verb. The partial orders SO, OV and SV were used by speakers of both languages more frequently than OS, VO and VS respectively. When three gestures were used, SOV was used far more often than any other order (roughly 80% of the time for Italian speakers and 90% for Turkish speakers). In the second experiment, the same general method is extended to vignettes showing more complex events which would typically be described with sentences containing an embedded clause in the object position of the main clause, e.g. "the man told the girl (that) the child caught the fish". The question of interest is whether or not Turkish (SOV) speakers will gesture such an event in a way consistent with standard Turkish phrase structure (so that the verb "told" comes *after* the verb "caught"). The result is that they do not. From this, Langus and Nespor conclude that the sort of improvised gestural communciation used in this experiment does not involve the FLN[2], and instead involves a direct interface between the CIS and SMS. This suggests that the SOV preference is not due to any properties of FLN but must reside in CIS and/or SMS.

In two further experiments, Langus and Nespor expand significantly on the Goldin-Meadow et al study. In their third experiment, digital video footage of the gestural communication performed by participants in the first experiment is edited to produce, for each vignette, a video of that vignette's event being signed in each of the six possible word orders. The segments of footage corresponding to individual gestural components are slowed down or sped up so that, across all vignettes and word orders, each component (and hence each overall gesture) lasts for the same duration. Participants are then shown these videos on a computer and, after the video has ended, are required to click on one of two vignettes according to which vignette the video they just saw described. The result is that speakers of both languages click on the correct vignette fastest when the video uses SOV word order, suggesting that this order is preferred not only for gesture production but also gesture *comprehension*. Interestingly, comprehension times for other word orders differ somewhat between participants according to their native language (e.g. Turkish speakers seem slower to comprehend OVS and OSV videos than Italian speakers), but both Turkish and Italian speakers were slower at comprehending VSO order videos than any other. In a fourth experiment, voice synthesising software is used to generate prosodically flat utterances using Italian and Turkish words in all six possible word orders, and participants are again required to choose between two vignettes according to an utterance they have just heard. In this experiment, Langus and Nespor conclude that the FLN *is* involved: Italian speakers respond fastest when their native SVO word order is used, and Turkish speakers respond fastest when their native SOV word order is used. They also note that Italian speakers comprehended any VO orders (SVO, VSO, VOS)

[2]Recall, as introduced in Chapter 3, that Hauser et al. (2002) decompose the language faculty into an abstract computational system (FLN), a conceptual-intentional system (CIS) and a sensory-motor system (SMS)

faster than any OV orders (SOV, OVS, OSV) and that, while Turkish speakers were fastest with their native SOV, they comprehended the VO orders faster than their non-native OV orders. Thus, Langus and Nespor conclude that communication which is facilitated by a direct interface between the CIS and SMS has a strong preference for SOV word order, whereas the FLN has a preference for VO over OV word order, but this preference can be overridden if one grows up in an OV speaking community.

The only conclusion that seems sensible to draw from these two studies is that there is something about the CIS-SMS interface in the human language faculty which provides a strong bias toward SOV word order for both production and comprehension when using improvised, gestural communication, regardless of the word order of any spoken language that a person may know. It is a relatively small inductive leap to suppose that the very same something behind these results can also explain the fact that home sign languages and ABSL both appear biased in the direction of SOV. A larger leap brings us to the possibility that the same something will also produce an SOV bias in improvised *spoken* communication, accounting for the apparent fact that all spoken languages today are descended from an initial language or stock of languages with an SOV bias. This last possibility is, of course, difficult to test empirically, as improvised spoken language is highly likely to be influenced by the native languages of the participants.

## 5.2 Explanations for majority descent and privileged status

### 5.2.1 Founder effects

Perhaps the simplest way to explain majority descent from SOV is to write it off as essentially being a historical accident. This can be done by appealing to a linguistic analogue of the so-called "Founder effect" in biology. This idea is originally due to Mayr (1942):

> The reduced variability of small populations is not always due to accidental gene loss, but sometimes to the fact that the entire population was started by a single pair or by a single fertilized female. These "founders" of the population carried with them only a very small proportion of the variability of the parent population. This "founder" principle sometimes explains even the uniformity of rather large populations, particularly if they are well isolated and near the borders of the range of the species. The reef heron (Demigretta sacra) occurs in two color phases over most of its range, a gray one and a white one, of which the white comprises about 10 to 30 percent of the individuals. On the Marquesas Islands and in New Zealand, two outposts of the range, only gray birds occur,

> while the white birds comprise 50 percent on the Tuamotu Islands, another marginal population. The differences in the composition of these populations is very likely due to the genetic composition of the original founders. The same explanation probably covers most of the cases in which isolated populations of polymorphic species have a much-reduced variability.

A rather colourful example of how a linguistic founder's effect might work can be found in Newmeyer (1998): "suppose that a nuclear war wiped out most of humankind and its written history, but spared the Amazonian region of Brazil. Some centuries later, a carefully constructed sample of the world's languages would in all probability show those with OVS order to be relatively common" (the reasoning being that the OVS word order, which is very rare globally, is actually relatively common in the Amazon). With regards to majority descent from SOV, the story would be that, in the distant past, the distribution of basic word orders was quite different to what it is today, possibly close to uniform. However, some sort of disasterous event lead to the near extinction of the human race, with only a small subpopulation surviving to become the "founders" of the now recovered race. That subpopulation, by nothing more than chance, happened to come from a language community where SOV was the basic word order (or, possibly, from a number of language communities where SOV was the most common basic word order). In the same way that all humans alive today would be ultimately descended from this small group of survivors, all languages spoken today would be descended from this small number of mostly SOV languages, explaining the high frequency of that basic word order today.

On the face of it, the founder effect hypothesis seems quite capable of explaining, in principle, the high frequency of SOV languages observed today. It also receives some indirect support from the study of human genetics, which has discovered evidence for a significant population bottleneck in our species past, just as the hypothesis requires. See, for example, Haigh and Smith (1972). Speculations as to the cause of this bottleneck include a disease epidemic, climate change, or a volcanic winter (the Toba supervolcano in Indonesia has been implicated in such a scenario (Ambrose, 1998), although this has been disputed (Petraglia et al., 2007)). As fascinating a topic to speculate on as it is, the cause of the population bottleneck is irrelevant from the perspective of the idea that this population lead to a linguistic founder's effect. What do matter are the estimates of exactly how small the human population became at the bottleneck. Recent estimates vary, with estimates at the low end including between 500 and 3000 (Harpending, Sherry, Rogers, & Stoneking, 1993) or less than 2000 (Rogers & Harpending, 1992) (although probably not less than 1000 (Rogers & Jorde, 1995)) breeding females, while estimates at the high end are typically for around 10,000 breeding females (Takahata, Satta, & Klein, 1995; Erlich, Bergström, Stoneking, & Gyllensten, 1996), the highest apparently being 18,000 (Sherry, Harpending, Batzer, & Stoneking, 1997). Nettle (1999)

suggests that during the Paleolithic period, languages may have been spoken by between 1,000 and 3,000 people (this estimate is based on analysis of what is known about language distribution in aboriginal Australian societies before first contact with Europeans). These estimates are consistent with as little as a single language, and probably no more than 36 languages, passing through the population bottleneck. This is a sufficiently small number for the possibility of all currently living languages being descended from a very small stock of ancestors with a strongly skewed distribution of basic word order (or any other feature, for that matter) to be taken seriously.

The biggest shortcoming of this explanation, in my eyes, is the fact that it is entirely unable to account for the prevalence of SOV word order in signed languages and the bias toward SOV in the various experiments described above. It is, of course, possible in principle that the high frequency of SOV word order and these results have entirely independent causes, but it would be quite a concidence if the language communities which survived the population bottleneck just happened to use the very same word order which some unrelated and presumably cognitive reason caused to be preferred for improvised gestural communication. A common cause seems much more likely.

### 5.2.2 Phylogenetic development of language

Givón (1979) offers an explanation for common descent from SOV based on the long-term phylogenetic development of human language. The central assumption of this explanation is that human language, rather than being an entirely novel communicative system which appeared de novo in our species, is built upon the basis of earlier and more primitive communicative systems (such as whatever communication systems were employed by the common ancestor of humans and the other great apes). As such, the path by which more primitive communication systems evolved into human communication is supposed to have left some mark on the structure of human language. Givón then, roughly, postulates a path from communicating only via utterances of the form O, to utterances of the form SO and then to the form SOV. This path is motivated in three ways: from an analysis of the communication of canines (which Givón argues is not significantly different from the communciation of great apes), from an analysis of the development of language in human children (invoking an updated version of Haeckel's principle of ontogeny recapitulating phylogeny), and from considerations of what are claimed to be vestiges of earlier communication systems in human language as used under certain situations.

Givón's account of the transition from O to SO to SOV is essentially driven by the principle that the agent, patient and action of an event are only explicitly vocalised (or "coded") when they cannot be readily inferred from the communicative context. Various changes in the physical and sociocultural environment during phylogenetic development are held to gradually change the extent to which the different elements of an event can be readily inferred. As

a starting point, Givón considers the communication of canines, noting two things. Firstly, that it consists entirely of communication about the "here and now", things which are in the immediate perceptual experience of all participants in the communication and are in a clearly defined context. Secondly, that canines in the wild live in small groups of genetically and behaviourally homogeneous conspecifics, and inhabit broadly the same kind of terrain throughout their lives, leading to a stable system of common background knowledge among the group. In these circumstances there is little need to code either the agent or the action of an event. The agent is always physical present and involved in the act of communication - third party agents never spoken of - and is either clear from context or, in the case of imperitive commands, irrelevant because all members of the social group have largely the same capabilities and responsibilities. Verbs are easily inferred given the object, due to the relatively limited range of activities which canines can engage in. Thus, canine communication consists primarily in communicating only the object of an imperitive command, with the intended agent and the desired action inferred from context: language begins with "O".

The progression from O to SO is taken to be associated with the change to communicating about agents which are displaced in either space or time, and hence cannot be inferred from the immediate context. This change is itself taken to be correlated with a progression from imperative to declarative and interrogative communication. At this point, a "choice" must be made as to whether to adopt SO or OS order. Givón motivates SO by recourse to the supposedly higher perceptual salience of actors over patients, as well as the possibility of agent coding arising simultaneously with topic coding, with topic-comment order taken to be obviously preferential to comment-topic. The final transition from SO to SOV takes place when actions can no longer be reliably inferred from agents and patients alone. This break down of inferrability is taken to coincide with the diversification of the range of possible human behaviours, such that is not necessarily one or a few stereotypical actions that one may perform on a given object, as well as the transition from a small "society of intimates", in which each potential participant of communication is familiar with the behaviour of any potential agent, to a larger "society of strangers", in which relatively little common background knowledge is present. The use of SOV order rather than VSO is apprently not motivated by anything more than the idea that, V having been incorporated into the communicative system later than SO, V should appear after SO.

This speculative phylogenetic progression from O to SO to SOV is held to be mirrored in the ontogenetic development of language in children. Givón notes that observational data provided by Bloom (1973) reveals that children at the age of 16 months vocalise objects more frequently than agents or verbs, and also notes that once children reach the stage of combining multiple words into rudimentary sentences, they are often of the agent-patient variety, such as "Daddy, chair!". Furthermore "vestigal" traces of the initial O stage of

language are held to surface in adult communication under circumstances where S and V may be readily recovered by inference, e.g. the use of "scalpel" in an operating theatre instead of, say, "nurse, hand me the scalpel".

This explanation has a *prima facie* air of plausibility about it, but is certainly not without its shortcomings. To begin with, it is something of an evolutionary "just so" story *par excellence*: the account of the development of human language from its earliest to currently attested stages can only ever be highly speculative. Furthemore, the progression from SO to SOV rather than VSO feels undermotivated to me. The principle of "that which arose later comes after that which arose first" does not hold in the case of the progression from O to S, where it is overridden by other principles: why could this not have happened with the progression from SO to SOV? To this explanation's credit, it seems able, unlike the founder's effect explanation, to account for the SOV bias in home sign languages, the early adoption of SOV in ABSL and the SOV bias in improvised gestural communication by hearing people. After all, these are all instances of communication and so presumably rely on much of the same cognitive machinery employed in verbal communication. However, like the founder effect explanation, it cannot account for the SOV bias in *non*-communicative tasks, as demonstrated by Meadow et al. (2008), which is a significant shortcoming.

### 5.2.3 Linguistic functionalism

Another quite general possibility for explaining both majority descent and privileged status of SOV is an appeal to linguistic functionalism. This may appear quite contradictory to the overall explanatory plan I outlined in Chapter 4. Recall that the general gist was that language started off being dominated by SOV for some reason, but has since then undergone a general shift toward SVO and VSO word orders due to these orders being *more* functional than SOV. It therefore seems contradictory to suggest that the initial prevalence of SOV may have been due to its functionality. However, recall that "overall functionality", to the extent that such an idea is coherent, is really a weighted combination of many individual functional principles. Furthermore, it is possible in principle that the relevant importance of these individual principles may vary depending upon what stage of language development we consider. Suppose, for instance, that the language faculty evolved in response to selective pressure to more efficiently process improvised communication which pre-lingual humans were engaging in on the basis of existing non-language-specific cognitive faculties. It is a coherent possibility that some functional principle A was more important than functional principle B in shaping that early improvised communication, but that once a language faculty had evolved and the way language was processed changed, principle B became more important than principle A. If principle A favours SOV over SVO or VSO, but principle B favours SVO and VSO over SOV, then everything works out as required.

In light of this, any combination of the functional principles presented in Chapter 3 which favour SOV over other word orders could in principle be used to explain majority descent and/or privileged status of SOV. Such an argument, however, would need to be fairly carefully constructed. The particular set of principles implicated would need to be accompanied by a carefully reasoned argument as to why these principles would have been of importance in shaping the word order of the earliest human languages, but less important than other principles which favour a different word order and which could have driven changes away from the initial state. This explanatory approach, in general, strikes me as quite precarious. In addition to the general issue of how delicate such a strategy is, it is actually quite difficult to instantiate in practice, for two reasons.

First of all, relatively few of the functional principles we saw in Chapter 2 explicitly favoured SOV above all other word orders: the accounts of Diehl, Mallinson and Blake, and Krupa all rank SOV *below* SVO. Tomlin's principles (without Song's reweighting of their importance) place SVO and SOV on even footing. Manning and Parker's explanation clearly places SOV above SVO, but as I already discussed, this is one of the least convincing and well-motivated functional accounts of word order. Second of all, if we wish to avoid the same shortcoming as the founder's effect explanation or Givón's phylogenetic explanation, we need to restrict the range of functional principles from which we can choose to those principles which could also reasonably apply to the non-communciative transparency stacking task in Meadow et al. (2008). This rules out, for example, principles related to working memory demand when processing sentences, such as Mallinson and Blake's heavy-to-the-right principle or Krupa's argument from sentence depth.

The only functional explanation for majority descent *and* privileged status which I am convinced can be made to stick is an iconicity explanation, and even then only one of the kind I advocated in Chapter 3. Recall that there I expressed concern over the fact that iconicity arguments are often expressed in terms of the similarity of the form of an utterance to some very vague, seemingly Platonic form of its meaning, and advocated that the froms of meaning used in iconicity arguments should be cognitively grounded. An argument of this form seems to me like it is able to account for all the evidence reviewed above, and indeed the very beginnings of such an argument can be found in the recent literature. In the following section I shall flesh this argument out somewhat.

### 5.2.4 Non-linguistic representation of events

A major explanada for accounts of majority descent from or privileged status of SOV is the fact that SOV appears to be preferred even in non-communicative contexts, as shown by multiple experiments discussed above. In light of this, Stowe and Meadow (2002) suggest: "finding consistent ordering patterns in a non-communicative context suggests that word order is not driven solely

by the demands of communicating information to another, *but may reflect a more general property of human thought*" (emphasis added). Along similar lines, Meadow et al. (2008) "speculate that, rather than being an outgrowth of communicative efficiency or the manual modality, ArPA [Actor Patient Action, or SOV] *may reflect a natural sequencing for representing events*" (emphasis added). These intuitions represent the beginnings of a hypothesis which is ultimately *non-linguistic* in nature, and which can therefore explain the SOV preference in non-communicative contexts. But what precisely would such a hypothesis look like, when developed in more detail, and can it be made to fit into the overall story I have argued for so far? That is to say, is a compatibility between the structure of human thought and basic word order the sort of thing which could be expected to play an important role in shaping early language but to be less important than some competing factor(s) after the evolution of a dedicated language faculty? I shall dedicate the next chapter to investigating these questions.

## 5.3 Summary

In the previous chapter, I argued, fairly quickly, that a convincing account of basic word order frequency must include, as one of two independent components, an explanation for why the proportion of SOV languages in the distant past was substantially higher than it is today. In this chapter I have tried to build a more convincing case for this claim by adding to the evidence based on language typology and historical linguistics (which could potentially be flawed in several ways) additional evidence from emerging signed languages and behavioural experiments, whose findings are much more robust. It is possible in principle that the SOV bias in both signed languages and improvised gestural communication can be explained by something which is totally irrelevant to word order in very early human languages. However, the fact that the more solid evidence happens to point in the same direction as the less solid lingusitic evidence is, I think, suggestive of a common cause underlying everything. In this chapter I have also reviewed some proposed explanations for either majority descent from SOV or the privileged status of SOV, and concluded that the most promising explanation which is consistent with *both* phenomena is that SOV word order is the "most compatible" order with whatever means by which humans (all humans, as part of their innate cognitive endowment) represent the meanings of utterances. That is, the SOV bias stems from some property of the CIS. More colourfully: "SOV is the basic word order of the language of thought".

While this explanation seems to have the potential to stick, I think it is noteworthy that so far it has only been demonstrated very indirectly, by noting an SOV preference in language. However, the entire idea of a language of thought is that it is used for *thinking*. In other words, entirely non-linguistic cognitive

processes, such as inductive inference, involve computations performed over representations stated in the LOT. If this is genuinely the case, then the SOV order in LOT should also be detectable by analysis of non-linguistic behaviour. Experimental findings along this line would make the hypothesis linking the privileged status of SOV (and the consequent majority descent from SOV) to the nature of non-linguistic representation of events much stronger. So far as I know, no experiment along these lines has been performed previously. The following chapter describes an experiment of just this kind which I have performed.

# Chapter 6

# Seeking SOV in the mind: experimental results

In this chapter I operationalise the hypothesis (put forward previously by Meadow et al. (2008) and Langus and Nespor (2010)) that the privileged status of SOV word order is the consequence of some property of the human conceptual/intentional system (CIS). I also seekingto experimentally support or falsify this hypothesis. I attempt this within the general framework of computational theory of mind, in particular relying very heavily on the idea of mental representations with compositional structure, as assumed by Fodor's famous "language of thought" (LOT) hypothesis (Fodor, 1975, 2008)[1]. After establishing a coherent conceptualisation of what it might mean for humans to "think in SOV", I present the results of an experiment in which reaction times are recorded for a variety of non-linguistic, memory-based tasks. The results from the most basic of these tasks show support for the hypothesis. That is, people *do* appear to "think in SOV" in at least one context. The results from the other tasks are not as clear cut, although they are not incompatible with the hypothesis and they are not compatible with what we might consider the most likely altenative - that people think in SVO. This finding paves the way for constructing a candidate E1 subexplanation for this thesis' main problem of explaining basic word order frequencies. The finding also has implications for distinct areas of philosophy and cognitive science. The findings presented in this chapter provide insight into the nature of the mind in general.

## 6.1 What does it *mean* to think in SOV?

Before one can design an experiment to test the hypothesis that people think in SOV word order, it is important to address what it actually *means* to think in SOV (or, for that matter, VSO or OVS). For spoken language, word order

[1]I shall make clear precisely which of Fodor's ideas I am using shortly, but to avoid any misunderstandings from the start, let me say here that I shall not be using the doctrine of radical concept nativism, nor shall I be necessarily rejecting the broad ideas of connectionism.

is a necessary consequence of the serial nature of the communication system: it's physically impossible to pronounce two words at the same time. But this is not the case at all for thought. When one holds in one's mind the thought of a cat chasing a mouse, it is not only logically possible but indeed feels intuitively obvious and perhaps even necessary that the individual elements of the thought, namely CAT, MOUSE and CHASE, come into the mind simultaneously, as a combined whole. If thought is not serial then at first blush it seems fundamentally incoherent to speak of it as having a "word order". Thus, we must think hard about how some property or properties of the CIS could make it sensible to speak of thinking in SOV. So far as I know, no attempt at such a definition has previously been presented and, of course, as a direct consequence this hypothesis has never been subject to empirical scrutiny.

The precise nature of the human CIS is, of course, a matter of no little mystery. It is difficult to speculate about what thinking in SOV might look like without making some basic assumptions about the architecture of the mind, and particularly what it is that actually happens when one holds a thought like that of a cat chasing a mouse in one's head.

### 6.1.1 Compositional representation of events

The computational theory of mind (CTM) holds that the various processes of thought are computational in nature. That is, thought consists of the transformation of some inputs into some outputs, according to formally-defined, rule-based manipulations of certain symbols, in which the input and output are specified. A necessary consequence of the CTM is that the brain constructs *representations* of external reality, in order to faclitate thought about that reality. This is a necessary consequence because a computational mind can only operate on symbols and not on reality directly. The broad picture of how the mind works according to CTM, then, is that the physical world interfaces with the mind exclusively through certain input channels (such as the eyes, ears, etc.). These channels translate physical properties of the world into patterns of neural activity. Computational processes then act on this raw information to construct higher level representations of the world. The storage and computational manipulation of these representations, and representations derived from them, are the basis for all our behaviours. For more detailed discussion of this view of the mind, see, e.g. (Fodor, 1983) or (Jackendoff, 1992).

I think it is fair to say that the CTM is fairly well accepted; as such the existence of mental representations is also taken for granted (though not the nature of those representations. However, the precise nature of these representations is the subject of some dispute. Generally, the controversy is characterised as involving a schism between those who believe that representations are symbolic or explicit and those who believe they are sub-symbolic or implicit. For the purposes of operationalising the hypothesis that SOV is the natural order of

the CIS, I shall make only two assumptions:

Firstly, I shall assume that there exists a kind of mental representation which is used to *represent* events which is distinct from those mental representations which are constructed when one directly *perceives* an event. A clear justification for this is the straightforward fact that it seems not only possible but quite routine to remember the substance of an event without having any memory whatsoever of its direct perception. For example, I know that in Shakespeare's play *Hamlet*, the character of Hamlet's uncle murders Hamlet's father in order to succeed him to the throne of Denmark. If we accept the CTM, this mean that ssome representation of the event (HAMLET'S UNCLE, HAMLET'S FATHER, KILL) exists in my long term memory. However, this representation is certainly not visual, aural, linguistic, etc. in nature: it has been almost 10 years since I read Hamlet and I do not recall any of the lines, or even parts of those lines, nor have I ever watched or heard the play performed. Long after the representation of the event which my mind constructed during my reading of Hamlet have decayed from memory, I am still able to recall the gist of the play. This implies that these memories of events are stored in some sort of representational format which is independent of any sensory modality. It is this format which I shall be considering in this chapter, and I shall refer to it as the *event-level representation*.

The second assumption I shall make is that the event-level representation is *compositional in nature*. That is, representations of the event (CAT, MOUSE, CHASE) are constructed in some systematic way out of representations of their individual elements, i.e. of CAT, of MOUSE, etc. This assumption is somewhat controversial, and many proponents of sub-symbolic representation deny it. Fodor has argued for the necessity of this compositionality on various grounds, primarily the "systematicity" of cognition (Fodor, 1975, 1998). It is not my intention to recap his arguments here. Rather, I shall justify this assumption simply because I cannot see any way to make sense of thinking in SOV without it. As far as I can tell, in so doing I am basically adopting Fodor's LOT hypothesis. I am, however, emphatically not accepting so-called "radical concept nativism", which Fodor has claimed to be a logical consequence of LOT, an idea which has, I think, been the cause of much of the controversy surrounding LOT. Furthermore, accepting LOT does *not* require a rejection of connectionism, as is commonly assumed: it is possible (although not logically necessary) to construct a connectionist system in which concepts are represented compositionally. Fodor himself has referred to the possibility of connectionism providing an "implementation architecture for a "classical" (language of thought) model". The only view of cognition which I am necessarily rejecting in this chapter is the view in which events are somehow represented "holistically", i.e., in some way which is not systematically derived from independent representations of their constituent semantic roles.

The LOT hypothesis does not, of course, by itself immediately give us a way to make sense of mental word order. The language of thought is a lan-

guage because, like natural language, it consists of meaningful structures which derive their meaning from the meanings of their substructures and from the way in which those substructures are combined (i.e., it has syntax as well as semantics), not because it involves serial ordering of components. Certainly there is no suggestion here that thoughts are literally written across the brain, left to right or in any other fashion! But at least now we can engage with the possibility of finding some way in which the elements of an event can be meaningfully considered to be ordered in their representation, since each element is represented separately, and the relationship between the representations of agents, patients and actions is systematic and fixed for all represented events. In the next section I propose one way in which this "word order of the language of thought" can be made meaningful.

### 6.1.2 Interpreting LOT word order as differential accessibility to processing

According to the CTM, any thought processes involving remembering events, recognising events, reasoning about events, generalising from observed events to make predictions about future events, etc., involves computational processes operating on representations of those events (although it is not necessarily the case that the same representation is used to facilitate all of these processes). This implies that the only way to assign an order to the roles of an event is to do so on the basis of how they are processed. That is, it may be possible to differentiate the roles by their *differential accessibility to cognitive processing*.

Under this interpretation, thinking in SOV means that our mental representations of the agents of an event are more accessible to processing than our corresponding representations of the patient or action, and that representations of patients are more accessible than those of actions. Strictly speaking, this is not "thinking in SOV" but rather "thinking in AgPaAc (Agent-Patient-Action)". I shall play it moderately loosely with regard to this distinction, occasionally using SOV in place of AgPaAc to emphasise the intended connection to language[2]. In particular, I shall use the label "SOVLOT" (subject-object-verb language of thought) to refer to the overall hypothesis we are considering.

Just what does "accessible to processing" mean? The points below illustrate the kinds of potential findings which capture what I have in mind:

- The agent is typically recalled faster than the patient, which in turn is typically recalled faster than the action.

- At a given time after events were stored in memory, agents can be remembered at a higher rate of accuracy than patients, which in can be remembered more accurately than actions.

[2]Recall in Chapter 2 I discussed the fact that agent and subject, and patient and object, are closely correlated.

- A particular event is recalled faster when queued or primed with a reference to the event's agent than when queued or primed with a reference to the event's patient (and likewise for patients and actions).
- The mind's "attentional spotlight" naturally falls on the agent of an event, such that mental effort must be exerted to redirect it to the patient or action, with the action requiring more effort than the patient. As a consequence, tasks like answering questions about the properties of the patient of an event are more difficult than answering similar questions about the properties of the agent (and likewise for patients and actions).
- When asked to identify commonalities or generalisations across a range of different events, commonalities involving the agents are recognised faster than commonalities involving the patients, which are in turn recognised faster than commonalities involving the actions.

I am not at this point suggesting that all or even any of the above statements are necessarily true. These points are illustrative examples of the kinds of ways in which we may meaningfully say that the representations of some elements of an event are more accessible to processing than others. It may turn out that not *all* statements which are "like" the statements above turn out to be true. This would not necessarily be a fatal problem for the hypothesis. If patients and actions were differentially accessible to processing in *some* psychological processes but not in others then there might still be an overall bias on thought in the direction of AgPaAc. Before discussing the experimental results, it is important to address the issue of how ones language of thought relates to word order in spoken language. This is the topic of the following section.

### 6.1.3 Establishing a link to language

The goal of this chapter is to experimentally investigate the possibility that the word order used in improvised communication - or even in non-communicative tasks involving the representation of events - is influenced by the representations and computational processes involved in non-linguistic thought. So far, I have not discussed how this might occur. This section addresses that issue.

Let us consider the sort of information processing which must occur between the holding of an event-level representation of an event in ones mind and verbally describing that event using spoken language. Psycholinguistics is still not able to provide a detailed knowledge of this process, but there are some subprocesses which we can safely assume to be logical necessities. One is a search for lexical information. The agent, patient and action of an event must each be mapped to sequences of phonemes before the appropriate muscle movements for pronouncing the utterance can be computed and executed. Thus there are three more or less independent searches which must be completed

It is fairly straightforward to see how differential accessibility to processing could influence the order in which the results of these searches become available.

For instance, the three searches proceed serially, with subsequent searches not beginning until the previous search is complete. Suppose the lexical information for the agent is found first, and then the search for the patient information is begun, and only once that information is found is the search for the action information begun. Then it is a strict necessity that the lexical information for the agent will be available for further processing before the information for the other roles is. If the word describing the agent is to be pronounced first, as in an SOV or SVO language, then this information can start being put to use immediately. However, if the word describing the action needs to be pronounced first, as in a VSO language, then the lexical information for the agent needs to be kept stored in a sort of "mental buffer" until it is needed. In this situation, pronounciation of the utterance begins slightly later than it would have if the agent information could be used right away. The same reasoning then applies for the second item of lexical information to be retrieved.

Similar situatons can arise even if the search for lexical information proceeds not in series but in parallel (i.e. with all three items of lexical information being searched for simultaneously). Even if the searches occur simultaneously, some searches may proceed faster than others, or be begun slightly sooner than others, again resulting in differential accessibility.

This discussion is not limited in any way to spoken language. If the communication is gestural, what is needed is not lexical information but information on how to represent agents, patients and actions in gestures. In this case the process may involve less searching and more "invention", but there are still three separate instances of the same process which may finish at differnet times and thus in different orders.

If we assume that improvised communication proceeds in a "greedy" manner, with information being used as soon as it becomes available, then it is clear that the most often used word order will correspond to the order in which the various linguistic processes most often terminate. How might this order come to correspond with the order in which various non-linguistic processes, such as those which operate entirely within the CIS? One possibility is that a search for linguistic (e.g. lexical) information may initiated as a result of the termination of some non-linguistic process. Consider, for example, the case in which one first has to retrieve the event to be verbalised from memory. If the agent is retrieved first, then the search for lexical information on the agent can begin first. As a result the agent will be the first role thatis spoken or gestured. Another possible scenario is that the neural hardware dedicated to a linguistic search process is similar to neural hardware which is dedicated to a similar but non-linguistic search process, perhaps due to more recent hadware developing out of older hardware. In both of these cases, improvised communication has a preferred word order which is a direct consequence of the differential non-linguistic accessibility of agents, patients and actions.

There are presumably multiple linguistic processes which could influence word order in improvised communication, and each of these processes could

be influenced by interactions with multiple non-linguistic processes. The field is still too underdeveloped for us to make with confidence any kind of claim regarding what the most important processes are. This makes it rather difficult to interpret the patterns of experimental results. Suppose that we record reaction times for 5 different non-linguistic processes which operate over events, and find evidence of differential accessibility to processing in all of them: in one, SOV is the empirical ordering; in two others it is SVO and the final two are VSO and OVS respectively. It may be the case that the process with SOV accessibility *does* have the most significant influence on most linguistic processes, and is the true explanation for the privileged status of SOV word order. However, we could not conclude this with any certainty based on this pattern of results. Our limited knowledge of the processes involved in linguistic and non-linguistic tasks forces us to take a rather hardline stance on interpreting experimental data. The only way we could have significant confidence in the SOVLOT hypothesis as an explanation for the privileged status of SOV would be if we recorded reaction time for a wide variety of independent non-linguistic psychological processes and found an SOV accessibility ranking for all or the majority of them. On the other hand, we cannot straightforwardly dismiss the SOVLOT hypothesis if we find only a few non-linguistic processes with SOV accesibility rankings, because these few processes could in principle be the ones that are most relevant. This seems to bring the SOVLOT hypothesis dangerously close to being unfalsifiable, but it is not. Testing a variety of non-linguistic processes and finding no evidence of an SOV accessibility ranking for any of them would strongly imply that SOVLOT is false.

This is all I have to say for now about interpreting the results of experimental investigations of the SOVLOT hypothesis. There are additional subtleties which require consideration, but I shall hold off on these until we have the results of specific experiments to provide context for these discussions.

### 6.1.4 Interim summary

I have considered the problem of experimentally testing the hypothesis that the privileged status of SOV word order in improvised communication and event representation tasks is a consequence of some more general property of human thought. This hypothesis has been proposed previously (Meadow et al., 2008), although without this detailed consideration, and is an economical way to explain the wide range of SOV-related biases surveyed in the previous chapter. I have argued that this fairly vague hypothesis can be rendered precise and testable if we interpret it in terms of the differential accessibility to cognitive processing of agents, patients and actions within a standard computational theory of mind framework. I have given the label SOVLOT to this operationalised version of the hypothesis, as it relies on the primary idea behind Fodor's language of thought hypothesis: that events are represented compositionally. I have discussed some of the issues associated with experimentally testing the

SOVLOT hypothesis, and delayed some discussion until later in the chapter in the context of our specific results.

## 6.2 Experimental investigation

Given the operationalisation of word order in the language of thought developed above, it is relatively straightforward to design an experiment to test the hypothesis that the human LOT word order is SOV. One needs to present participants with stimuli depicting events, causing the instantiation of appropriate event-level representations in memory. Then the participants complete various tasks involving processing the representations. If the tasks manipulate the relative importance of processing the individual elements of the events, then analysis of the response times should allow us to discover any differential accessibility to processing and thus the LOT word order [3]. In this section I describe an experiment which instantiates this broad paradigm. The results offer some support for the SOVLOT hypothesis.

### 6.2.1 Overview

The stimuli for this experiment were 32 pictured vignettes depicting a variety of simple events, taken from Langus and Nespor (2010). Drawings were used instead of written or spoken linguistic descriptions of events in an attempt to minimise any influence due to the SVO word order of English. The 32 events depicted by the vignettes are shown in Table 6.2.1. The events are constructed from 8 distinct agents, 8 actions and 8 patients. Each agent is depicted performing two distinct actions, to two distinct patients. There is no overlap between the set of agents and the set of patients.

Participants completed three distinct tasks after being shown the 32 events in random order. The tasks were also presented in a random order, and there is no significance to the order in which the tasks are described below. The first task, which I shall refer to as the Role Recognition task, involved recognition of individual semantic roles from the events shown in the stimuli. Participants were required to give a yes or no answer to questions such as "did you see a cat?" or "did you see any throwing?". The second task, which I shall refer to as the Role Association task, involved being shown one semantic role which was present in the training stimuli and then choosing from two presented semantic roles which one was seen in the same event as the given role. In effect, the participants are being asked questions such as "you saw a cat. Was the cat involved in an event with a chimp or an elephant?". The third task, which

[3]The practice of analysing response times to make inferences about underlying mental representations and computational processes has a long and established history in cognitive psychology, dating back at least to the 1869 work of Donders, recently republished as (Donders, 1969). Notable publications in this vein include (Sternberg, 1966) and (Shepard & Metzler, 1971).

Table 6.2.1: 32 events depicted by vignettes used in experiment

| Agent | Patient | Action |
|---|---|---|
| Girl | Fish | Fish up |
| Girl | Fish | Throw |
| Girl | Ball | Fish up |
| Girl | Ball | Throw |
| Boy | Fish | Fish up |
| Boy | Fish | Throw |
| Boy | Ball | Fish up |
| Boy | Ball | Throw |
| Old man | Dog | Pat |
| Old man | Dog | Feed |
| Old man | Cat | Pat |
| Old man | Cat | Feed |
| Chimp | Dog | Pat |
| Chimp | Dog | Feed |
| Chimp | Cat | Pat |
| Old woman | Crate | Push |
| Old woman | Crate | Pull |
| Old woman | Unicorn | Push |
| Old woman | Unicorn | Pull |
| Robot | Crate | Push |
| Robot | Crate | Pull |
| Robot | Unicorn | Push |
| Robot | Unicorn | Pull |
| Woman | Elephant | Hunt |
| Woman | Elephant | Fence with |
| Woman | Rabbit | Hunt |
| Woman | Rabbit | Fence with |

I shall refer to as the Event Recognition task, involved being shown all three semantic roles of an event. The event either occured in the training stimuli or was created by taking a trained event and modifying one of its semantic roles.

For each of the tasks, the question of interest is how response times varied as a function of the particular semantic roles involved in each question. For example, in the Role Recognition task, do participants recognise the agents of the events in the training stimuli faster than they recognise the patients or actions of those events? Linear modelling and analysis of variance are used to formally address this question, with the `R` program being used to perform all statistical work.

### 6.2.2 Method

The participants in the experiment were 40 individuals drawn from the University of Adelaide's School of Psychology's paid participant pool, who were paid $5 for participating in the experiment, plus 5 friends and family of the author, making 45 participants in total. Gender was roughly balanced across participants (19 male, 26 female). Participants were asked about their native language prior to completing the experiment, and each participant's language was looked up in the WALS database (Dryer, 2008) to find its basic word order. A strong majority of 41 participants spoke an SVO native language (typically English or some form of Chinese), with the remaining 4 participants speaking an SOV native language (Dari, Urdu, Bengali or Malayalam). No participants spoke a VSO native language. Participants completed the experiment using a personal computer after reading a brief written explanation of the experiment's process.

Before being exposed to the stimuli or completing any tasks, participants were trained on the use of the terms "agent", "patient" and "action". This involved being shown a series of pictures of events in which each of the three semantic roles was explicitly identified in a caption. Participants were then shown three pictures of novel events and asked to identify the agent, patient or action by using the mouse to click on one of three buttons, each of which was labelled with one of the semantic roles from the event. Participants did not proceed to the experiment proper until they correctly answered all three of these questions. Incorrect responses caused the participants to be shown a written explanation of the terms and then to repeat the test.

Upon successful completion of the terminology test, participants were exposed to the stimuli vignettes. The vignettes were displayed one at a time, in the center of an otherwise blank white screen, in a random order. Each vignette was displayed for 5 seconds. The left-right orientation of the vignettes was randomised by applying a mirror-image transform to each original image file with probability 0.5. This ensured that the agents and patients of the events were not consistently the leftmost or rightmost objects in the event.

Participants were then presented with the three tasks described above in a random order. Screen captures showing the appearance of the three tasks are shown in Figure 6.2.1. Participants were asked 24 questions in the Role Recognition task, 12 of them involving roles which were present in the training stimuli (4 as agents, 4 as patients and 4 as actions), and 12 of them involving roles which were not present in the training stimuli (8 objects which could conceivably have been agents or patients, and 4 actions). Participants were asked 12 questions in the Role Association task, with 2 instances of each of the 6 possible pairings of base and target roles, and 12 questions in the Event Recognition task, with 3 events derived from an event in the training stimuli by changing the agent, 3 by changing the patient and 3 by changing the action, along with 3 events which were actually present in the training stimuli. In

all tasks, the screen was blanked for 2 seconds between successive questions. Response time was measured beginning from the appearance of each question. In each case responses were given by striking either the 'f' or 'j' keys of a standard QWERTY keyboard. Participants were instructed by an onscreen message before the beginning of the task to leave their fingers resting over these keys at all times throughout the task, so that response times would not include the spurious factor of time taken to move between keys. Participants were permitted to take an arbitrarily long break between the conclusion of the stimulus training and the first task, and in between consecutive tasks, by having to click on a button to advance to the next task once they were ready.

In completing any of the three tasks, the participants must be shown some representation of individual semantic roles of the events, either agents, patients or actions. The chosen modality for this was written English words. In some sense this is unfortunate, as it seems to encourage the participant to construct linguistic representations of the events in the training stimuli, and this may introduce the possibility that the word order of a participant's native language will contaminate the response time data. However, the choice is inevitable, in that the only feasible non-linguistic representation which could be used - drawings - is difficult to apply to the case of actions by themselves. Drawing a picture of a girl throwing a ball is easy, given adequate artistic ability, as are drawing pictures of only a girl and only a ball. Drawing a picture of "throwing" in isolation is, however, quite challenging. Note also that the scope for contamination from SVO native languages is essentially limited to the Event Recognition task, as this is the only task in which participants must be shown a representation of all three roles in an event, such that an ordering is in fact present. In the Role Recognition task, patients must respond to only a single word, so that no ordering can be present or implied. In the Role Association task, patients must respond to three roles, but these represent only two of the roles of a complete event, so that only a partial ordering of agent, patient and action can be implied. Also, the different roles are arranged not side-by-side in the manner of a written English sentence, but with the base role positioned above the two alternative target roles. This factor is also expected to have minimised the influence of native language.

### 6.2.3 Preprocessing of data

Prior to analysis, a number of manipulations were made to the data. These manipulations are described in this section in detail, with the results of the analyses being presented in the following section.

#### Removing data from poorly performing participants

In order to remove the influence of uncooperative participants or participants who did not adequately understand any of the three tasks involved, it was decided to discard the results of any combination of a task and participant where

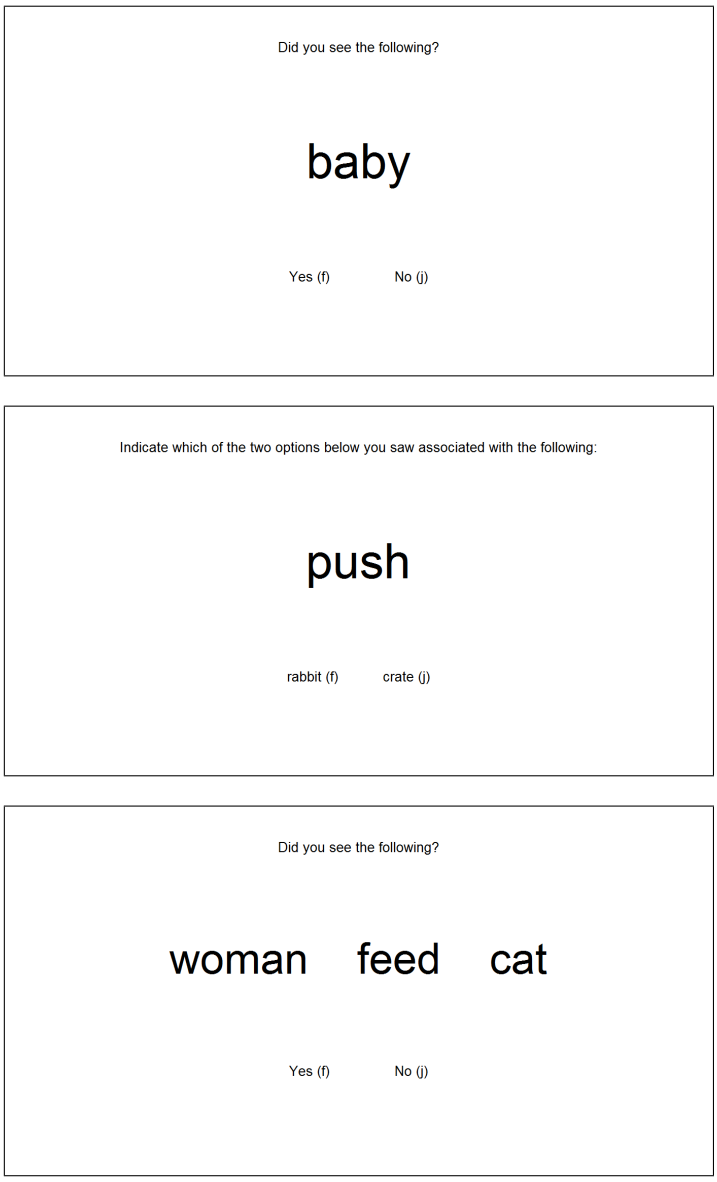


Figure 6.2.1: Screen captures of the the question asking screens for each of the three tasks in the experiment

the participant did not provide a correct answer to at least some threshold percentage of the questions for that task. In order to determine an appropriate threshold, the distribution of task accuracies for each of the three tasks must be considered. Figure 6.2.2 shows histograms for each task indicating the range of accuracies that participants attained. The histogram for the Role Recognition task shows that participants found the task relatively straightforward. No participant provided correct answers for less than 75% of the questions, which is well above chance (50%). As such, no participants were excluded from the analysis of the Role Recognition task on the grounds of having not understood the task or not made a genuine effort to perform the task. The histograms for

tasks 2 and 3 show a very different picture. In both cases there are participants providing answers with an accuracy in the vicinity of chance. In fact, in the Event Recognition task, there are participants whose performance is *below* chance. Given this it seems wise to exclude some data points, however the issue of choosing a cut-off threshold requires some consideration.

On the one hand, the higher the threshold is set, the more certain it is that the reaction times left come from participants who are consistently and accurately performing the psychological process of interest. On the other hand, the higher the threshold the less data left to analyse, and so the less the power of any statistical analyses. It is also important to note that the overall spread of the accuracy histograms for tasks 2 and 3 is significantly greater than for the Role Recognition task, suggesting that the latter two tasks are genuinely more difficult. On first consideration, we may not wish to exclude too much data produced by participants who understood the task and were cooperative but achieved low seeming accuracy due to the difficulty of the task, as these reaction times could still be informative about the process of interest. However, it is important to consider the question of how task difficulty leads to decreased accuracy, and how this affects reaction time. If incorrect responses are due to complete but "unlucky" executions of a processing routine which provides correct and incorrect responses stochastically with relatively fixed running times, then reaction times from low-accuracy participants are likely to still be of interest. However, if incorrect responses are generated, say, in response to "hunches" developed after only partially completing a processing routine then inclusion of the corresponding reaction times may obscure any reaction time regularities in the partially completed routine.

Without commiting to a model of the psychological processing which underlies responding to the questions in these tasks, there seems no principled way to decide upon a cut-off for task accuracy. As such, I have decided fairly arbitrarily to use the median task accuracy as a threshold. This approach eliminates the most extreme low-performing participants but guarantees that we will not lose more than half the data collected. On the one hand, this may be an extreme cut-off in that half the collected data is not used. On the other hand, the median cut-offs corresponds to threshold accuracies of around 77% for the the Role Association task and 68% for the Event Recognition task, which are certainly not unreasonably high. Much of a lower cut-off for the Event Recognition task and we will be including data from participants performing at around chance. I have repeated my analyses with less stringent cut-offs and found no change in the significance of the results, and so here I use the median cut-off so that the estimated values produced by linear regressions are based on data which is not of questionable reliability.

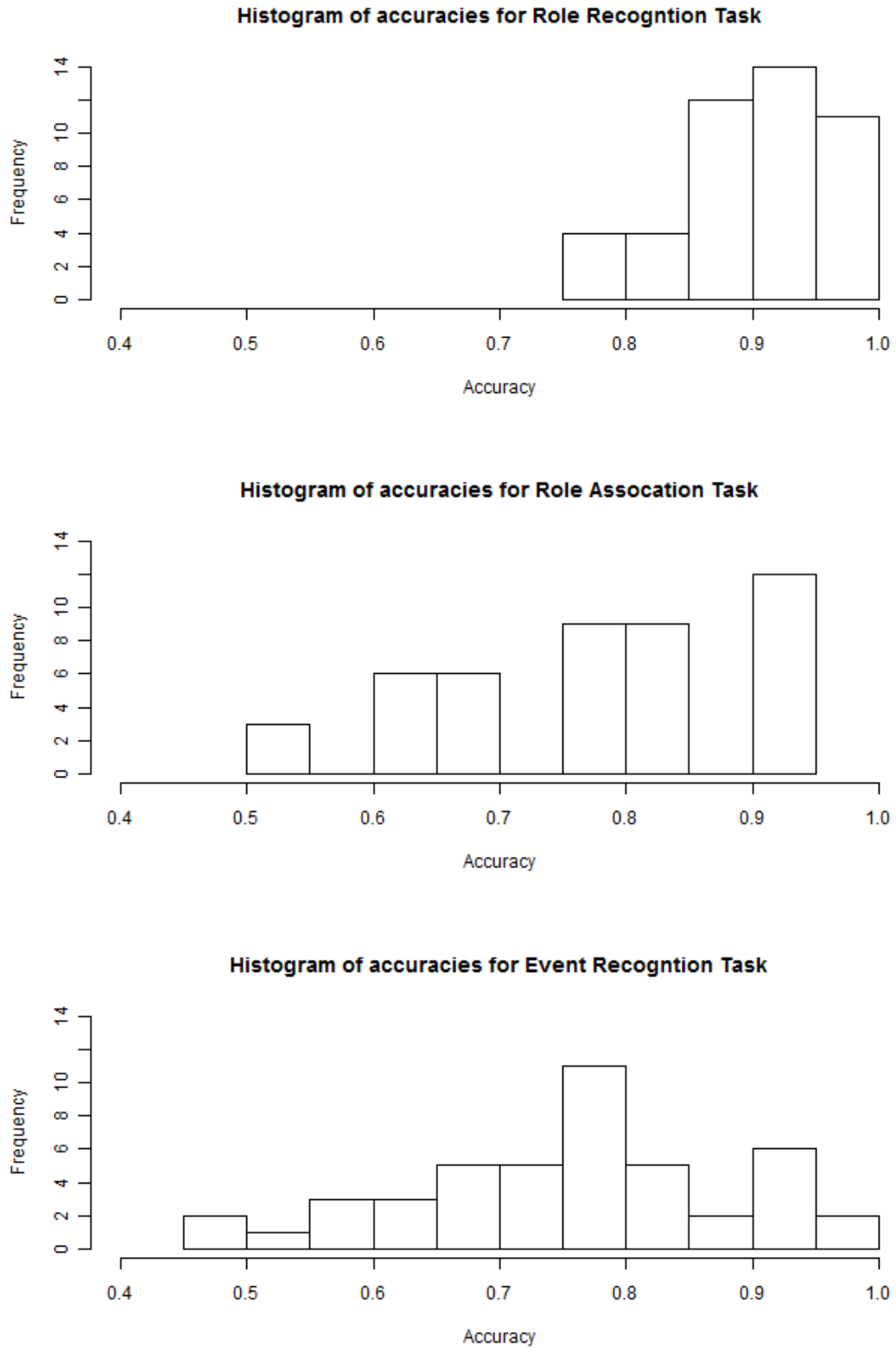


Figure 6.2.2: Histogram of overall participant accuracies for the three tasks in the experiment.

### Removing outlier data points

Having removed data from poorly-performing participants, I turn my attention to the issue of outliers. Outlying data points are a common occurrence in response time data, with various causes, including lapses of attention, inability to make a decision, quick responses on the basis of gut instincts, and more. Figure 6.2.3 shows the response time distributions for the three tasks in this experiment. In general, it can be very difficult to detect the presence of outliers on the basis of visual inspection of response time distributions, as response time distributions typically have long tails, and infrequent outliers drawn from a distribution with a mean greater than the mean for genuine responses can hide in this tail. However, in this case there are some data points which appear as obvious outliers even with only a casual inspection. This is especially true of the Role Association task, where there are some response times exceeding one minute.

There are a wide variety of techniques for deciding which points to consider as outliers and exclude from analysis, including rejecting all points which exceed some particular threshold value, rejecting all points which exceed the mean of the distribution by more than a certain number of multiples of the standard deviation of the distribution, and replacing all datapoints above some threshold with datapoints at that threshold (known as "Winsorising"). In deciding how to deal with outliers, I have been guided by the work in (Ratcliff, 1993). This work uses simulated response time data (using the ex-Gaussian distribution) both with and without outliers to investigate the effect that different strategies for excluding outliers have on the power of ANOVA analyses. The clear finding in this work is that eliminating points above a particular threshold quantity usually results in much higher power than eliminating those points above some multiple of the standard deviation. Of course, the exact cut-off point which yields maximum power varies from distribution to distribution. Ratcliff advises that tests should be performed using a variety of cut-off points and only results which remain significant over a range of non-extreme cut-off points should be reported, and also that cut-offs should be chosen so that no more than about 10% of data points are discarded.

I have decided to be fairly conservative in estimating my cut-offs. I have chosen cut-off response times of 8 seconds for the Role Recognition task, 20 seconds for the Role Association task and 12 seconds for the Event Recognition task. This reduces the number of data points for the three tasks from 1080 to 1075, from 540 to 536 and from 624 to 613, respectively, corresponding to a loss of data of less than 2% in all cases. These cut-off times were chosen simply on the basis of visual inspection of the response time distributions. In all three cases, the cut-offs eliminate the most obvious and extreme outliers from the distributions. It is entirely conceivable that data points still remain after this cut-off procedure which should properly be considered outliers, in the sense that their response times are influenced by processes beyond those

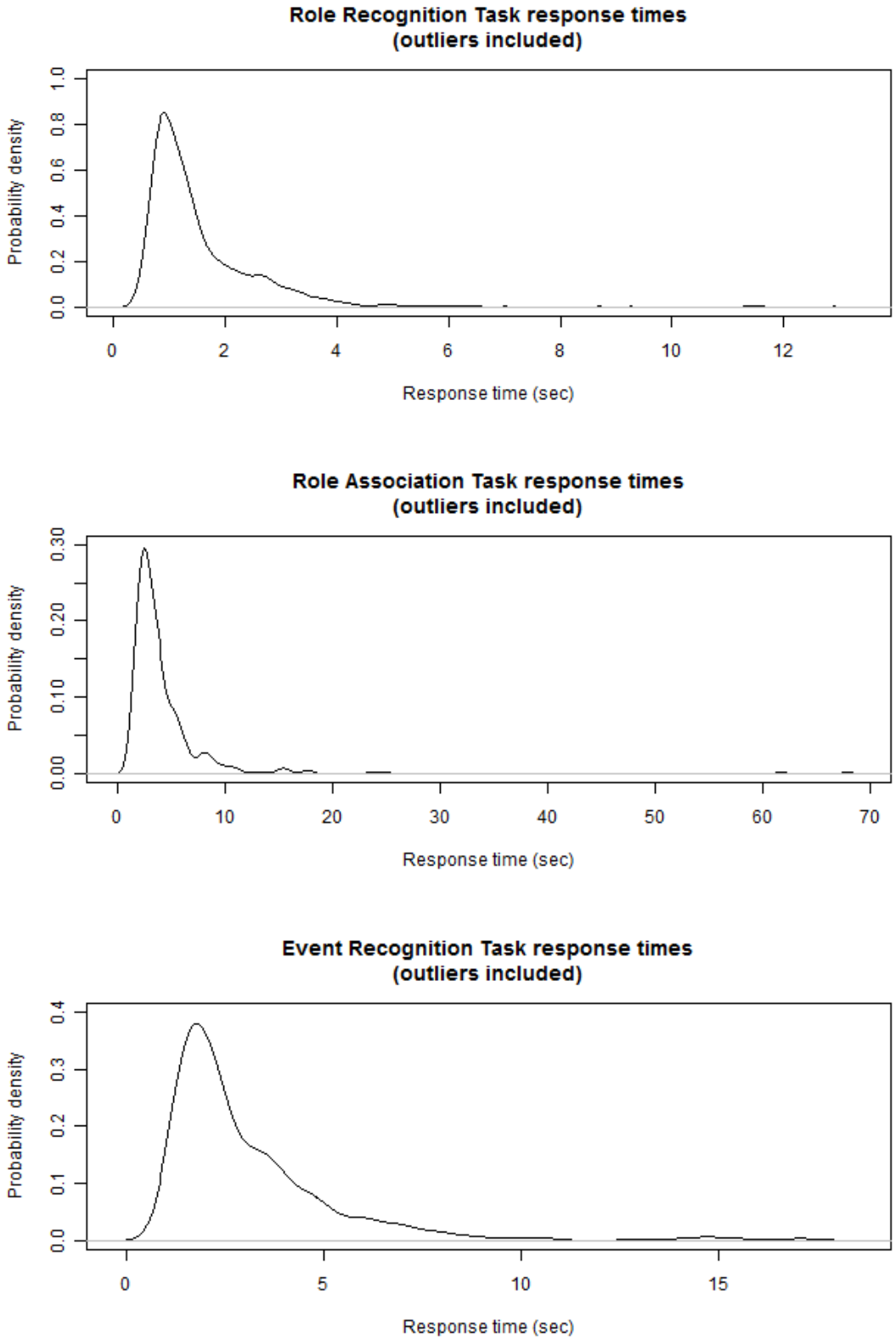


Figure 6.2.3: Response time distributions for different tasks, with outliers included.

Table 6.2.2: Shapiro-Wilk normality test p-values for various transformations of response time data

| Transform | Role Recognition | Role Association | Event Recognition |
|---|---|---|---|
| No transform | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ |
| Logarithm | $< 2.2 \times 10^{-16}$ | $2.15 \times 10^{-07}$ | $2.612 \times 10^{-09}$ |
| Square root | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ | $< 2.2 \times 10^{-16}$ |
| Inverse | $1.016 \times 10^{-10}$ | $9.68 \times 10^{-11}$ | $4.423 \times 10^{-10}$ |
| Inverse square root | $9.428 \times 10^{-10}$ | 0.1486 | $2.374 \times 10^{-05}$ |

we are interested in measuring. Indeed, the reaction time distribution plots in Figure 6.2.4 which show the results of the cut-off still include some density which looks likely to correspond to outlying points. However, as these outlying points are much closer to what is presumably the true distribution, decreasing the cut-offs to exclude these points carries a significant risk of excluding data points which really ought to be kept. I am happy leaving the remaining outliers in for reasons which will be discussed in the following section.

### Transforming data to achieve normality

The distribution of response times for each of the three tasks was investigated by visually inspecting density plots produced by the R system's `density` command, which uses normal distribution kernels to estimate probability densities from histograms . These density plots are those shown above in Figure 6.2.4. The plots show that the response time distribution is distinctly non-normal, with a strong rightward skew and long tail. This is typical of response time data. Since standard statistical analysis techniques assume normally distributed data, the first step in analysis is to investigate various transformations of the response time data in an attempt to produce data which is closer to normally distributed. The transformations considered were the logarithmic transform ($t \mapsto \log(t)$), the square root transformation ($t \mapsto \sqrt{t}$), the inverse transform ($t \mapsto 1/t$) and the inverse square root ($t \mapsto 1/\sqrt(t)$). The inverse logarithm transformation was not considered as it is not well-defined for all of the recorded data - respones times very close to one second get mapped to logarithmic response times very close to zero, which causes extremely large inverse values, exceeding the capacity of a 32 bit floating point number. For each transform, visual inspection of normal quantile-quantile plots and application of the Shapiro-Wilk normality test were used to assess the extent to which the transformed data was normally distributed. The QQ plots are reproduced in Figure 6.2.5 and the Shapiro-Wilk $p$-values are shown in Table 6.2.2. For all 3 tasks, the inverse square root transformation was shown to yield the most normal distribution of response times, and so was used for all analyses.

There is a happy side-effect of choosing an inverse transformation to bring the response time distribution close to the normal distribution. In addition

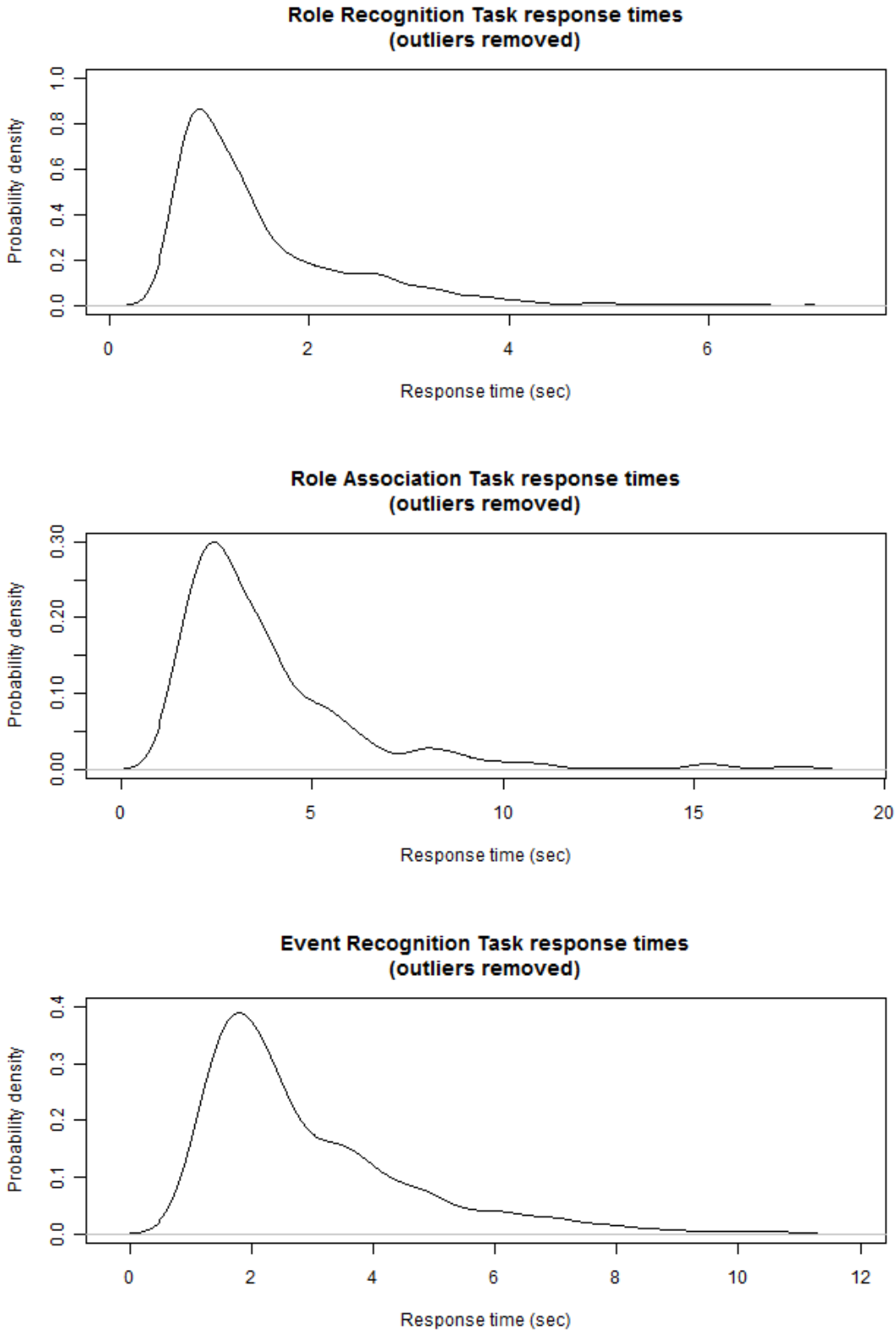


Figure 6.2.4: Response time distributions for different tasks after removal of outliers.

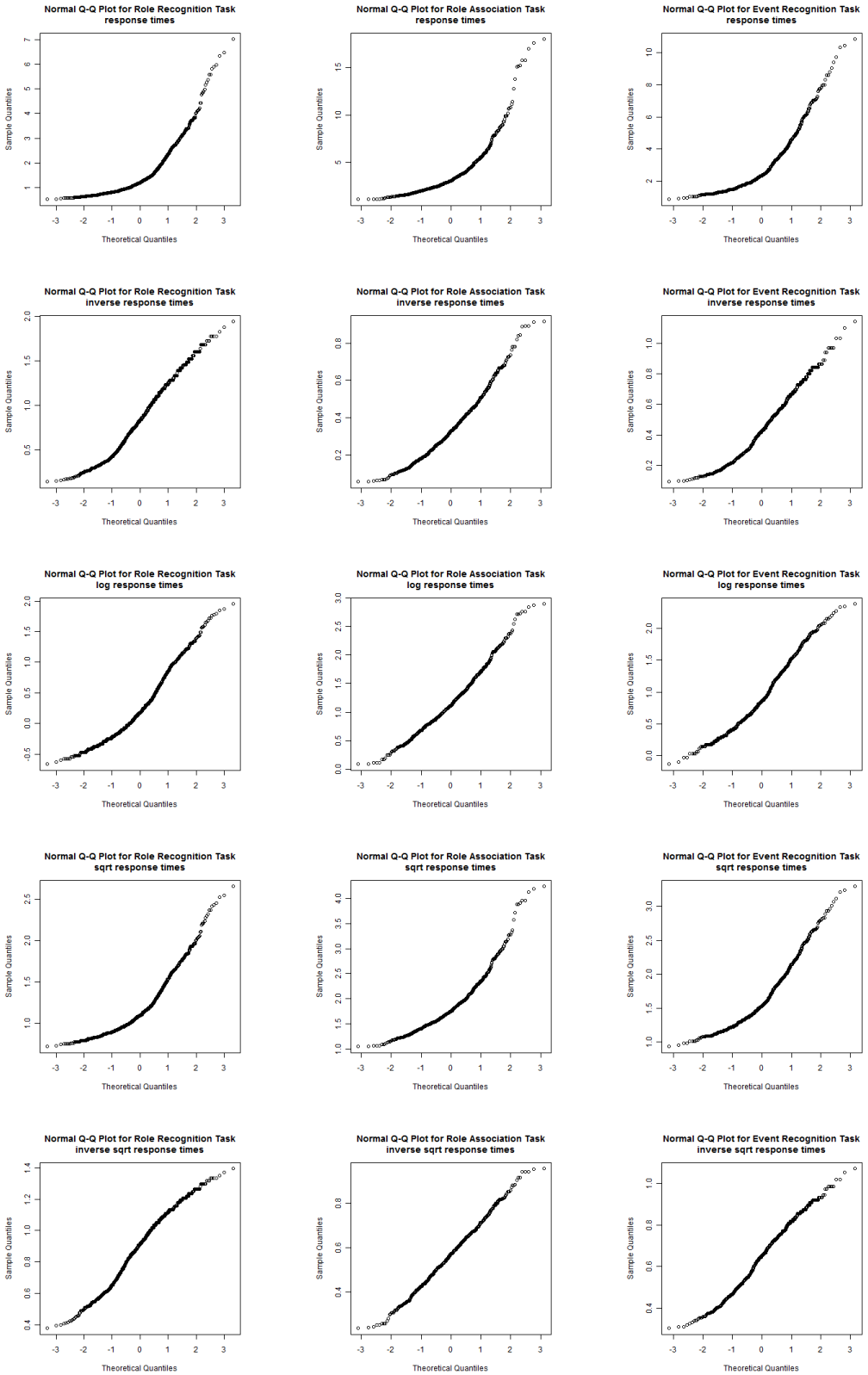


Figure 6.2.5: Normal quantile-quantile plots for various transformations of the response time data. The three experimental tasks are arranged in columns, the five transforms are arranged in rows.

to various methods of excluding data points, Ratcliff (1993) has also considered the effects on statistical power of applying various transformations to data which have the effect of minimising the influence of outliers, and one such transformation is the inverse (non-square root) transformation. Because response time distributions usually have quite low means and outliers are typically higher than the mean rather than lower, the inverse transform maps large outliers to comparatively small values which are much closer to the transformed value of the mean. In fact, the simulations in Ratcliff (1993) show that performing ANOVA on inverse transformed data yields only slightly less power than testing non-transformed data which has been trimmed using the optimal cut-off value for a specific distribution. The inverse square root transform should have the same effect (the square root operation prior to inversion compresses, rather than expanding, the range of response times, so it does not interfere with the relevant effect of the inversion). This is the reason I was happy to perform only a very conservative trimming of the data in the previous section. The most extreme and obvious outliers have been removed completely, and the remaining less-extreme outliers have had their effects minimised. This approach allows a significant reduction in the risk of coming to a false conclusion due to the influence of outliers, without running the risk of choosing too extreme a cut-off and throwing away useful data.

### 6.2.4 Results

#### Role Recognition task

The response time data for the Role Recognition task, after the pre-processing described above, was subject to an analysis of variance based on a linear model in which inverse square root response time was the response variable and which included the following explanatory variables: experimental condition (Ag, Pa, Ac, FalseObj or FalseAct), correctness (whether the response given was correct or not), order (the order in which this task was performed relative to the other tasks), the participant's gender, whether or not the participant was a native English speaker and the participant's native language word order (SVO or SOV). The output of the R `summary(lm)` call is shown in Figure 6.2.6.

Almost all of the factors included in the model were significant at the 0.05 level, the only exception being the native English speaker indicator. The result of primary interest here is the significant difference in response time across conditions. In particular, observe that the true agent, true patient and true action conditions all yielded significantly different reaction times, entirely consistent with the hypothesis of a differential accessibility to processing of the individual semantic roles of an event. Looking at the estimated coefficient values, and remembering that faster inverse square root response times correspond to slower absolute response times, we can see that participants responded fastest to true agents ($\simeq 0.16$), second fastest to true patients ($\simeq 0.13$) and slowest to true actions ($\simeq 0.07$). In other words, the ranking of roles by accessibility is Ag

```
Call:
lm(formula = invsqrtRT ~ Condition + Correctness + Order + Gender +
    NativeEnglish + NativeWordOrder, data = exp1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63208 -0.13359  0.02172  0.13900  0.45961

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.54628    0.03252  16.796  < 2e-16 ***
ConditionFalseObj   0.09210    0.01728   5.329 1.20e-07 ***
ConditionTrueAc     0.06729    0.01992   3.379 0.000755 ***
ConditionTrueAg     0.16452    0.01996   8.244 4.87e-16 ***
ConditionTruePa     0.13059    0.01991   6.559 8.42e-11 ***
Correctness1        0.08139    0.02413   3.373 0.000770 ***
Order1              0.14738    0.01504   9.802  < 2e-16 ***
Order2              0.11911    0.01426   8.351  < 2e-16 ***
GenderM             0.02053    0.01200   1.710 0.087575 .
NativeEnglish1      0.00224    0.01297   0.173 0.862866
NativeWordOrderSVO  0.12415    0.02207   5.625 2.36e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1878 on 1064 degrees of freedom
Multiple R-squared: 0.2018,     Adjusted R-squared: 0.1943
F-statistic:  26.9 on 10 and 1064 DF,  p-value: < 2.2e-16
```

Figure 6.2.6: Results of linear model fitting for Role Recognition Task response time data

> Pa > Ac. This is precisely what the SOVLOT hypothesis would predict. When we also consider the significant difference in response time to the false stimuli, false objects (which, in general, could have conceivably been agents *or* patients) were responded to faster than false actions. This too is consistent with an Ag > Pa > Ac accessibility ranking.

### Role Association task

The response time data for the Role Association task, after the pre-processing described above, was subject to an analysis of variance based on a linear model in which inverse square root response time was the response variable and which included the following explanatory variables: the base role (Ag, Pa or Ac), the target role (Ag, Pa or Ac), interactions between base and target role,

```
Call:
lm(formula = invsqrtRT ~ Base * Target + Correctness + Order +
    Gender + NativeEnglish + NativeWordOrder, data = exp2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.3444007 -0.0970311 -0.0001048  0.0817910  0.3724920

Coefficients: (3 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.485276   0.036214  13.400  < 2e-16 ***
BaseAg             -0.037692   0.020457  -1.842 0.065968 .
BasePa             -0.025918   0.028781  -0.901 0.368262
TargetAg           -0.071820   0.028803  -2.493 0.012957 *
TargetPa           -0.016202   0.020320  -0.797 0.425597
Correctness         0.040460   0.016474   2.456 0.014372 *
Order               0.000408   0.007165   0.057 0.954609
GenderM            -0.013013   0.012310  -1.057 0.290944
NativeEnglish1      0.038256   0.013100   2.920 0.003646 **
NativeWordOrderSVO  0.086653   0.022353   3.877 0.000119 ***
BaseAg:TargetAg           NA         NA      NA       NA
BasePa:TargetAg     0.040974   0.035230   1.163 0.245349
BaseAg:TargetPa           NA         NA      NA       NA
BasePa:TargetPa           NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1359 on 525 degrees of freedom
Multiple R-squared: 0.1068,     Adjusted R-squared: 0.08979
F-statistic: 6.278 on 10 and 525 DF,  p-value: 4.047e-09
```

Figure 6.2.7: Results of linear model fitting for Role Association Task response time data

correctness (whether the response given was correct or not), order (the order in which this task was performed relative to the other tasks), the participant's gender, whether or not the participant was a native English speaker and the participant's native language word order (SVO or SOV). The output of the R `summary(lm)` call is shown in Figure 6.2.7.

In contrast to the Role Recognition task, very few of the factors included in the model were significant at the 0.05 level. The significant factors were the correctness of the response, whether the participant was a native English speaker and the participant's native language word order. None of the values of

the base role, target role or combination of the two were significant. However, it is interesting to note a degree of consistency in the non-significant trends of the data. Looking at the estimated coefficient values, and again remembering we are modelling an inverting function of response time, we can see that, all else being equal, participants responded to agent bases ($\simeq -0.04$) slower than patient bases ($\simeq -0.03$), and these slower than action bases (0, contrast condition), and also that participants responded to agent targets ($\simeq -0.07$) slower than patient targets ($\simeq -0.02$) , and these slower than action targets (0, contrast). That is, when considering both base roles and target roles, there was a consistent (though non-significant) tendency to respond fastest to actions, then patients then agents. This is precisely the *reverse* ranking which was found in the role recognition test, an unexpected result which I shall discuss later, in the additional context of the results of the analysis of the Event Recognition task.

### Event Recognition task

The response time data for the Event Recognition task, after the pre-processing described above, was subject to an analysis of variance based on a linear model in which inverse square root response time was the response variable and which included the following explanatory variables: experimental condition (Ag, Pa, Ac, FalseObj or FalseAct), correctness (whether the response given was correct or not), order (the order in which this task was performed relative to the other tasks), the participant's gender, whether or not the participant was a native English speaker and the participant's native language word order (SVO or SOV). The output of the R `summary(lm)` call is shown in Figure 6.2.8.

The results of this analysis are similar to that of the Role Association task, in that most of the the interesting factors included in the model were not significant at the 0.05 level. The significant factors were the "true event" condition (i.e. the condition in which the displayed event *was* present in the training stimuli), the correctness of the response and whether the participant was a native English speaker. There was no significant difference in the response times, all else being equal, to false events in which the agent, patient or action had been changed from that of an otherwise identical true event. However, it is interesting to once again note a degree of consistency in the non-significant trends of the data. Looking at the estimated coefficient values, and again remembering we are modelling an inverting function of response time, we can see that, all else being equal, participants responded to false events constructed by changing the agent ($\simeq -0.012$) slower than those created by changing the patient ($\simeq -0.006$), and these slower than those where the action was changed (0, this being the contrasted condition). This is precisely the reverse of the ranking which was found in the role recognition test, and which the SOVLOT hypothesis predicts, and precisely the *same* ranking which was found for both the base and target conditions of the Role Assocation task.

```
Call:
lm(formula = invsqrtRT ~ Condition + Correctness + Order + Gender +
    NativeEnglish + NativeWordOrder, data = exp3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.358944 -0.116017  0.008378  0.114439  0.425868

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         0.538378   0.028825  18.677  < 2e-16 ***
ConditionAg        -0.012084   0.021156  -0.571 0.568067
ConditionPa        -0.006290   0.020906  -0.301 0.763616
ConditionTrue       0.031670   0.017207   1.841 0.066182 .
Correctness1        0.043812   0.013387   3.273 0.001126 **
Order1              0.068006   0.018264   3.723 0.000215 ***
Order2             -0.001245   0.015011  -0.083 0.933940
GenderM             0.012367   0.013133   0.942 0.346728
NativeEnglish1      0.071792   0.014521   4.944 9.93e-07 ***
NativeWordOrderSVO  0.015520   0.023175   0.670 0.503298
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.15 on 603 degrees of freedom
Multiple R-squared: 0.1202,     Adjusted R-squared: 0.1071
F-statistic: 9.154 on 9 and 603 DF,  p-value: 4.701e-13
```

Figure 6.2.8: Results of linear model fitting for Event Recognition Task response time data

### 6.2.5 Alternative analyses

All of the analyses described above for the three experimental tasks use simple linear regression. While this technique is widely known and use, making results such as those given above easily accessible to a broad audience, there are some definite technical shortcomings. Strictly speaking, one of the assumptions underlying the statistical validity of linear regression is violated in this experiment. Linear regression assumes that all samples are statistically independent of one another. This is not true in the present case as I have collected multiple samples from each individual paticipant. These samples are not statistically independent because the myriad individual-level factors which influence a participant's response times (such as individual differences in memory, attention, reaction time, etc.) remain constant or near-constant for the duration of the experiment.

The ideal treatment of this data using more modern statistical techniques would involve the use of mixed-effects modelling (see, e.g., (Baayen, Davidson, & Bates, 2008)), in which multiple samples of data for individual participants and individual stimuli can be analysed without violating assumptions of the model. However, the currently established best-practice for using mixed-effects modelling is that the model's random effects structure should be maximised. This means that the model should be afforded every allowance to attribute variation in response time to peculiarities associated with individual participants, individual stimuli and individual experimental conditions. Unfortunately, data collection during the experiment was not rigorous enough to allow this appropriate level of structure in mixed-effects modelling. In particular, no record was kept of the particular stimuli for each response time, only whether it corresponded to an 'S', 'O' or 'V' experimental condition. Between this shortcoming of data collection, other shortcomings of the stimuli discussed below and the relatively low sample sizes for the experiment in general, I have decided to defer a more modern analysis of data for the time being. A replication of this experiment with more carefully selected stimuli, more participants and better data collection, analysed using mixed-effects modelling, would constitute much more reliable evidence on the SOVLOT hypothesis than that presented here. I view this replication as a prime candidate for future research continuing the work started in this thesis.

### 6.2.6 Discussion

Participants in this experiment were required to perform a total of three distinct tasks. One of these, the Role Recognition task, can be considered substantially simpler than the other two, on such grounds as the fact that substantially less complex stimuli were presented to participants in each instance of the task (one semantic role vs three) and the fact that participants typically achieved higher accuracy with this task, with all participants performing well above chance. As one should perhaps expect, analysing the reaction times from this task gave the most satisfying results. There were very significant differences in response times, all else being equal, corresponding to the interesting differences in experimental condition. Furthermore, these differences were all completely consistent with the most straightforward interpretation of the hypothesis that SOV is the word order of the language of thought: response times for agents were faster than those for patients, which were in turn faster than those for actions. One limitation of the Role Recognition task which should be acknowledged is the fact that the sets of all agents and all patients from the events in the stimulus set are completely non-overlapping. This raises the possibility that the different response times for agents and patients was not driven by their different semantic roles, but rather by variation in some other property between the two sets. For instance, all of the agents are, of necessity, animate (either people or animals), whereas some of the patients are inanimate (such

as balls or crates). If response time is influenced by degree of animacy then the fact that patients have a lower mean animacy than agents may explain at least some of the observed effect. It is unclear how this shortcoming could be easily avoided, since the set of agents will always have a higher mean animacy than the set of patients unless highly contrived stimuli are used. Some participants can, of course, conceivably play the role of either agent or patient. A future experiment may involve two distinct sets of stimuli, with some participants playing different semantic roles in each set, so as to potentially rule out this possibility. If a given participant elicits a comparable mean response time regardless of the role in which it is presented to participants, then an explanation based on something such as animacy or salience must be appealed two. However, if the same participant elicits different response times in different roles then the result reported here is substantially strengthened.

The other two tasks, Role Association and Event Recognition, are clearly more complicated than Role Recognition, as participants were presented with substantially more complex stimuli, took longer on average to respond and often achieved substantially lower accuracy than for the Role Recognition task. Correspondingly, analyses of the results for these tasks were more difficult to interpret. Essentially no significant effects on reaction time were found for any of the experimental conditions (with the exception in the Event Recognition task of true events having significantly different respones times from false events). However, it is very interesting to note that three separate non-significant trends - in the effects on response time of the base role and target role in the Role Association task and of the changed role for false events in the Event Recognition task - are all trends in precisely the same direction, so that response times for agents were slower than those for patients, which were in turn slower than those for actions. The fact that this exact same trend appeared three times, rather than three distinct and unrelated trends, suggests that these trends may represent a genuine effect and are non-significant only due to the decreased sample size for these tasks (as many participants had their data rejected due to low accuracies).

In discussing the results of this experiment, the first thing I want to draw attention to is the simple fact that the analysis of the recorded data for the Role Recognition task yielded a statistically significant result which is very clearly consistent with the existence of an SOV accessibility ranking for the process or processes underlying this task. This may seem like a fairly insignificant thing to focus on compared to the more complicated cases of the Role Association and Event Recognition tasks, but one should take care to properly appreciate it. There was absolutely no *a priori* reason why this had to be the case: it was entirely conceivable that none of the tasks would show evidence of such an accessibility ranking, and the SOVLOT hypothesis would be very much in doubt. There appears to be nothing in the design of the task which could have pushed the results in this direction. The finding is all the more surprising in light of the fact that the overwhelming majority of the participants spoke

a native language with SVO basic word order. Perhaps the least surprising result possible for the experiment as a whole would have been to find an SVO accessibility ranking for all 3 tasks, which could be straightforwardly attributed to native language influence. As it is, there seems to have been no significant native language influence and no other way to account for the results of the Role Recognition task other than to posit some genuine differential accessibility of agents, patients and actions in the relevant psychological processes. The fact that the simplest task used in the experiment has yielded such a clean result is definitely encouraging for the SOVLOT hypothesis.

Let us now turn our attention to the more difficult to interpret results of the experiment, namely the finding that there were no statistically significant differences in accessibility for the different semantic roles in either of the two more complex tasks in the experiment, but that precisely the same non-significant trend was found three times in the data collected from these tasks. In principle, any detailed consideration of this finding should be taken with a grain of salt. None of the differences in accessibility were statistically significant and as such the safe interpretation is that there is no differential accessibility to processing of agents, patients and actions in the processes which underly the Role Association and Event Recognition tasks. This safe stance is actually not a bad one at all for the SOVLOT hypothesis: one significant result which is very much compatible with the hypothesis and two null results. This view is relatively favourable for the SOVLOT hypothesis, and certainly better than one in which the more complicated tasks yielded significant results suggesting, say, SVO or VSO accessibility rankings. Nevertheless, the fact that precisely the same non-significant trend turned up three times, rather than three different random trends as we might expect if there genuinely was no effect, invites us to consider the possibility that, with more data, this trend would have become significant. This would leave us facing a rather different situation: one significant SOV accessibility ranking and two significant VOS accessibility rankings. What would this mean for the SOVLOT hypothesis?

It is difficult to say precisely. One possibility is that the two different accessibility rankings reflect the existence of two quite distinct sets of representations. If the two rankings were very different - say SOV and VSO or SOV and OVS - then this would probably be the most sensible interpretation. However, given that SOV and VOS happen to be mirror images of one another, it is intriguing to consider a second possibility: that there is a single set of common representations underlying all three tasks and that the two different apparent accessibility rankings are due to the nature of the processing that is defined over those representations. To see how this could happen, let us consider the case of the Event Recognition task.

On the basis of the clear SOV accessibility ranking for the Role Recognition task, we may conclude that a sensible model for the processing underlying this task is three parallel searches, initiated simultaneously, one searching all of the agents of the events in the training stimuli, one searching the patients and

another searching the actions. As soon as any of these searches encounter the stimuli presented during a task question, all searches terminate and the participant responds with "yes". If all three searches cover the entire set of events in memory, the participants responds with "no". The search over agents proceeds the fastest and the search over actions proceeds the slowest, leading to the observed SOV accessibility ranking. There is an obvious extension of this process to the Event Recognition task, with three parallel, simultaneously initiated searches over each of the roles, with the different searches taking place at different speeds. Whenever one search finds a match with the appropriate role of the event presented during a task question, the other roles of the matching event are checked, and the overall process terminates with a "yes" response if all roles check. The process can be terminated with a "no" response as soon as any one of the searches exhausts the set of events in memory. If the agent search is fastest and action search is slowest in this process, as per the Role Recognition task, we should find an SOV accessibility ranking, not a VOS ranking, so obviously this is not what is happening. However, let us consider a fairly small variation on this process, in which the process can terminate earlier under certain conditions.

Observe that, in the training stimuli events, any pair of two semantic roles, such as Girl as agent and Ball as patient, occurs either not at all or twice. Suppose that the process underlying responses to Event Recognition task questions is able to take advantage of this fact. If any of the three search processes finds two events which match the question event in two roles, it can immediately terminate the whole process with a response of "no". Let us call this "short circuiting" a search. For example, refer back to Table 6.2.1 and suppose that this table is a diagrammatic depiction of a set of event-level representations stored in a person's memory. Suppose that this person is participating in the Event Recognition task and is shown the stimulus event (Girl, Fish, Push), which has been derived from either the first or second event in the table by changing the action. Imagine a little homunculus in the person's mind scanning down the agent colum of the table, keeping an eye out for the object Girl. Whenever the homunculus encounters Girl in the agent column, it checks the corresponding patient and action values and compares them to Fish and Push, seeking an event matching the stimulus. In this case, the homunculus finds only a partial match, on Fish. After examining just the first two rows of the table, if the homunculus is wise enough it can conclude that (Girl, Fish, Push) is not in the table. This is because there are at most two events in the table involving a girl fishing, and the homunculus has seen them both. It can therefore short circuit the search. It is fairly straightforward to see that an identically functioning homunculus which is bound to the patient column of the table can also short circuit the search in a similar way, after finding two partial matches on Girl.

Now consider a homunculus who searches down the action column of Table 6.2.1. This homunculus scans down the action column looking for actions of

PUSH, which it does not find until near the bottom of the table. When it does encounter PUSH, it finds no partial matches and so cannot short circuit the search. The search, therefore, terminates only after the entire action column has been searched. This homunculus is unable to short circuit the search because the false event has been derived by making a change in the column which it is searching. This holds true in general: an agent-searching homunculus cannot short circuit an Event Recognition search if the stimulus event is a false one derived by changing a true event's agent, and a patient-searching homunculus cannot short circuit a search for a false event with a changed patient.

What implications does this have for Event Recognition Task response times? Suppose that the difference in speed between the parallel processes scanning over agents, patients and actions is small compared to the difference in how much memory must be scanned through between processes which are short circuited and those which run until all memory has been searched. In this caes, responses will most often be generated by short circuiting search processes. If we assume that search processes over agents are fastest and those over actions are slowest, in accordance with the SOVLOT hypothesis and the results of the Role Recognition Task, then it follows that short circuits by search processes scanning agents will result in the fastest response times. Second fastest will be short circuits coming from processes scanning patients, and slowest those coming from processes scanning actions. This means that when the stimulus event is one with a false agent, a response can be generated only by a short circuiting patient or action search, and this is the slowest pairing of searches! Similarly, if the stimulus event has a false action, a response can be generated by a short circuiting agent or patient search, and this is the fastest pairing of searches. Consequently, we get a "backward" response time distribution with false actions responded to fastest of all, then false patients and finally false agents. So, this slightly modified (and more efficient) process for responding to Event Recognition Task questions predicts precisely the pattern of response times found in the experiment. It is, however, still the case that agents are the most accessible role to processing and actions the least, so that the data structure can be interpreted as having SOV word order. The inverted response times is attributed to the particular details of the overall computational process applied to the data structure.

Of course, all of this is highly speculative, and smacks somewhat of special pleading. Afterall, it is entirely conceivable that we could contrive a variety of fantastical search processes such that any set of response time data could be claimed to come from the application of these processes to a single representation. The plausibility of this approach drops off the more strained the collection of processes seems, although it is hard to say just when we cross the threshold from plausible to implausible. All we can really say is that the probability of substantially distinct representations and processes, rather than subtle variations on the operation of common underlying representations and processes, increases both as we see more distinct accessibility rankings and as

the similarity between the rankings decreases. Exactly what counts as "similar" here is fairly vague: two accessibility rankings are more "similar" the easier it is to make a small adjustment to a process and thus change one of the rankings into the other. For example, it seems intuitive to me that making a small process change which mirror images the accessibility ranking is easier than making one which replaces an accessibility ranking with a completely unrelated one. In light of all this, the fact that this experiment yielded only two distinct rankings, and that they are fairly similar to one another (being mirror images), means the SOVLOT hypothesis would probably be better of under significant versions of the observed trend than it would be under any other significant results for the two more complex tasks, other than, of course, more SOV accessibility rankings. While this is perhaps cause for some mild optimism, the most appropriate response to this experiment for now seems to be threefold. Firstly, to be encouraged by the clearly SOVLOT-compatible results obtained for the Role Recognition task; secondly, to make no particular conclusions (and in particular to not rule out the SOVLOT hypothesis) on the basis of there being no significant differences in accessibility for the Role Association and Event Recognition tasks; and thirdly, to insist on more and better experiments probing this issue.

Putting aside the SOVLOT hypothesis for now, I wish to emphasise the fact that the results of this experiment have some significance for some broader issues in cognitive science. One of cognitive science's all-too-common "great debates" is over the extent to which the language which one speaks influences the way that one thinks. The stance that language has a substantial influence on thought has come to be known as the Sappir-Whorf hypothesis (and often it is defined in no more precise terms than those I have used here). The most extreme form of the Sappir-Whorf hypothesis would suggest that people literally think in their native language: that is, the representations used in, say, reasoning out the logical consequences of the propositions "Socrates is a man" and "All men are mortal", are precisely the same representations used in, say, comprehending those spoken sentences. As such, an English speaker who hears these propositions and concludes "Socrates is mortal" is not doing exactly the same things "under the hood" as a German speaker who hears "Alle Menschen sind sterblich" and "Sokrates ist ein Mensch" and concludes "Sokrates ist sterblich". At the other extreme is the stance, similar to that I have adopted here, that all humans think in a common, innate language of thought, and thus think in the same way, though they may speak about their thoughts differently. I think it is most likely the case that in this (and most of the field's other much publicised dichotomies) that neither extreme is correct and the truth is a more nuanced combination of the two[4], although closer to one

[4]For an excellent discussion of the tension between these two extremes, including a history of how consensus has swung back and forth between the two and recent experimental evidence in favour of different nuanced stances for different specific problems, see (Regier, Kay, Gilbert, & Ivry, 2010)

extreme than the other. In this case, my money is on the truth being closest to the non-Sappir-Whorf extreme, and the results of this experiment seem to be a data point in favour of this. Almost all of the participants in this experiment spoke an SVO native language, and yet in none of the tasks considered were there any significant results suggesting an SVO ranking of accessibility. Nor were there any non-significant but consistent trends in the data suggesting such a ranking either. This is entirely the opposite of what we would expect if what participants were actually storing in their memory after seeing the training stimuli were linguistic descriptions of the events in their native language.

### 6.2.7 Conclusion

I think it is fair to conclude on the basis of the results of this experiment that I have shown the existence of a moderate amount of evidence for a differential accessibility to cognitive processing of the separate semantic roles which is compatible with the word order SOV. The results are not as perfectly clear cut in this regard as one would like, as only one experimental task yielded such evidence, with the other two showing no statistically significant differences in response time for the agent, patient and action conditions. There are interesting non-significant trends in the data for the other two tasks, conceivably quite compatible with the SOVLOT hypothesis, but it would be difficult to attach strong interpretations of these trends without specific modelling of the psychological processes underlying the generation of responses for these tasks. The results found are particularly striking in light of the fact that the vast majority of participants were native speakers of an SVO language, yet there is no evidence in the data of any differential accessibility to processing compatible with SVO word order.

There is obviously tremendous scope for additional work in this area, and given both the promising nature of this early work and the fact that the SOVLOT hypothesis plays such a potentially strong role in explaining the cross-linguistic distribution of basic word orders, this additional work should be considered valuable. A replication of the above experiment with substantially more participants would be an obvious first step, as this may yield significant results for the Role Association and Event Recognition tasks. The development of stochastic models of memory search and stimulus comparison processes would also be valuable, as it may suggest an interpretation of the intriguing fact that there is evidence for both SOV and VOS accessibility rankings.

## 6.3 Summary

In this chapter I have proposed a relatively clearly defined way of operationalising the intuitive notion of "thinking in SOV", which has been previously proposed as an explanation for the observed privileged status of SOV word

order in various experiments. Namely, one can meaningfully consider humans to think in the order SOV if agents and then patients and then actions are relatively more accessible to cognitive processing than each other. This operationalisation depends crucially on the idea that events are mentally represented in a compositional fashion, i.e. that the mental representation of an event is constructed in some way from the individual representations of the semantic roles in the event. I have loosely identified this requirement with the endorsement of Fodor's infamous language of thought hypothesis, although not with the strongest form of this hypothesis or its consequences responsible for said infamy. I have also reported the results of an experiment designed to test the idea that SOV is the word order of the language of thought. This experiment involved recording the reaction times for 3 distinct memory-based tasks, and these reaction times were analysed using linear regression. The first task yielded a statistically significant result indicating that participants were able to recognise the agents of events they had been shown faster than they were able to recognise objects, and were able to recognise objects faster than actions. This result is in complete compliance with the idea that SOV is the word order of the language of thought. The second and third tasks did not yield statistically significant results, however precisely the same non-significant trend was found three separate times within this data, that trend being one of decreasing accessibility whe moving from actions to patients to agents. This trend is the reverse of the significant trend, although this is conceivably due to a different pattern of access to the same underlying data structure (which could equally well be thought of as having SOV or VOS word order depending upon perspective). These results are particularly surprising in light of the fact that the majority of the participants were native speakers of SVO languages. On the whole, the results of the experiment can be considered at the very least to be suggestive of an SOV word order in LOT. This makes the experimental work described here an important first step in moving the hypothesis that this is the reason for the observed eged status of SOV from a conjecture proposed on the basis of indirectly-relevant psycholinguistic observations to a directly observed and tested psychological phenomenon. There is obvious and extensive scope for further experimental work in this area.

# Chapter 7

# Discussion of SOV representation

## 7.1 A proposal for subexplanation E1

In Chapter 4, I established the following as one of two independent subexplanations required to construct a complete explanation of basic word order frequencies which was compatible with both synchronic and diachronic evidence:

**E1** An explanation for why the initial distribution of basic word order in human language was heavily skewed toward SOV.

Based on the ideas developed and the experimental results presented in the previous chapter, I am now in a position to offer the following proposal for subexplanation E1:

The human Conceptual-Intentional System contains, among other things, a mental representation of what I have been calling an event - that is, an occurrence involving an agent, an action and a patient - which is independent of those representations which are constructed when an event is directly perceived. This event-level representation is the format in which events typically persist in long term memory, and I have taken it to be the format which is most directly interfaced with the narrow faculty of language. In other words, the linguistic representations involved in the computational processes by which one verbally describes an event are ultimately derived from an event-level representation of the event, and comprehension by a listener of a verbally described event ultimately involves a transformation from linguistic representations to an event-level representation. The nature of the event-level representation and of various fundamental computational processes which operate on it are such that the different semantic roles in an event are differentially accessible to processing, in the sense that some roles are, e.g. processed more quickly than others, or are processed strictly before or after others, etc. In particular, agents are more accessible to processing than patients, and patients are more accessible

than actions. In other words, the event-level representation can be thought of as having the order agent-patient-action (AgPaAc). Because of this, in many behavioural tasks whose underlying information processing involves manipulation of event-level representations and whose essential physical nature involves producing a serial ordering of the agent, patient and actor of an event, the AgPaAc order is that most likely to be used. As a special case of this, improvised communication involving direct interfacing of the CIS and SMS leads to communication in the AgPaAc order or, in the case of spoken language, the SOV word order, on account of the strong statistical correlation. Assuming that human language developed out of this kind of improvised communication, we can reasonably expect early languages to have had SOV basic word order more often than not. In short and with style: SOV is the word order of the language of thought.

## 7.2 Future research

### 7.2.1 Explanada all the way down

I have argued that an SOV word order in the language of thought constitutes a compelling explanadum for the explanada of why the initial distribution of basic word orders should have been heavily skewed toward SOV. However, it is a truism that every explanadum is, itself, an explanada. My account above leads directly to an obvious follow up question: why is SOV the word order of the language of thought? This is not to say that connecting majority descent from SOV and the privileged status of SOV to facts about the human CIS is not a significant achievement: I believe it is. This connection pushes the explanatory frontline for questions about basic word order *deeper* into the mind, which is both enlightening and exciting. However, on some level this progress lacks a degree of satisfaction, in that it doesn't really tell us just what it is that is "special" about SOV. Of course, the possibility exists that there *is* nothing truly special about SOV, and the fact that the LOT has SOV order is nothing but a historical accident of human evolution. This would be, I think, something of a disappointing result, although there would be nothing to be done about it. Significant cognitive insights cannot be squeezed out of accidents, no matter how badly we may want them. Still, I can think of no reason to assume *a priori* that this will be the case, and as such it makes sense to me that research directed toward this question would be worthwhile.

In light of this, I have decided to include this fairly short section which conducts some preliminary discussion on how we might go about answering the question of why the LOT is SOV. This is obviously a question of substantial depth and scope, and as such my treatment here is going to necessarily be brief and speculative. In writing this chapter, I have been motivated by an observation due to Givón (1979): "While observed facts and facts deduced from facts are the flesh and bone of scientific inquiry, its heart and soul is

creative speculation about the facts". In the sections below I shall outline two distinct but compatible possible avenues for explaining the SOV word order of LOT.

### 7.2.2 Phylogenetic development of LOT

Recall that in Chapter 5 I discussed Givón (1979)'s explanation for common descent from SOV basic word order in terms of phylogenetic progression from systems of communication based on O utterances, through those based on SO utterances and then to those based on SOV utterances. It seems to me that one can straightforwardly tell a very similar story about the phylogenetic development of the language of thought. That is to say that, at various stages of the evolution of the human language of thought, it may not have been the case that all of the agent, patient and action of an event were always able to be represented. Perhaps the ability to represent the individual elements evolved one at a time, in response to a changing environment requiring more faithful representations of reality in order to be able to specify behavioural rules conducive to survival.

For example, consider the case of an animal evolving an instinctive response to flee upon seeing, or otherwise perceiving, a predator. The very simplest instinct of this type is equivalent to a rule instructing, e.g., "if you see a lion, run". This rule is insensitive to issues such as what the lion is doing at the time it is seen: lions chasing other members of your species and lions feeding their young are seen as equally threatening. As such, an animal does not need the ability to represent entire events in order to execute some computational process which initiates behaviour consistent this rule. It only needs to be able to represent agents (or S). Behaviour like this is sub-optimal however, in that it leads to a fleeing response (and hence energy consumption, loss of territory, etc.) when one may not be strictly necessary, such as when a lion is engaged in some behaviour which reliable indicates that it is not likely to attack. A more discerning rule requires more complicated representations, such as one in which a lion is either doing some specific action or is interacting in any way with some specific patient. A good improved rule may be "if you see a lion doing anything to another member of your species, run", since this is presumably a good scenario in which to stay put. This requires representing agents and patients (S and O). The most fine-grained behavioural rules, of course, require representation of actions too (S, O and V), so that animals can run, say, only when they see a lion stalking a conspecific, rather than merely looking at a conspecific. Of course, it may turn out to be the case that "if you see a lion stalking anything, run" (requiring S and V) works better as an instinct than "if you see a lion doing anything to a conspecific, run", so that the more typical evolutionary path is through representational schemes involving just S, then S and V, then S, O and V. Presumably the question of which progression is most likely is an empirical one for ethologists to answer. My point is simply

that representational systems in which only 1 or 2 semantic roles in an event can be represented facilitate different kinds of behaviour, and so the progression from single role representation to whole event representation may proceed along lines dictated by the differential reproductive fitness which results from these behaviours.

In order for progressive evolution arguments such as these to actually make predictions about constituent ordering, a principled account is required of where new elements are positioned relative to existing elements. The simplest principles here seem to be either consistent prepending or consistent appending. This leads to SOV order being explicable either by the pathway S, SO, SOV or the pathway V, OV, SOV. The former seems eminently more plausible, as the ability to represent actions but not their agents or patients would intuitively be of minimal utility, and it is not even immediately clear that this *could* be done.

The question arises of how we might be able to experimentally test the hypothesis that the SOV order of LOT is the consequence of evolutionary development from S to SO to SOV. Certainly this is not straightforward, and as far as I can see the answer can only come from ethology. By making careful study of the behaviour of various non-human species and inferring the computational procedures which underly this behaviour we can then infer the sorts of mental representations which must underly these computations. If this kind of research were to be performed on species of increasingly distant common ancestry with humans, and species were found which showed only evidence of SO and S representations, then this would represent a fairly strong verification of the phylogenetic development of LOT hypothesis. Needless to say, actually seeking this verification would require undertaking a research program of significant scale, with no guarantees of success.

### 7.2.3 Rational explanations

An influential perspective in modern cognitive science has been the so-called "rational" perspective, as famously outlined by Anderson (1990). Rational theorising proposes that the various faculties of the human mind should resemble optimal solutions to particular computational problems - that the mind is well adapted to its environment. The experiments of the previous chapter were based on the idea that a language of thought can only coherently be thought of as having a word order if the different "words" of the representation of an event are in some meaningful way differentially accessible to processing. A direct consequence of this is that different LOT word orders will make different sets of tasks faster or easier than others. It seems fair then to designate as "optimal" that word order which facilitates fastest or easiest completion of those tasks which are most important for survival.

As in the case of the phylogenetic development hypothesis, once again the question arises of how the rational explanation for the SOV order of LOT might

be tested experimentally.

# Part III

# Dynamics: systematic drift away from SOV and UID functionality

# Chapter 8

# Uniform information density and word order functionality

## 8.1 Introduction

In this chapter I derive a novel account of word order functionality, which in the subsequent chapter I shall endeavour to show meets the criteria established in Chapter 4. That is, it ranks SVO and VSO as more functional than SOV, which is consistent with what is known about long term trends in word order change. My account of word order functionality is based on a relatively recent hypothesis from psycholinguistics, called the uniform information density (UID) hypothesis. The UID hypothesis is a rational hypothesis of language use, which has wide-ranging scope and enjoys substantial empirical support. It is phrased in terms of the mathematical formalism of information theory, and relies crucially upon the concept of a probabilistic language. Since an understanding of these issues is essential to a proper understanding of my account of word order functionality, this chapter begins with a section which attempts to offer a brief introduction to the relevant concepts. After this introduction, the UID hypothesis is presented, with discussion of its theoretical motivations and the evidence supporting it. Finally, I develop the functionality account, first informally using intuitive examples, and then formally via a mathematical model.

## 8.2 Theoretical prerequisites

### 8.2.1 A brief information theory primer

For the most part, the information theoretic prerequisites for understanding this thesis boil down to the definition of a few functions of random variables, combinations of random variables or events in the sample spaces thereof. However, in order to provide some intuitive feel for the relevance of these functions, and also to make clearer the motivation for my eventual use of them, I shall

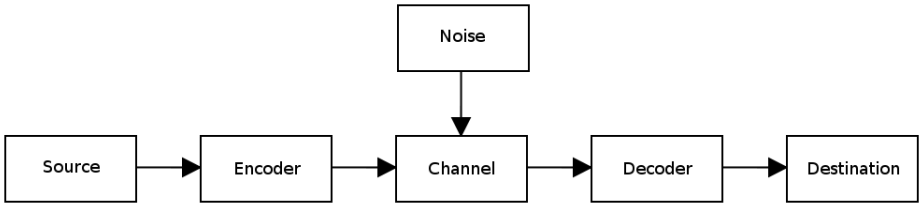


Figure 8.2.1: Block diagram showing the standard situation of interest in information theory: information from a source is encoded before being transmitted over a noisy channel, at the other end of which it is decoded to provide information at the destination which, ideally, should match the source information.

now spend some time describing what is essentially the "standard context" of information theory. This context is the problem of providing efficient and reliable communication over a noisy channel.

### Motivation

Figure 8.2.1 shows the standard situation of interest in information theory. On the left hand side of the figure is a *source*, which is essentially a discrete stochastic process, which produces a sequence $\{a_0, a_1, a_2, \ldots\}$ of symbols $a_i$ from some finite *source alphabet* $X$. The source is usually something which would be informally thought of as a "source of information", such as a newspaper being read one letter or one word at a time, a computer file being received one byte at a time over a network connection or a time series of measurements from a sensor on a space probe. The problem of interest is to reproduce the source sequence $\{a_i\}$ at some *destination*, shown on the right hand side of the figure, where the source cannot be directly observed. This reproduction is achieved by means of a *channel*, which can be thought of as representing some communication medium such as a telephone line, radio link, etc. Mathematically, a channel is defined by a finite *channel alphabet* of symbols $Y$ and a conditional probability distribution relating the channel's input to its output, $P(y_j|y_i)$. That is to say, if we put the channel symbol $y_i$ "into one end of the the channel", then we will get some other symbol $y_j$ "out the other end" with probability $P(y_j|y_i)$. If the channel's probability distribution is such that $P(y_i|y_i = 1)$ and $P(y_j|y_i = 0)$ for any $i$ and $j$, then the channel is a deterministic one whose output is always equal exactly to its input. Such a channel is said to be *noiseless*.

With a noiseless channel, the problem of reproducing the source sequence $\{a_i\}$ at the destination via the channel is simply a matter of translating symbols or strings of symbols from the source alphabet into symbols or strings of symbols from the channel alphabet, in such a way that the translation can be reversed at the destination (at least with a high degree of probability). Such a translation process is known as a *code*. A code consists of a process for *encoding* (translating strings of source symbols into strings of channel symbols) and a corresponding process for *decoding* (translating in the opposite direc-

tion).The boxes labelled as encoder and decoder in Figure 8.2.1 indicate where these processes take place in the overall process of transmission. In the case of a noiseless channel, the typical problem of concern in information theory is choosing a code which minimises the average number of channel symbols which must be sent per source symbol, so that information can be sent as quickly as possible. Codes designed to serve this purpose typically exploit statistical properties of the source process, so that common source symbol strings are assigned to shorter channel symbol strings than uncommon source symbol strings. Perhaps the most familiar example of this sort of code is the compression of files on a computer, as performed by applications like WinZip. In this case, the source is the original uncompressed file, and the channel is the computer's hard drive (on which the compressed file is stored) and the destination is the file that results from decompressing the compressed file. The source alphabet and channel alphabet are the same, and could be taken simply as the space of bits, $\{0, 1\}$, or a space of larger units such as strings of 8, 16, 32 or 64 bits. The channel is noiseless (assuming the hard drive is working properly), and the goal is to encode (i.e. compress) the file so that as few channel symbols must be sent as possible (i.e. so that the compressed file is as small as possible). The compression algorithms used by WinZip and similar programs achieve this by exploiting statistical properties of the input file, which is why some files compress better than others. This sort of encoding process is also what makes it possible for modern audio files, such as MP3s, to be so small (relative to raw audio data).

The more complicated case considers a *noisy* channel, in which there is some probability that the channel's output does not exactly match its input. This corresponds to real world situations where, say, electromagnetic interference can introduce errors in information transferred via radio waves, or sent over long unshielded cables. For example, consider the case where information is passed through the channel in a binary code, i.e. where the channel alphabet is $\{0, 1\}$. A radio link with some amount of interference may be modelled by the conditional probability distribution $P(i|i) = 0.99, P(j|i) = 0.01$ for $i, j = 0, 1$. This channel produces output matching the input 99% of the time, with a "bit flip" error occuring once every 100 bits on average. In this situation, the typical problem of concern is choosing a code which minimises the probability of an error in the reproduction of the source sequence at the destination. Codes designed to serve this purpose must necessarily insert *redundancy* into the information sent across the channel. The simplest example are the so-called *repetition codes*. As an example, if the source and channel alphabets are both $\{0, 1\}$, then a repetition code may map the source letters 0 and 1 to the strings of channel letters 00000 and 11111, respectively. If, at the destination the string 10111 is received, this is decoded as the the source letter 1, under the assumption that a single bit flip from 1 to 0 has occurred, rather than 4 separate bit flips from 0 to 1 happening in close succession. If this encoding is used to send information over a noisy channel, the source message will be reproduced

accurately at the destination unless 3 bits in 5 successive bits happen to be flipped, which is highly unlikely if the channel is only slightly noisy. Much more complicated noisy channel codes exist than these repetition codes, of course, and just as codes for noiseless channels can exploit statistical properties of the information source, good noisy channel codes typically exploit any statistical properties of the channel which might exist. For example, if a flip from 0 to 1 is more probable than a flip from 1 to 0 for some reason, then channel strings dominated by 1s are more resistant to error and so if these are used as translations of the most frequent source strings, then the overall probability of error is decreased.

Of course, in real world applications, we are usually interested in *both* of the kinds of encoding described above. On the one hand, we want each source symbol to require as few average channel symbols as possible, so that information can be sent along the channel as fast as possible. But on the other hand, we require codes to insert some degree of redundancy into the channel signal so that we can recover from errors. These two goals are, unfortunately, generally at odds with one another (notice that the repetition code discussed in the previous paragraph obtains its resistance to errors at the price of slowing down the rate at which information can be sent over the channel by a factor of five). Consequently, good code design is largely a matter of finding good compromises between these two competing goals.

As the discussion above will have made clear, the main area of application for information theory is telecommunications, and carefully designed codes underly the functioning of many modern conveniences, such as cell phones, wireless internet and digital television. However, there is nothing in the mathematics of information theory which limits its applicability to this realm: information theory works well for any kind of serial communication process. In this thesis I shall be interested in the application of information theory to human language. Here, roughly, the source and destination of concern are the minds, and in particular the conceptual-intentional systems (CIS), of two communicating people, and the source alphabet is the set of ideas which can be expressed by the CIS. The channel is comprised of the sensory-motor systems (SMS) of the communicating parties and also the physical medium of the atmosphere. The channel alphabet is the set of phonemes which the human vocal and aural machineries are capable of both producing and perceiving. Distinct human languages represent distinct codes for sending sequences of symbols from a common source alphabet (since, while natural languages vary between people, we can be assumed to all share the same language of thought[1]) through a

[1]Advocates of the more extreme forms of the Sapir-Whorf hypothesis may balk at this suggestion, but I am already committed to a universalist view of thought by my work in Part II of the thesis. Besides, even if we grant that there are some thoughts which people can/cannot think by virtue of their native language, it seems clear that the majority of thoughts which are expressed in day-to-day spoken communication are well within the conceptual repertoire of all humans.

common channel, though typically different languages use a different subset of the available channel alphabet (i.e. have different phoneme inventories). The channel is a noisy one, as words can be mispronounced and misheard due to ambient noise. The channel noise also has some statistical structure: the sound of the English letter 'b' is more likely to be misheard as the letter 'v' than it is to be misheard as 's' or 'o', for instance. The source process also has statistical structure, because not all thoughts which can be formed by the CIS are equally likely to be expressed vocally, especially in particular contexts. If we assume that languages are well-designed codes (an assumption I shall discuss in more detail later), then we should expect to be able to find evidence that languages exploit the statistical structure present in the source process and the channel noise to make good compromises between time efficiency (making it as quick as possible to express a thought) and reliability (making the comprehension of vocalisations robust against mispronounciation and mishearing). This expectation opens the door to a potentially wide-ranging series of functional explanations for language universals, where we can consider a universal explained, or at least well on the way to being explained, if it can be shown to be related to code quality.

### Concepts

In this section I shall formally define a number of information theoretic concepts which will be used throughout the rest of this part of the thesis. For the remainder of this section, unless stated otherwise, let $X$ be a discrete random variable, with outcome set $\Omega = \{x_1, \ldots, x_n\}$ and probability distribution $P$.

The *entropy* (or *uncertainty*) of $X$ is defined to be:

$$H(X) = \sum_{i=1}^{n} -P(x_i) \log(P(x_i)).$$

The entropy of $X$ can be thought of as a qualitative measure of how uncertain we are about the outcome of $X$, prior to observing it, or equivalently as a measure of how difficult it is to correctly guess what the outcome of $X$ will be. It is easy to verify that if the distribution $P$ has all of its probability mass assigned to a single outcome, so that $X$ takes a fixed value every time, the entropy of $X$ is zero, reflecting our total certainty of the outcome. Similarly, it can be shown that $H(X)$ is maximised when $P$ is a uniform probability distribution, so that every outcome in $\Omega$ is as likely as any other. Intuitively, it is impossible to be less certain of the outcome than we are in this situation. Note that the base of the logarithm in the definition of entropy is arbitrary. The most commonly used value is 2, and in this case entropy is measured in units of *bits*. All entropies and derived quantities in this thesis should be assumed to be in bits unless stated otherwise.

A common use of the idea of entropy in information theory is to consider the entropy of a source. Consider the simplest kind of source, in which each symbol

$A_i$ emitted is an independent sample from a fixed probability distribution. In this case the entropy of the source is simply $H(A_i)$. Source entropy represents how hard to predict the symbols which the source produces are on average. It is possible to define more complicated sources than the IID example considered here, for instance it is straightforward to define a source based on a Markov process. In these more complicated situations it is usually still not too difficult to define a source entropy which we can give the same intuitive interpretation.

Let $Y$ be a second discrete random variable with outcome set $\omega = \{y_1, \ldots, y_m\}$. The *mutual information* between $X$ and $Y$ is defined to be:

$$I(X;Y) = H(X) - H(X|Y),$$

where

$$H(X|Y) = \sum_{i=1}^{m} P(Y = y_i) H(X|Y = y_i)$$

and

$$H(X|Y = y_i) = \sum_{i=1}^{m} P(Y = y) \left( \sum_{j=1}^{n} -P(x_j|y_i) \log(P(x_j|y_i)) \right). \tag{8.2.1}$$

In words, the uncertainty of $X$ minus the uncertainty that remains in $X$ after the value of $Y$ is given. The mutual information between $X$ and $Y$ is a measure of the extent to which knowing the value of $Y$ reduces our uncertainty as to the variable $X$, on average, and the standard terminology is to say that $I(X;Y)$ is the *amount of information that $Y$ provides about $X$*. Note that if $X$ and $Y$ are statistically independent, then $H(X|Y) = H(X)$ and so $I(X;Y) = 0$, or in the standard terminology: statistically independent variables provide no information about one another, which is something we would intuitively expect to be true of a measure of information. A standard result in information theory is that mutual information is symmetric, i.e. $I(X;Y) = I(Y;X)$: one variable always tells us as much on average about another variable as that other variable tells us about the original variable on average.

Note that the mutual information between a variable and itself is straightforwardly equal to the entropy of that variable: $I(X;X) = H(X) - H(X|X) = H(X) - 0 = H(X)$. Thus we can interpret the entropy of a variable not only as a measure of *a priori* uncertainty as to the variable's value, but also as the amount of information required to resolve that uncertainty. Bearing this interpretation in mind, observe that the definition of $H(X)$ is in fact the expected value of the function:

$$I(x_i) = -\log(P(x_i)),$$

with the expectation taken over all values of $x$ in $X$'s outcome set. This function is called the *self-information* of the event $x_i$, or alternatively the *surprisal* of $x_i$. Similarly to $H(X)$, we can interpret $I(x_i)$ as either a measure of how uncertain we are that $x_i$ is the value of $X$ in a particular trial, or as a measure of how

much information we need to resolve that uncertain. The entropy of a random variable, then, is how uncertain on average we are about whether or not each possible outcome is the actual value. As the term surprisal suggests, $I(x_i)$ is also a measure of how surprised one should be to observe that $x_i$ is the actual value of $X$.

The final definition of interest relates to channels. The *capacity* of a channel is the maximum possible value of the mutual information between its input and its output, where the maximum is taken over all possible probability distributions for channel input. That is, it is the amount of information that the output of the channel provides us about the input of the channel, on average, when the input to the channel is distributed in the way which is most favourable to the channel in regard to this quantity.

These four concepts - entropy, mutual information, surprisal and channel capacity - are the essential information theoretic concepts which will be used throughout the rest of the thesis. Of course, in presenting them here I have only scratched the surface of information theory proper, and even with regard to those concepts which I have introduced, I have not discussed them in anything like their full depth. For more details the interested reader should consult a textbook on information theory, such as (Gallager, 1968).

### The noisy channel coding theorem

The most celebrated result in information theory is the so-called *noisy channel coding theorem*, which I shall now state loosely. Consider a source process with entropy $H$ and which outputs $t_S$ source symbols per second, as well as a noisy channel with capacity $C$ and which can transfer $t_C$ channel symbols per second. If $H/t_S < C/t_C$ then there exists a code for this source and channel such that the information from the source transmitted over the channel with arbitrarily low probability of error at the destination. The converse to this result is also true: if $H/t_S > C/t_C$ then no code exists which gives an error probability below some certain unavoidable probability (which varies as a function of $H/t_S$). This result was first stated by Shannon (1948), with an outline of a proof. The first rigorous proof is due to Feinstein (1954). The details of the proof are not important for my purposes here. To some extent, the result is intuitive, or at least its converse is. For simplicity suppose $t_S = t_C$ so that one channel symbols must be sent for each source channel in order for the channel to "keep up" with the source. $C$ is, by definition, the absolute maximum amount of average information about each input symbol which each output symbol can provide. $H$ is the average amount of information required to completely remove all uncertainty associated with each input symbol. If $H > C$ then on average there will be some non-zero uncertainty remaining about the input symbol after each output symbol is received, and as such there is some inevitable probability of an error.

### 8.2.2 Probabilistic language models

In the previous subsection I provided a brief introduction to information theory and stated that in this thesis I would be applying information theory to human language. In doing this we treat each successive word of an utterance as a random variable generated by some source process, but what exactly does this process look like? If we wish to actually answer a solid question like "what is the probability of hearing the word "jumped" as the next word in the string "the quick brown fox"?", how should we go about it?

Computing the probability of words in context requires a *probabilistic language model*, or PLM. A PLM is nothing more than an assignment of a probability to each of the sentences in a particular language. That is, for each sentence $S$ in a language, a PLM is a way of establishing $P(S)$. With this given it is easy to derive a probability distribution for individual words. The probability of hearing "jumped" after "the quick brown fox" is equal to the sum of the probabilities of all sentences beginning with "the quick brown fox jumped" divided by the sum of the probabilities of all sentences beginning with "the quick brown fox".

Individual PLMs are, of course, much more interesting than large lists of (string, probability) pairs, as a simple matter of practicality: languages are typically conceived of as being infinite, so that it is not possible to directly specify a probability for each string. Instead, PLMs consist of stochastic processes for generating sentences, and the probability of a sentence is simply the probability of its being generated by the process. Perhaps the simplest example of such a process is a Markov process, where each word is sampled from a probability distribution which depends upon some number of previous words. Markov language models in which each word is conditioned only on the previous word are typically referred to as bigram models, and models where the previous two words influence each subsequent word are referred to as trigram models. Using a bigram or trigram model, it is straightforward to compute quantities such as the entropy of each word in a sentence. More complicated PLMs are both possible and common, with standard examples being Hidden Markov Models and Probabilistic Context Free Grammars (PCFGs). A discussion of these language models is beyond my present scope as I do not use either of these classes of model in this thesis, but the interested reader is directed to (Bod, Hay, & Jannedy, 2003).

## 8.3 The UID Hypothesis

### 8.3.1 Definition

In this section I shall introduce an idea from the psycholingusitics literature called the Uniform Information Density (UID) hypothesis. Later I shall use the UID hypothesis as the basis of a new theory of word order functionality.

For now I will describe the theoretical motivations for the UID hypothesis as well as the empirical evidence to support the hypothesis. As far as I know, no particularly formal or rigorous definition of the UID hypothesis has been given in the literature, so I shall use the definition below, which is adequate for my purposess and which I believe should be compatible with all previous uses of the hypothesis.

> **The Uniform Information Density Hypothesis**: Let us consider conversation in a spoken language as a sequence of units $u_1, u_2, \ldots$, where the units may be phonemes, words, sentences, etc. Let $H(u_1), H(u_2), \ldots$ be the entropies of the successive units, where the probabilistic language model which defines the entropies is the implicit model used by the mind for the purposes of language comprehension, and the entropy of each unit is defined taking into account any relative context from preceding units. Let $t_1, t_2, \ldots$ be the time duration of each unit, i.e. $t_i$ is the time taken by the language producer to pronounce $u_i$. Then the spoken language is more functional the more uniform the amount of information conveyed per unit time, i.e. the closer $H(u_i)/t_i$ is to being constant. Language producers will unconsciously use a variety of means in an effort to bring their speech closer to this ideal (this is the "online" UID hypothesis), and languages themselves will change over time so as make this ideal easier to achieve (this is the "offline" UID hypothesis).

Strictly speaking, the hypothesis as stated above is different for each different choice of unit, as it is possible to keep information density constant at, say, the level of words while allowing it to vary at the level of phonemes (though not vice versa). As shall be clear once I have discussed the motivations for the hypothesis, the most appropriate scale at which to state the hypothesis is at whatever scale the human language faculty is sensitive to probabilities. However, for practical investigations and applications of the hypothesis, phonemes and words are the most commonly used timescales and there seems to be no great danger in this. The precise probabilistic language model or models underlying human language processing are, of course, unknown, so typically models such as $n$-grams trained on large corpora are used, the assumption being that the relevant psychological model will be well approximated by such models.

### 8.3.2 Theoretical motivations

The UID hypothesis is a *rational* hypothesis about language production in the sense of Anderson's program of rational analysis (Anderson, 1990): it hypothesizes that language use represents an optimal or near-optimal solution to an underlying problem specified at Marr's computational level (Marr, 1982). In

fact, communicating with UID is an optimal solution to *two* independent communication problems. The discussion of these problems belows closely follows Levy and Jaeger (2007).

### Noisy channel motivation

First, UID is an optimal solution to the problem of communicating efficiently (i.e., at high speed) and reliably (i.e., with minimal risk of error). This follows fairly straightforwardly from the discussion of the noisy channel coding theorem above. Recall that I earlier showed that spoken language can be fairly straightforwardly considered to be communication over a noisy channel as conceived by information theory, where the source of noise in the channel is a combination of articulation and perception problems and ambient noise. The noisy channel coding theorem states that error-free communication over such a channel is guaranteed to be possible if the source entropy is less than or equal to the channel capacity. If the source entropy exceeds the channel capacity, then some probability of error is unavoidable, and this probability of error increases the more the source entropy exceeds the channel capacity. As such, if we consider only the fact that spoken language should be a reliable communication medium, there is a functional pressure to keep the entropy of each unit of language as low as possible, so as to avoid exceeding the channel capacity. However, it is also desirable that communication should be time efficient, in the sense that as much information should be conveyed per unit time as possible. If we consider only this requirement, there is a functional pressure to keep the entropy of each unit of language as high as possible, so as not to waste time and energy verbalising units of language which listeners can reliably guess. These two functional pressures are in conflict with one another and the optimal trade-off is a language in which the entropy of each unit of language is just slightly less than the channel capacity. If any unit has a higher entropy than this, the probability of error is introduced. However, if any unit has a lower entropy than this, the time taken to pronounce that unit is not used as efficiently as possible, since more information could have been conveyed in that time without introducing a risk of error. As such, information theoretically optimal languages (and optimal codes in general) exhibit uniform information density which is as close as possible to the channel capacity.

### Processing effort motivation

The second independent perspective from which UID can be considered a property of rational languages is related to the information processing effort which a listener must expend in incrementally comprehending an utterance. Under certain assumptions about how sentences are processed, in particular about how ambiguity is resolved on-line, it can be shown that UID languages minimise total listener effort. My discussion here follows (Levy, 2005), which in turn builds on (Hale, 2001).

An elementary fact about human language comprehension is that it proceeds sequentially: people do not wait until the end of a sentence to begin the information processing work involved in comprehending it. Instead, each word of a sentence is processed, to the extent that it can be, immediately upon being heard. An elementary fact about incremental sentence comprehension is that not all parts of a sentence are equally easy to process in the context of those parts which have preceded it: some parts are easier to process and some are harder. In other words, human language comprehension displays *differential difficulty*. Accounting for these differences and explaining why certain sentence components are more or less easy to process is a central problem for psycholinguistics.

The traditional approach to this problem, dating back to (G. A. Miller & Chomsky, 1963) is the *resource limitation* theory: that some syntactic structures require more of some finite resource or set of finite resources (for instance, working memory) than other structures, and those structures which are more demanding with respect to this resource are correspondingly more difficult to process. At a finer grain, the individual words in those structures which contribute most to the high demand are the most difficult to process. These theories have developed hand in hand with theories about ambiguity resolution. It is hypothesised that the human parser is serial (i.e. pursues only a single candidate parse of a sentence at a time). If at some point during incremental comprehension of a sentence ambiguity should arise as to which of several syntactic structures compatible with the words heard so far is correct, a choice has to be made as to which of the possible parses should be pursued. In resource limitation theory, it has been proposed that the parser's strategy is to pursue the compatible option which minimises the consumption of the limited resources.

A less dominant school of thought on differential difficulty and ambiguity avoidance, the *resource allocation* theory, holds that the parser works in parallel: in the face of ambiguity, multiple compatible syntactic analyses are pursued simultaneously, and limited resources are shared across competing analyses in proportion to the parser's judgement of which analysis is the most likely to be correct (this judgement being made on a variety of grounds, e.g. syntactic, semantic, pragmatic, etc). In this account, difficulty in incremental comprehension comes from the effort involved in redistributing resources in the light of changes to the assessment of each parse's likelihood which occur after each word is processed. Levy (2005) develops a mathematically formalised version of this account, in which the distribution of resources across competing analyses is represented by a probability distribution over syntactic structures which are compatible with the sentence so far, and the difficulty of processing a word is proportional to the Kullback-Leibler divergence[2] between the distribution over

[2]The Kullback-Leibler divergence is a measure of the dissimilarity between two discrete probability distributions, whose definition is motivated by information theoretic concerns. The precise details of the KL divergence are not necessary for our purposes here.

structures before and after processing it. Levy shows that this formulation is in fact equivalent to one in which the difficulty of processing a word is in fact proportional to that word's surprisal, in the context in which it appears. Since entropy is nothing other than expected surprisal, it follows that the entropy of the random variable corresponding to an upcoming word is also the expected difficulty of processing that word. Levy assesses this surprisal based account of syntactic processing effort against available evidence on differential difficulty and find that it fares at least as well as and often better than the predictions derived from resource limitation theories.

From here, there are two paths to the conclusion that UID is a feature of rational languages. Levy suggests that it is natural to expect processing difficulty to be superlinear in surprisal, i.e. proportional to $(-\log(P(w_i|w_{-i}))^k$ for some $k > 1$, rather than just $-\log(P(w_i|w_{-i}))$. Given this assumption, if we define the total difficulty in processing a sentence to be the sum of the difficulties of processing each individual word (as is natural for incremental parsing), then it can be proven that total difficulty is minimised when all words of a sentence have equal surprisal (see (Levy & Jaeger, 2007) for the proof). Whether or not processing difficulty is genuinely superlinear in surprisal is an empirical matter. Alternatively, and regardless of how difficulty scales with surprisal, we can reason more or less as we did for the noisy channel motivation. In the same way that communicative reliability deteriorates if the source entropy exceeds the channel capacity, such that the adaptive thing to do is to avoid entropy exceeding this threshold, we can require that the difficulty of processing any individual word not exceed some certain threshold (say, the amount of processing effort which the average human can, under typical conditions, bring to bear on a problem in the timespan of an average word) in order to avoid problems with processing. If we strive to keep difficulty under this amount, but keep information as high as possible below this bound for the sake of efficiency, uniform information density is a consequence.

### 8.3.3 Empirical evidence

In addition to these theoretical justifications, there is extensive empirical evidence for the UID hypothesis. Because the hypothesis had its origins in the speech research community, much of the early support for the idea originates from phonetics, where the empirical phenomena were identified before any theoretical explanations were offered. Aylett (1999) found that words which are very frequently uttered, or which are easy to predict from context (i.e., have low entropy and convey little information) tend to be spoken less clearly and more quickly than words which are infrequent or hard to predict (i.e., have high entropy and convey more information). Shortening words with low information content and elongating those with high information content acts to bring the information density of a sentence closer to uniformity. Bell et al. (2003) found similar results for function words. The first work in the speech research

community to explain these findings in terms of of an information theoretically optimal strategy for communication is Aylett and Turk (2004), which postulates that the purpose of the link between a word's predictability and spoken duration is to facilitate reliable communication over a noisy channel. They term this the *smooth signal redundancy hypothesis*.

The first application of the UID idea outside of speech and phonetics is (Genzel & Charniak, 2002), which provides evidence of UID (which they call *entropy rate constancy*) in corpora of written text, which are analysed using $n$-gram models. Evidence is presented to suggest that UID is achieved through both lexical means (i.e., choice of which words are used) and non-lexical means (i.e., how the words are used), although specific non-lexical means are not identified. (Genzel & Charniak, 2003) reproduces the results of (Genzel & Charniak, 2002) in more generality, including using corpora from several genres of writing and in Russian and Spanish in addition to English. Keller (2004) corrected a methodological error in these previous two papers but still found support for UID, and presented additional support on the basis of eye-tracking data.

The first evidence of a specific tactic for achieving UID via non-phonetic or lexical means (and the origin of the term "uniform information density") was presented in (Jaeger, 2006) and (Levy & Jaeger, 2007). These works suggest that speakers of English make strategic use of the optionality of the word "that" at the beginning of relative clauses in order to achieve UID. For example, presented with a choice between sentences like "how many people are in the family you cook for?" and "how many people are in the family **that** you cook for?", speakers will use the optional "that" if the relative clause has high information content. More evidence for this effect is presented in (Jaeger, 2010). (Frank & Jaeger, 2008) shows that speakers make use of optional contractions, such as "isn't" from "is not", in a way consistent with UID as well.

All of the evidence cited for the UID hypothesis so far has related to choices which people make during language production, i.e. they are online effects. There is also some evidence of offline UID effects on language, i.e. of languages themselves changing so as to facilitate UID. (Piantadosi, Tily, & Gibson, 2009) and (Piantadosi, Tily, & Gibson, 2011) present and finds evidence for the *communicative lexicon hypothesis* (CLH): that "human lexical systems are efficient solutions to the problem of communication for the human language processor". It is shown that the length of a word is better predicted by that word's average predictability in context than by its frequency of occurence. Words which are more predictable in context (i.e. which convey little information) tend to be shorter than words which are less predictable, in complete consistenty with the UID hypothesis.

## 8.4 Linking word order and information density

In the previous section I reviewed a number of ways in which languages appear to have changed in response to the functional demands of unifomation information density. The changes so far have been relatively straightforward, in that they manipulate information density quite directly by either slowing down or speeding up the pronounciation of syllables or words, inserting or ommitting optional words, or decreasing or increasing the length of a word. In this section I shall present the hypothesis that the information density of an utterance also be manipulated in a less direct manner by changing the order of the words in the utterance. This hypothesis leads fairly directly to the idea that some basic word orders are more functional than others. In this section I shall develop my hypothesis in a fairly informal, intuitive manner. In the next section I shall construct a formal mathematical model which will facilitate emprical testing of the hypothesis in the subsequent chapter.

The idea that information density is influenced by word order relies entirely upon the semantic contribution to the entropy of language, rather than the syntactic contribution. To cleanly illustrate this, consider the following two sentence beginnings:

(22) The man bought the...

(23) The boy smashed the...

Both of these sentences have the same (partial) syntactic structure, so that no word can be considered more likely to follow "the" than any other for syntactic reasons. However, the same is not true semantically. 22 can be completed by any words such that the final noun phrase refers to something which can be purchased by a man: "book", "vase" and "cushion" are all acceptable possibilities. 23, on the other hand, can only be completed by words such that the final noun phrase refers to something which can be smashed by a boy. While "cushion" is a pefectly acceptable completion of 22, this is not the case for 23, by virtue of the fact that cushions, owing to their essential property of being soft and malleable, cannot be smashed. If we accept the intuitive principle that everything which can be smashed can be bought but not everything which can be bought can be smashed, then we must conclude that a rational language user's uncertainty regarding the completion of 23 is less than regarding the completion of 22. This is an example of a purely syntactic contribution to uncertainty, and this effect would not be captured by a purely semantic probabilistic model of language such as a typical PCFG.

Having established how semantic considerations can influence the uncertainty of a word given some mid-sentential context, it is relatively straightforward to see how the order of words can have an effect. When the words of an utterance are arranged in a different order, then each word other than the first word of the utterance must be anticipated by the listener in a different

context. Consider the sentence "pizza is my favourite thing to eat", and the reordering "my favourite thing to eat is pizza". At some point in hearing these two sentences, the listener is presented with the following sentence beginnings:

(24) Pizza is my favourite thing to...

(25) My favourite thing to eat is...

The uncertainty of the completion of 24 is intuitively lower than that of the completion of 25: there are relatively few things one is likely to do to pizza, but a great many things that one is likely to eat. This situation demonstrates that the distribution of information throughout an utterance can be changed by doing nothing other than reordering the constituents of the sentence, due to entirely semantic conributions to entropy.

The considerations above lead fairly directly to the possibility that some of the six logically possible basic word orders may systematically lead to more uniformation distributions of information throughout utterances than others, so that the UID hypothesis implies a functional ranking of word orders. This motivates the statement of the following hypothesis:

> **The UID word order functionality hypothesis**: all else being equal, some basic word orders are more functional than others by virtue of the fact that they lead to, on average, more uniform distribution of information throughout an utterance.

I have established that this hypothesis is possible in principle. Whether it is true or not is an empirical question and in order to test the hypothesis it is necessary to formulate it more precisely. I shall do so in the next section.

Before continuing, I wish to point out that I consider this account of word order functionality to be a true example of what the traditional school of functionailsm thought functionality should be. Under the noisy channel motivation for the UID hypothesis, there is nothing in the logic leading to the conclusion that languages ought to distribute information uniformly which is specific to the implementation of language in the human mind. UID is a desirable quality not only for human language but for any system of serial communication, including many artificial communication systems in telecommunications which bear very little similarity to human language. That is to say, it is a functional requirement which applies to the abstract task of communication, by virtue of the nature of that task (so long as we take error resistance and time efficiency to be requirements of good communication). Note, however, that the processing effort motivation for the UID hypothesis is certainly human specific, making it an instance of cognitive functionalism.

## 8.5 Mathematical formalism

### 8.5.1 Basic model structure

The first component of the mathematical model I shall construct to test the UID word order functionality hypothesis is a model *world*. There are two kinds of entities in the model world: *objects*, which are members of a set $\mathcal{O}$, and *actions*, which are members of a set $\mathcal{A}$. An *event* in the world is a triple consisting of two objects and an action, and shall be denoted $e = (o_1, o_2, a)$, where $o_1, o_2 \in \mathcal{O}$ and $a \in \mathcal{A}$. The convention shall be that $o_1$ is the agent of the action, and $o_2$ is the patient. By way of example, the event which might be described by an English speaker as "the dog bit the man" is (Dog, Man, Bite). The set of all events defined in the world, or the "event-space", will be denoted $\mathcal{E}$, and is simply the Cartesian product $\mathcal{O} \times \mathcal{A} \times \mathcal{O}$.

This model world is inhabited by *agents*, who use languages to describe events which occcur in the world. The agents produce *utterances* which consist of three words, drawn from a vocabulary $\mathcal{V}$. Utterances shall be denoted $u = (w_1, w_2, w_3)$, where $w_i \in \mathcal{V}$ and the convention shall be that the words of the utterance are uttered in left to right order. Two of the words in each utterance refer to the objects of an event (i.e. the agent and patient), and one of the words refers to the action. By way of example, the English sentence "the dog bit the man" corresponds to the utterance ("man", "bite", "dog"). The set of all utterances defined in the language, or the "utterance-space", will be denoted $\mathcal{U}$.

Agents use one utterance to describe one event, without ambiguity. The mapping from events to utterances is as follows. Each object and each action in the model world corresponds to precisely one word in the vocabulary, i.e. $|\mathcal{O} \cup \mathcal{A}| = |\mathcal{V}|$. In this way, each event $e$ corresponds uniquely to an unordered set of words $\{w_{ag}, w_{pa}, w_{ac}\}$. In order to derive an utterance from this set, the agent must decide upon an ordering of these three words: a word order. A language for the agents is defined completely by the vocabulary $\mathcal{V}$ and a word order parameter $\theta$. To describe an event an agent utters the three words in the appropriate order.

I now describe the manner in which agents select events to talk about. The model world is equipped with a probability distribution $\Phi$, defined over the event-space $\mathcal{E}$: each possible event is associated with a probability. The distribution $\Phi$ decsribes the likelihood of the different events being described by an agent. Some events will receive higher probabilities than others, and some events may have a probability of zero. This is intended as a reflection of the fact that, in the real world, some events are more probably (and hence more probably discussed) than others (for instance, (Dog, Man, Bite) probably happens somewhere every day, whereas (Man, Dog, Bite) is much more rare), and some events are in fact impossible (such as Cat, Car, Drink)). The distribution $\Phi$ is a property of the world and is known by all agents. A

| | Apple | Bread | Cake | Rice | Coffee | Cola | Juice | Water |
|---|---|---|---|---|---|---|---|---|
| Alice | 0.05 | 0.00 | 0.03 | 0.02 | 0.07 | 0.03 | 0.00 | 0.00 |
| Bob | 0.02 | 0.00 | 0.04 | 0.04 | 0.02 | 0.04 | 0.02 | 0.02 |
| Eve | 0.00 | 0.01 | 0.00 | 0.9 | 0.03 | 0.01 | 0.00 | 0.06 |
| Mallory | 0.04 | 0.04 | 0.01 | 0.01 | 0.00 | 0.01 | 0.9 | 0.00 |
| Trent | 0.02 | 0.00 | 0.01 | 0.07 | 0.02 | 0.03 | 0.03 | 0.02 |
| | | | EAT | | | | DRINK | |

Figure 8.5.1: A diagrammatic representation of the toy world event distribution.

conversation between two agents consists of agents making a series of statistically independent samples of events from $\Phi$ and verbalising these events with their languge's word order.

To make this idea more concrete, I shall construct a simple toy world consisting of thirteen objects and two actions, and carry this world through the rest of the section to use in examples. Five of the objects in this world represent individual people (Alice, Bob, Eve, Mallory, Trent) and the other eight represent items which are either food (Apple, Bread, Cake, Rice) or drink (Coffee, Cola, Juice, Water). The two actions are Eat and Drink. The events in this world represent particualar people eating or drinking particular items. Some of the logically possible events in this world are either impossible or incoherrent, such as (Coffee, Cake, Eat) or (Alice, Apple, Drink). These impossible events are given a probability of zero in the event distribution $\Phi$. The sensible events are given non-zero probabilities, in such a way as to ensure the following features: each of the five people are equally likely to be the actor of an event; the two actions are equally likely, regardless of the actor; each person has their own particular idiosyncratic distribution over which foods they prefer to eat and which drinks they prefer to drink. A diagrammatic representation of all the non-zero probabilities of $\Phi$ is shown in Figure 8.5.1.

At this point, I have described a complete probabilistic model in which agents converse with one another in a principled away about the events which occur in the world they inhabit. The specification is sufficiently detailed that conversations can be readily simulated on a computer. I now turn my attention to the information theoretic aspect of the model.

### 8.5.2 Information theoretic analysis

From the perspective of the receiving agent in a conversation, i.e. the agent which is not producing the utterance, each word that the transmitting agent utters can be thought of as a random variable, in so far as the receiving agent does not, in general, know in advance what each word will be. Thus, each word has an entropy associated with it. For every utterance $(w_1, w_2, w_3)$, there is a corresponding sequence of entropies, $(H_1, H_2, H_3)$. I shall call this sequence the *entropy profile* of the utterance.

For a fixed word order $\theta$, the first term in the entropy profile is always the same. It is equal to the entropy of the marginal distribution over either agents, patients or actions derived from $\Phi$. For example, if $\theta = \text{SOV}$, then the first word of each utterance describes the agent, and so, if $w(o_i)$ is the word used to describe object $o_i$ then $H_1$ is the entropy of the distribution:

$$P(w_1 = w(o_i)) = \sum_{(o_1,o_2,a)\in\mathcal{E}} \frac{\Phi(o_1, o_2, a)\delta(o_1, o_i)}{Z}, \tag{8.5.1}$$

where $\delta(\cdot, \cdot)$ is equal to 1 if its arguments are equal and 0 otherwise, and $Z$ is the appropriate normalising constant. Note that this is simply the distribution $\Phi$ marginalised to give a probability distribution over agents.

The entropy of the second word of an SOV utterance is the entropy of the distribution derived from $\Phi$ by first conditioning on the agent, which is known due to the first word of the utterance, and then marginalising over patients, i.e. if $o(w_i)$ is the object described by $w_i$ then $H_2$ is the entropy of the distribution:

$$P(w_2 = w(o_i)|w_1) = \sum_{(o_1,o_2,a)\in\mathcal{E}} \frac{\Phi(o_1, o_2, a)\delta(o_1, o(w_1))\delta(o_2, o_i)}{Z'}. \tag{8.5.2}$$

The entropy of the third word is, as should by now be clear, the entropy of the marginal distribution over actions, conditioned on the agent and patient, which are both now known due to the first and second words of the utterance.

$$P(w_3 = w(a_i)|w_1, w_2) = \sum_{(o_1,o_2,a)\in\mathcal{E}} \frac{\Phi(o_1, o_2, a)\delta(o_1, o(w_1))\delta(o_2, o(w_2))\delta(a, a_i)}{Z''}. \tag{8.5.3}$$

It should be intuitively obvious that if we restrict our attention to a single event $e$ and consider the six different choices of word order which can used to describe it, each word order can in principle lead to a distinct entropy profile, because we are conditioning and marginalising over $\Phi$ in different ways each time. Since each of the word orders gives rise to a different entropy profile, the UID hypothesis suggests that each event in one of our model worlds has a corresponding "best" and "worst" word order.

### 8.5.3 Measuring UID functionality

To formalize the notion of each event having a best and worst word order, I introduce the concept of a *UID deviation score* $D(H)$ for any given information profile $H$. Let:

$$D(H) = \sum_{i=1}^{3} |H_i - \frac{H_1 + H_2 + H_3}{3}|. \tag{8.5.4}$$

It is easily verified that the UID ideal information profile, with $H_1 = H_2 = H_3$, has a deviation score of zero. The function $D(\cdot)$ provides an objective measure of how close a given information profile is to the UID ideal. The best word order for an event is that which minimises the value of $D(\cdot)$ and the worst is that which maximises it.

The deviation score allows us to measure UID word order functionality on a per-event basis. The question now arises of how this per-utterance score can be transformed into a global measure of UID functionality of a given word order for the entire model world. It is not clear that there is a single correct way to do this. There are infinitely many ways to derive a global measure, and the correct one to use for empirically testing the UID word order functionality hypothesis is determined by the details of precisely how it is that languges change their word order in response to the functional pressure for UID. Since it is not clear precisely what this global measure looks like, I shall resort to considering just two reasonable seeming candidates. If investigation using these global measures suggests that the UID word order functionality hypothesis is promising, then the issue of how to derive a more principled global measure can be dealt with. If the hypothesis does not fare well against the data with these candidates, however, it is difficult to imagine that it would do substantially better with whatever the true global measure turned out to be.

The first global measure of UID word order functionality I wish to propose is simply the probability that a given word order will yield the best possible UID deviation score for a randomly selected event in a conversations' event stream. That is, for a given $\theta$, if we randomly sample an event from $\mathcal{E}$ according to $\Phi$, what is the probability that the UID deviation score of the entropy profile associated with verbalising that event with word order $\theta$ is equal to or less than the UID deviation score associated with verbalising that event with any of the five other word orders? It seems intuitive that the higher this probability, the more functional the word order is for the given model world.

The second measures is simply the mean value of the distribution over deviation scores that each word order induces, i.e. for each word order $\theta$ the distribution $Q_\theta$ given by:

$$Q_\theta(d) = \sum_{e \in \mathcal{E}} \Phi(e)\delta(D(H_{e,\theta}), d). \tag{8.5.5}$$

It once again seems intuitive that the higher this probability, the more functional the word order is for the given model world.

Although it is perhaps obvious, I want to emphasise the following: if the event distribution $\Phi$ were uniform, then every pairing of an event and a word order would produce precisely the same information profile. In such a situation, the UID hypothesis would not prefer any basic word order over any other and it would be impossible to derive an account of word order functionality from the UID hypothesis. Of course, this is not the case in reality: the event distribution of the real world is highly *non*-uniform and so the UID hypothesis *does* impose a functionality ranking on word orders. The particular ranking imposed is determined by the structure of the non-uniform event distribution. I say all of this to emphasise the fact that, according to my hypothesis, the distribution of basic word orders is a reflection of the *relational structure of the physical world* that languages are used to describe. In a nutshell, the basic claim is that "languages are the way they are because the world is the way it is". This line of thinking is in no way novel. For instance, there is evidence that colour words (Regier, Kay, & Khetarpal, 2007) and spatial terms (Khetarpal, Majid, & Regier, 2009), despite varying significantly cross-linguistically, represent near-optimal partitions of universal spaces arising from our perception of the physical environment.

## 8.6 Summary

The essence of my offered account of word order functionality is hopefully now clear. I have illustrated that word order has implications for information density and developed a simple mathematical framework for quantifying this effect: given a simplified representation of the relational structure of the world, in the form of an event distribution $\Phi$, UID deviation scores can be used to produce a ranking of the six possible word orders by functionality. The obvious question to ask at this point is what sort of word order rankings emerge when the event distribution $\Phi$ more accurately approximates reality, and do they satisfy the requirements established in Chapter 3? I shall pursue precisely these questions in the next chapter.

# Chapter 9

# Estimating UID functionality from corpora and an experiment

In this chapter I take up the task of instantiating model worlds of the form described in the previous chapter in such a way that their event distributions $\Phi$ more realistically resemble the structure of the event distribution in the real world. For each of the worlds $(\mathcal{O}, \mathcal{A}, \Phi)$ which I instantiate in this chapter, I shall evaluate the two global measures of UID functionality I proposed in the previous chapter, as well as present various graphical representations of how the different word orders distribute information differently.

## 9.1 Corpus analysis

In this section I estimate $P$ on the basis of a variety of corpora, covering a a range of languages as well as different kinds of language such as spoken and written, child-directed and adult-directed. The languages used are English and in one case Japanese. I have chosen these languages primarily on the basis of familiarity, but a supporting motivation is the fact that they have distinct basic word orders. English and Japanese are uncontroversially considered to have basic word orders of SVO and SOV, respectively, and these are by far the most commonly used word orders for these languages. Using corpora with different patterns of word order use is valuable because it helps guard against the possibility of spurious results if it turns out that language users have a tendency to preferentially produce those utterances which are UID functional for their language's basic word order.

I estimate $\Phi$ from a corpus a fairly straightforward way. Most of the corpora I shall use come in the form of *treebanks*, that is collections of sentences which have been syntactically parsed into tree-structures (and also part-of-speech tagged). I use the `tregex` tool[1] to extract from these treebanks sets of triples,

[1] `tregex` is described in (Levy & Andrew, 2006) and is available from `http://nlp.stanford.edu/software/tregex.shtml`

each consisting of two nouns and one verb which are part of the same transitive verb phrase. Thus I am left with a simplified corpus of (subject, object, verb) triples, where triples may occur more than once. For corpora which are not in treebank format, I have manually produced corpora of (subject, object, verb) triples of the same kind (these corpora are of course much smaller due to the effort involved in manually extracting these triples). The triples are then transformed so that plurality, tense, case, etc. are not distinguished. For English treebanks, this is done using the Porter stemming algorithm (Porter, 1980). For non-treebank corpora this was done by hand. To summarise, the overall process is such that the utterances "the black cat chased the mouse", "the cats are chasing the mice" and "the cat chased the mouse around and around the new sofa" would all be mapped to the same triple (cat, mouse, chase), which I shall identify with the event (CAT, MOUSE, CHASE). Thus, at the end of this process I am left with a set of "event corpora", with each event potentially occuring more than once per corpus. Note that utterances involving the verbs "is", "are", "am", "was", etc. and the pronouns "that", "this", "it" were not included in event corpora. "Is" and related verbs are not actions in the sense that my formalism attempts to model. Impersonal pronouns can refer to wildly differening objects so that it does not make sense to track the statistics of their involvement with actions. Personal pronouns such as "I" and "you" were included, as these always refer to humans who are broadly similar in the sorts of actions they are agents and patients of.

For any given event corpus, the set of all agents and patients in the corpus define the sets $\mathcal{O}$ and $\mathcal{A}$, which determine the eventspace of the model world. I make the assumption that each event corpus represents a set of independent, identically distributed samples from an event distribution $\Phi$ and take a maximum likelihood estimate of $\Phi$, i.e. I take:

$$\hat{\Phi}(e) = \frac{n(e)}{n}, \tag{9.1.1}$$

where $n(e)$ is the number of times event $e$ occurs in the event corpus and $n$ is the total size of the corpus.

With the estimate $\hat{\Phi}$ computed, it is straightforward to compute distributions of deviation scores for particular combinations of model worlds are word orders, in both the unstructured and discourse modes of communication. I present the results of these computations in this section for a variety of corpora. Note that I shall leave a discussion of why it is that particular word orders are shown to be more or less UID functional until the next Chapter. Note that entropies in this chapter are in units of "nats" (i.e. using natural logarithms) for ease of programming.

### 9.1.1 English CHILDES data

The first corpus I shall consider is an English corpus of child-directed speech, in particular a subset of the Adam transcripts from the Brown corpus (Brown,

1973) in the CHILDES database (MacWhinney, 2000). These are transcripts of spontaneous speech between a young (27 months at the beginning of the study) boy "Adam" and his parents (and occasionally the researchers producing the transcripts), recorded in the child's home in fortnightly sessions of around 2 hours extending over about 2 years. The particular subset used is that used in (Perfors, Tenenbaum, & Regier, 2006), as that subset was readily available to me in a processed form which made building an event corpus more straightforward than working with the entire corpus in raw form. I am not aware of anything in the selection criteria for this subset which should bias the results if it is used for my purposes here. Note that only speech by the parents or investigators directed at the children was used, not the speech of the children. The total size of the event corpus derived is 544 events, which is quite small. The work using this corpus was the first empirical test of the UID word order functionality hypothesis which I performed, and the small corpus size was considered acceptable due to the preliminary nature of the investigation and the effort involved in manually constructing an event corpus.

I have chosen to present two different graphical summaries of the results of analysing this event corpus. Figure 9.1.1 shows the deviation score distribution $Q$ for each of the six different word orders, while Figure 9.1.2 shows box and whisker plots which indicate how the entropies of the first, second and third words are distributed for each of the word orders. It is fairly straightforward to get a feel for the relative functionality of the word orders by looking at the deviation score plots, as these tend to have quite pronounced modes. However, these plots provide no insight into *why* a particularly disfunctional word order has the high deviation scores it does. The box and whisker plots make up for this shortcoming, by allowing us to see which words typically have high or low entropy.

Table 9.1.1 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get SVO > SOV > VSO > VOS > OVS = OSV. If we rank the word orders by their mean deviation scores, we get SVO > VSO > VOS > SOV > OVS > OSV. Examining the plots in Figure 9.1.2, we can see that the first of these rankings has much of its structure (S-initial > V-initial > O-initial) determined by the entropy of the first word in each utterance. For SOV and SVO this entropy is 1.584 bits, for VSO and VOS it is 2.808 bits and for OVS and OSV it is 4.955 bits. Clearly one of the important defining factors in UID functionality is minimising the entropy of the first word in each utterance.

### 9.1.2 Japanese CHILDES data

The second corpus I consider is a Japanese corpus of child-directed speech, in particular "Asato", "Nanami" and "Tomito" transcripts in the MiiPro corpus in the CHILDES database. This corpus was, like the English CHILDES data

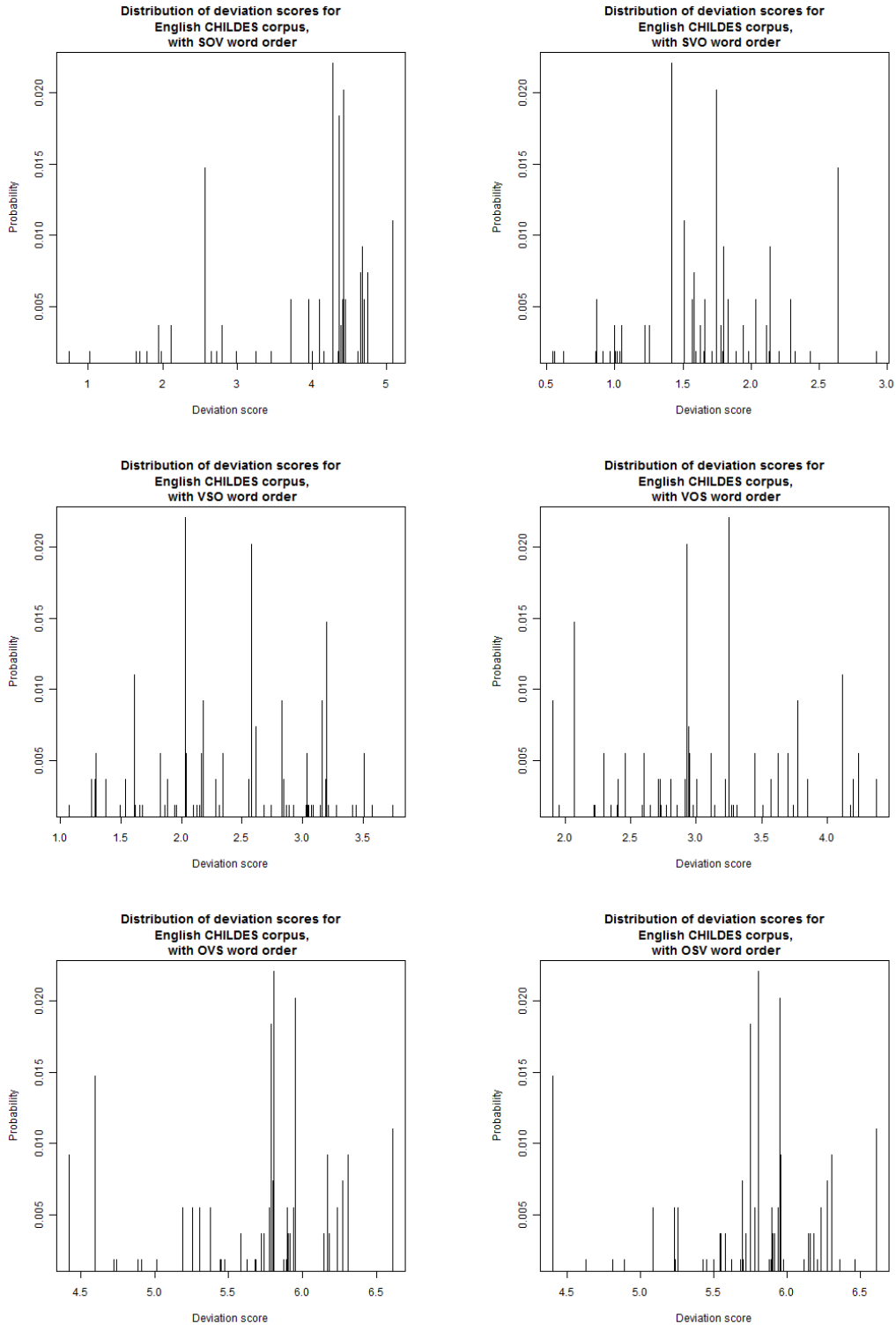


Figure 9.1.1: Distribution of deviation scores for all six basic word orders in the model world instantiated from the English CHILDES corpus.

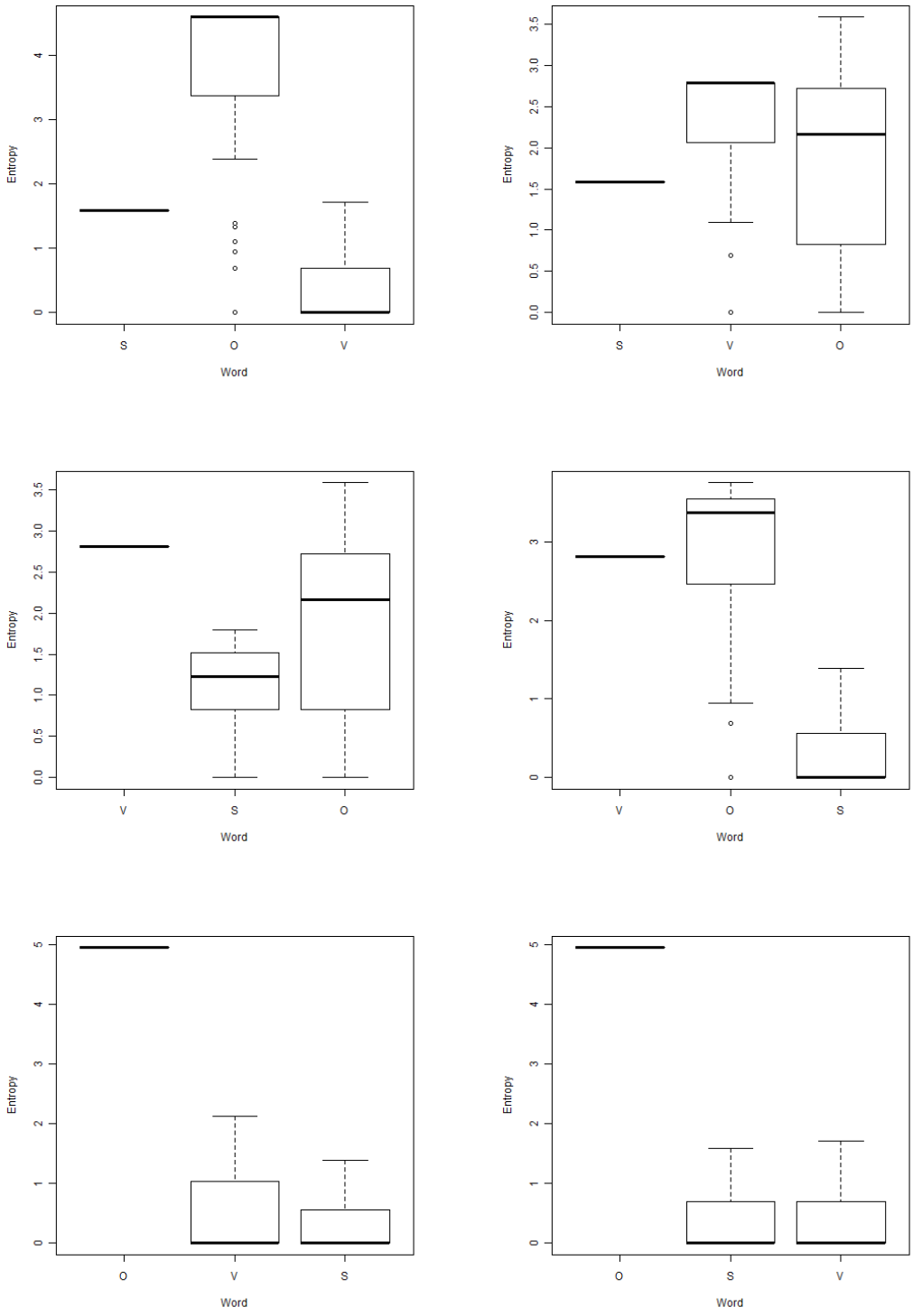


Figure 9.1.2: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the English CHILDES corpus.

Table 9.1.1: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the English CHILDES corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.103 | 4.003 |
| SVO | 0.888 | 1.648 |
| VSO | 0.055 | 2.274 |
| VOS | 0.037 | 3.478 |
| OVS | 0.000 | 6.038 |
| OSV | 0.000 | 6.038 |

Table 9.1.2: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the Japanese CHILDES corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.358 | 3.464 |
| SVO | 0.813 | 2.955 |
| VSO | 0.000 | 4.494 |
| VOS | 0.000 | 4.507 |
| OVS | 0.000 | 5.236 |
| OSV | 0.000 | 5.236 |

discussed above, used for a quick preliminary test of the hypothesis, and so it too is quite small. In this case the size of the event corpus is even smaller, at only 134 utterances. This is due primarily to two factors. Firstly, the extra difficulty involved in comprehending a non-native language, and secondly the fact that subjects are often "dropped" (i.e. omitted) in Japanese where they can be reliably inferred from context. This is a feature of the language and is considered entirely grammatical. This naturally means that a corpus of Japanese language of some given size will typically yield fewer (subject, object, verb) triples than, say, an English corpus of comprable size.

Figures 9.1.3 and 9.1.4 show the same kind of plots that were shown for the English CHILDES corpus.

Table 9.1.2 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get SVO > SOV > VSO = VOS = OVS = OSV. If we rank the word orders by their mean deviation scores, we get SVO > SOV > VSO > VOS > OVS > OSV. Again, the box and whisker plots enable us to see that the ranking is largely influenced by the entropy of the first word in each utterance. For this model world these entropies are 1.310 bits for SOV and SVO, 3.824 bits for VSO and VOS, and 4.222 bits for OVS and OSV.

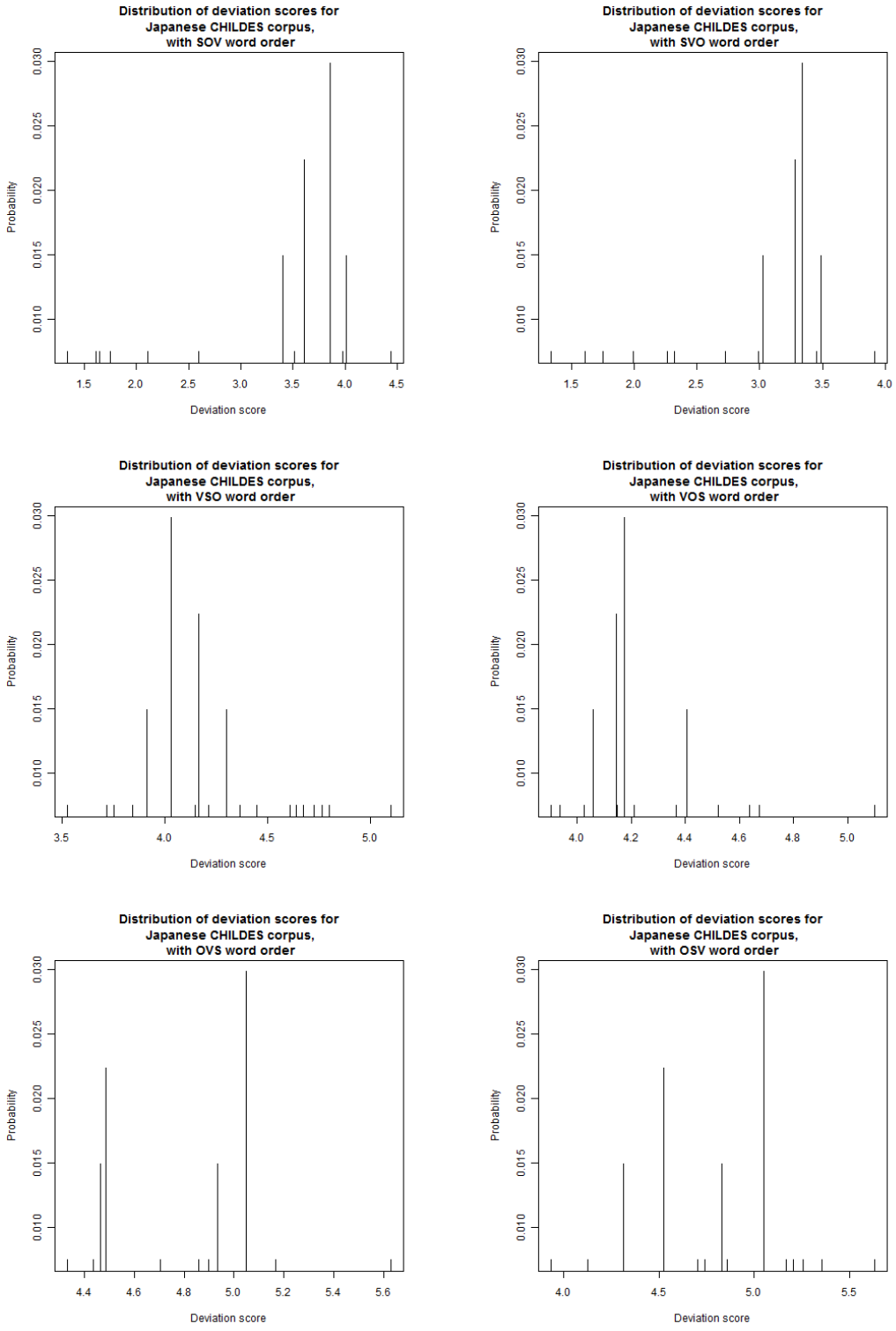


Figure 9.1.3: Distribution of deviation scores for all six basic word orders in the model world instantiated from the Japanese CHILDES corpus.

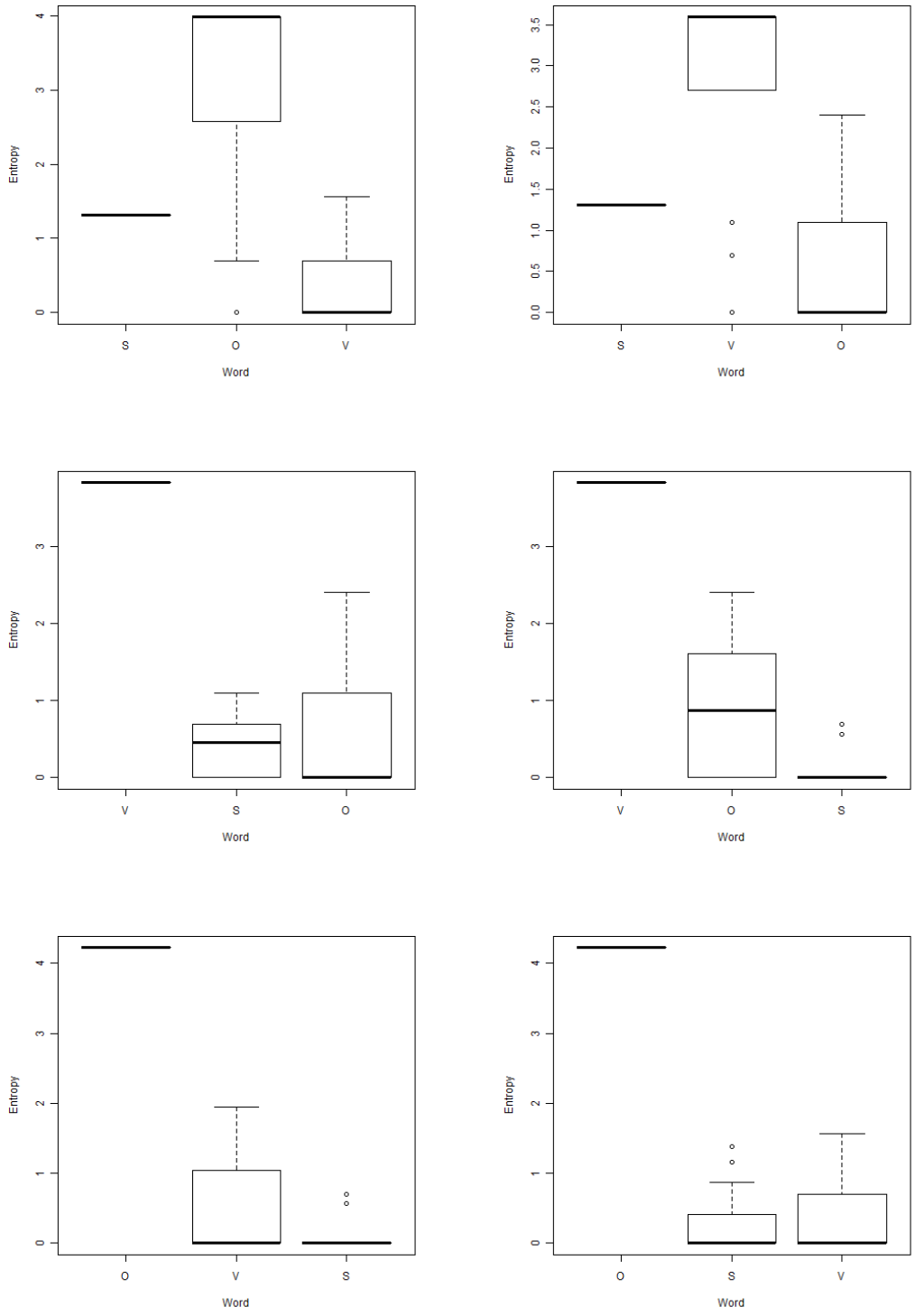


Figure 9.1.4: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the Japanese CHILDES corpus.

Table 9.1.3: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the Brown corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.005 | 7.192 |
| SVO | 0.001 | 6.719 |
| VSO | 0.787 | 4.346 |
| VOS | 0.530 | 5.593 |
| OVS | 0.000 | 8.039 |
| OSV | 0.000 | 8.074 |

### 9.1.3 Brown corpus

The first serious test of the UID word order functionality test I shall present is based on the subset of the Brown University Standard Corpus of Present-Day American English (commonly known as "the Brown corpus", not to be confused with the corpus produced by Roger Brown referred to above) as included in the Penn English Treebank. The corpus consists of around 24,000 sentences, taken from a variety of sources of written English, clasified into categories such as "popular lore", "adventure and western fiction" and "humor". Using `tregex` I derive an event corpus of size 2316.

Figures 9.1.5 and 9.1.6 show the now standard plots for this model world.

Table 9.1.3 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get VSO > VOS > SOV > SVO > OVS = OSV. If we rank the word orders by their mean deviation scores, we get VSO > VOS > SVO > SOV > OVS > OSV. As per the two CHILDES corpora, the box and whisker plots suggest that the ranking is shaped significantly by first word entropy. For this model world the entropies are 6.500 bits for SOV and SVO, 6.364 bits for VSO and VOS, and 5.604 bits for OVS and OSV. Note that in this case initial verbs are less uncertain than initial subjects, in contrast to earlier cases, and that there is less difference between word orders than previously.

### 9.1.4 Switchboard corpus

The next large test is based on the subset of the Switchboard corpus as included in the Penn English Treebank. This is a very large corpus, consisting of spoken English. Using `tregex` I derive an event corpus of size 926.

Figures 9.1.7 and 9.1.8 show the now standard plots for this model world.

Table 9.1.4 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get VSO > VOS > SVO > SOV >

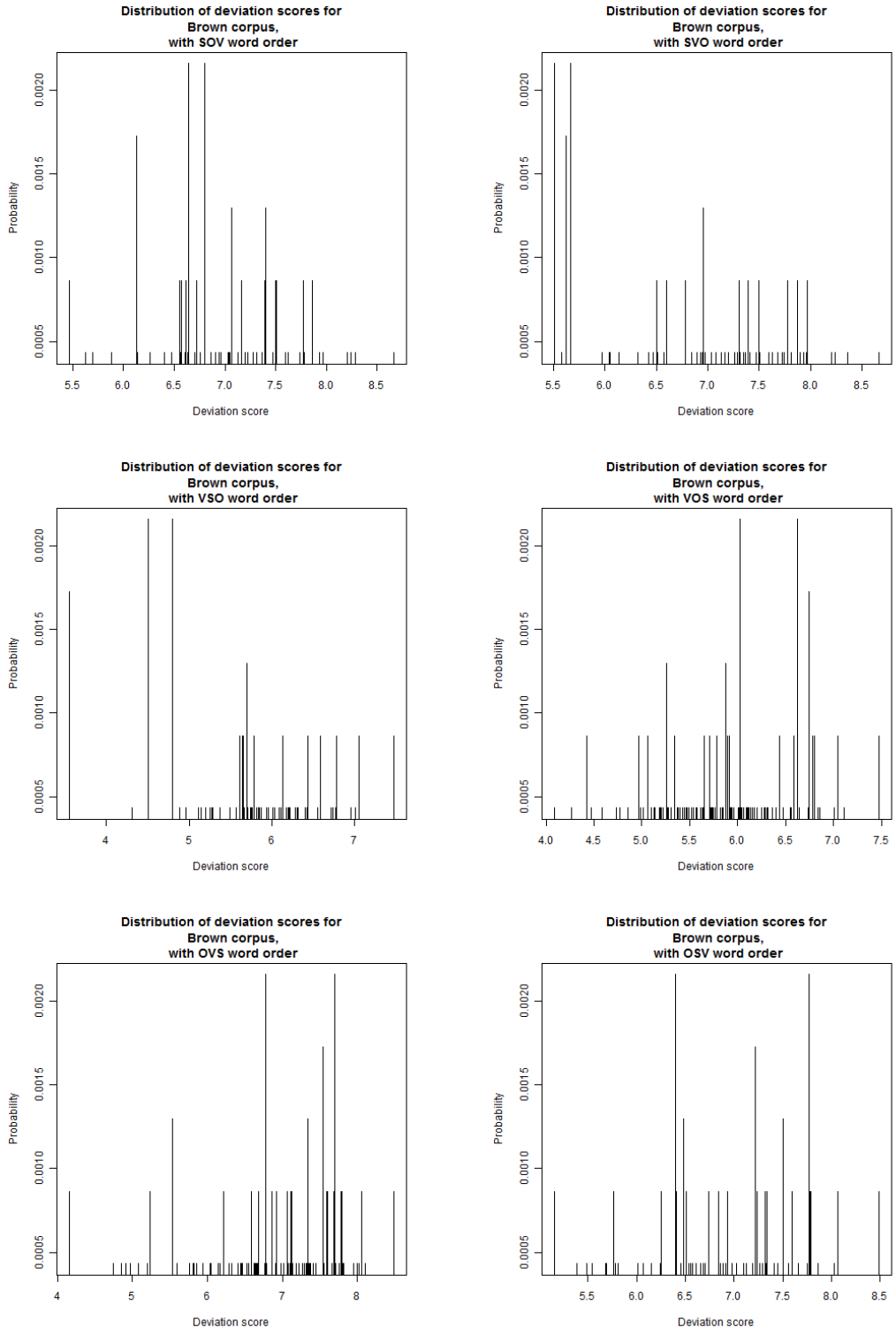


Figure 9.1.5: Distribution of deviation scores for all six basic word orders in the model world instantiated from the Brown corpus.

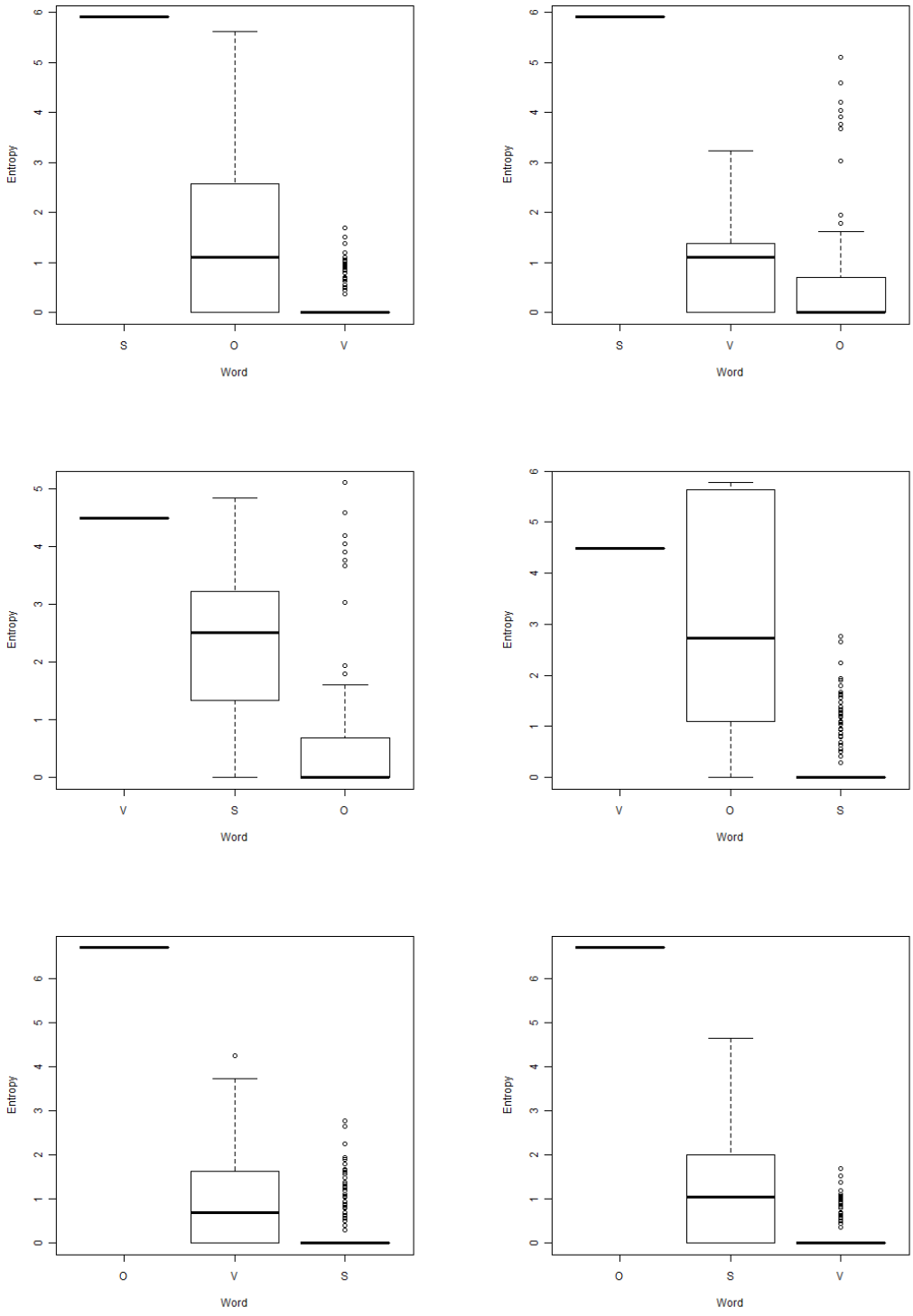


Figure 9.1.6: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the Brown corpus.

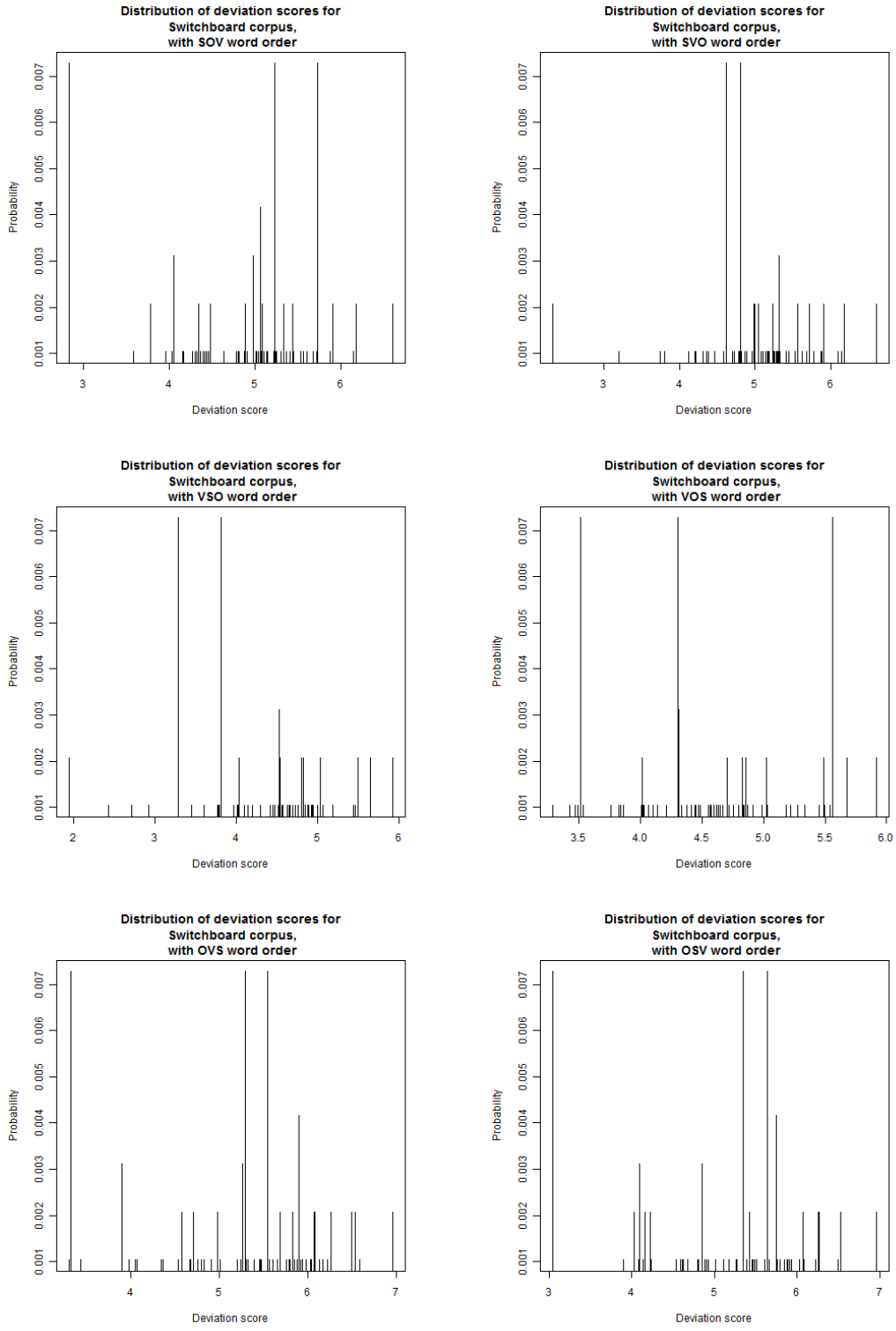


Figure 9.1.7: Distribution of deviation scores for all six basic word orders in the model world instantiated from the Switchboard corpus.

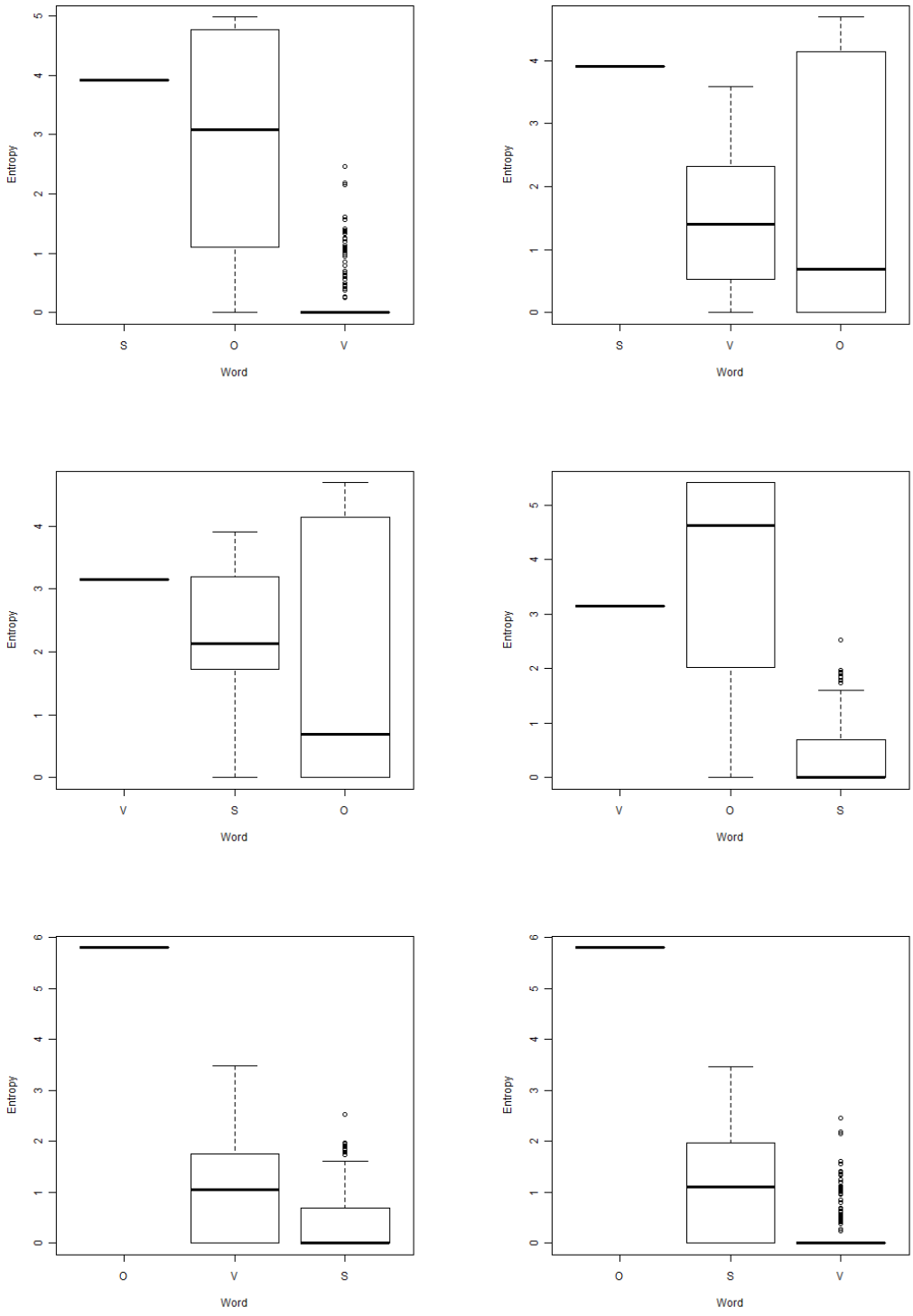


Figure 9.1.8: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the Switchboard corpus.

Table 9.1.4: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the Switchboard corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.045 | 4.861 |
| SVO | 0.129 | 3.902 |
| VSO | 0.725 | 2.832 |
| VOS | 0.204 | 4.443 |
| OVS | 0.000 | 6.639 |
| OSV | 0.000 | 6.636 |

Table 9.1.5: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the Wall Street Journal corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.012 | 7.920 |
| SVO | 0.004 | 7.681 |
| VSO | 0.625 | 5.389 |
| VOS | 0.592 | 5.954 |
| OVS | 0.000 | 8.022 |
| OSV | 0.000 | 8.064 |

OVS > OSV. If we rank the word orders by their mean deviation scores, we get VSO > SVO > VOS > SOV > OSV > OVS. For this model world the first word entropies are 4.954 bits for SOV and SVO, 4.437 bits for VSO and VOS, and 4.437 bits for OVS and OSV.

### 9.1.5 Wall Street Journal

In this section I instatiate a model world on the basis of the Wall Street Journal (WSJ) corpus included in the Penn English Treebank. This corpus consists of approximately 49,000 sentences of written English taken from 1989 editions of Wall Street Journal. Using `tregex` I extract an event corpus of size 7319.

Figures 9.1.9 and 9.1.10 show the now standard plots for this model world.

Table 9.1.5 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get VSO > VOS > SOV > SVO > OVS > OSV. If we rank the word orders by their mean deviation scores, we get VSO > VOS > SVO > SOV > OVS > OSV. For this model world the first word entropies are 6.867 bits for SOV and SVO, 6.013 bits for VSO and VOS, and 6.899 bits for OVS and OSV.

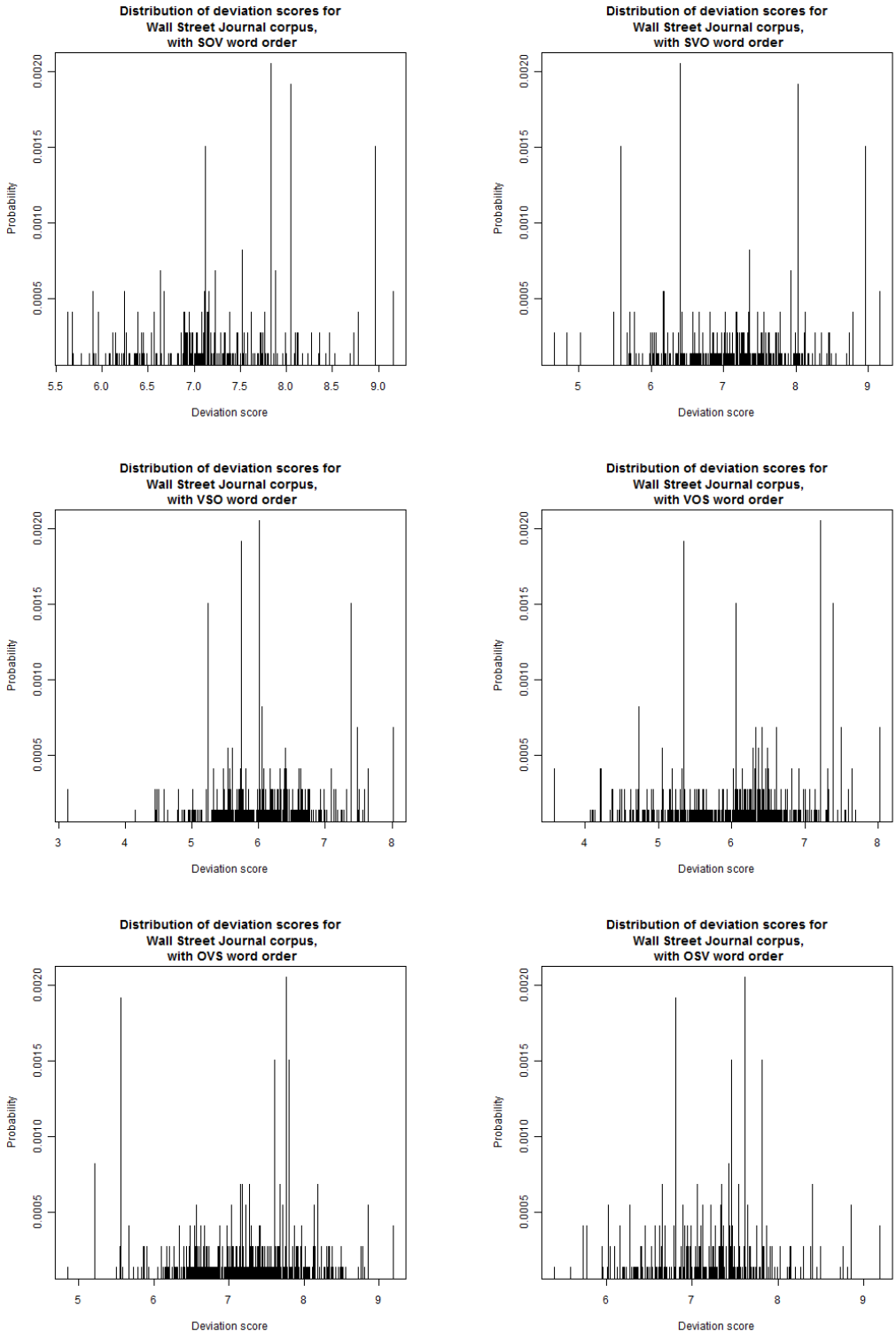


Figure 9.1.9: Distribution of deviation scores for all six basic word orders in the model world instantiated from the Wall Street Journal corpus.

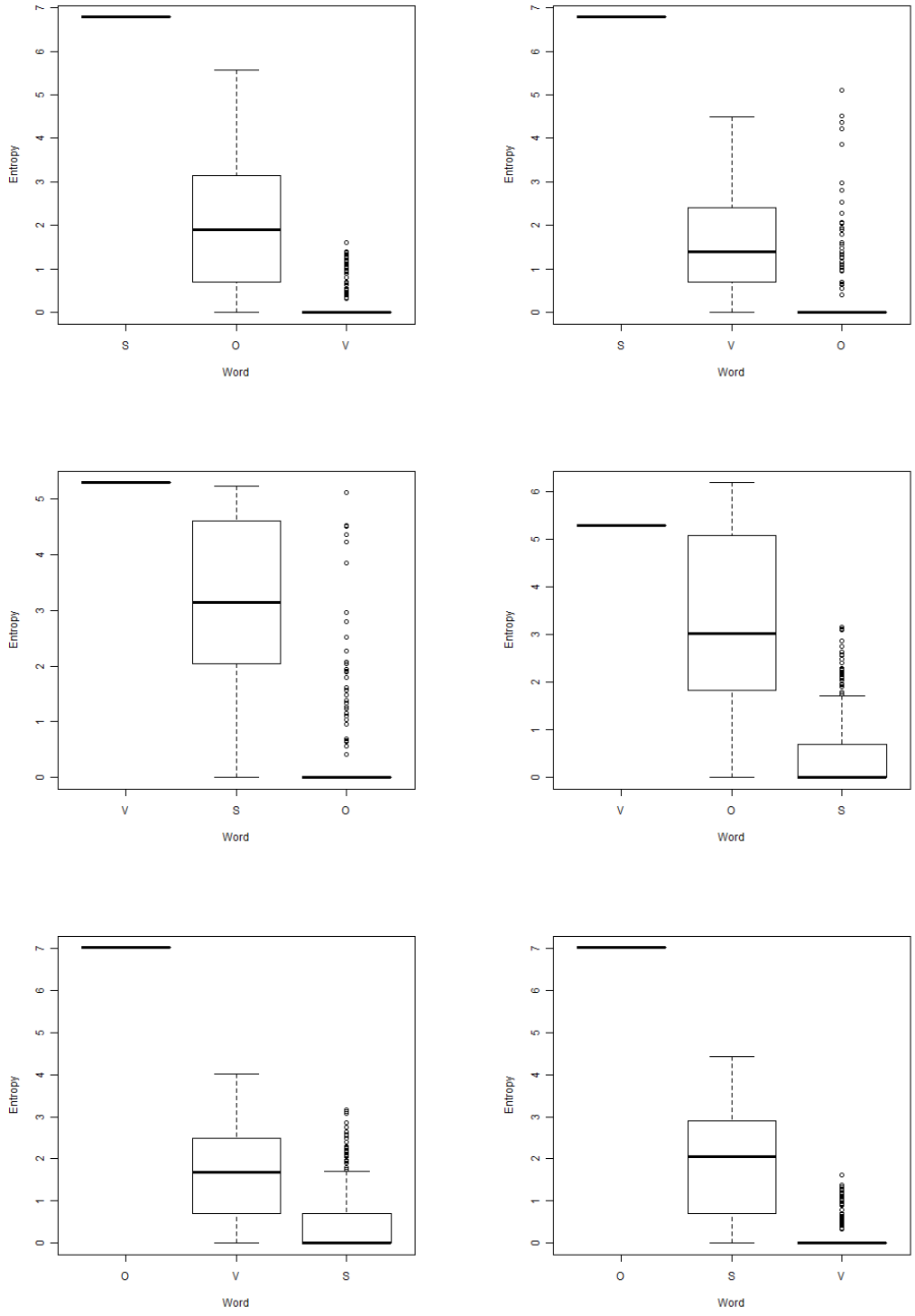


Figure 9.1.10: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the Wall Street Journal corpus.

Table 9.1.6: Global UID word order functionality measures for all six basic word orders in the model world instantiated from the merged English corpus.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.100 | 6.554 |
| SVO | 0.137 | 5.785 |
| VSO | 0.532 | 4.880 |
| VOS | 0.396 | 5.927 |
| OVS | 0.000 | 7.800 |
| OSV | 0.000 | 7.812 |

### 9.1.6 Merged English corpora

In this section I construct a "merged English corpora" by the simple process of concatenating the event corpora for all of the English corpora discussed previously. Because the same actions and objects are sometimes present in multiple corpora, concatenating them provides additional information about these common roles. This merged event corpus has just over 14,000 events in it.

Figures 9.1.11 and 9.1.12 show the now standard plots for this model world.

Table 9.1.6 shows the two global UID functionality measures for this model. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get VSO > VOS > SVO > SOV > OVS > OSV. If we rank the word orders by their mean deviation scores, we get VSO > SVO > VOS > SOV > OVS > OSV. For this model world the first word entropies are 5.670 bits for SOV and SVO, 5.022 bits for VSO and VOS, and 6.899 bits for OVS and OSV.

## 9.2 Experimental elicitation of the event distribution

In the previous section, the event distribution $\Phi$ was estimated on the basis of corpora of spoken and written language. In this section I describe the instantiation of a model world on the basis of the results of an experiment designed to measures people's perceptions of the relative likelihood of events.

The experiment consists of three parts. In the first part I choose the objects $\mathcal{O}$ and relations $\mathcal{R}$ for the model world based on the first words learned by English-speaking children, on the assumption that those words reflect objects and relations which are common and salient. The MacArthur Communicative Development Inventory (Fenson et al., 1994) provides a list of those words, along with norms for when they are learned. I identified all of the words that were either singly-transitive verbs or nouns that were potential subjects or ob-

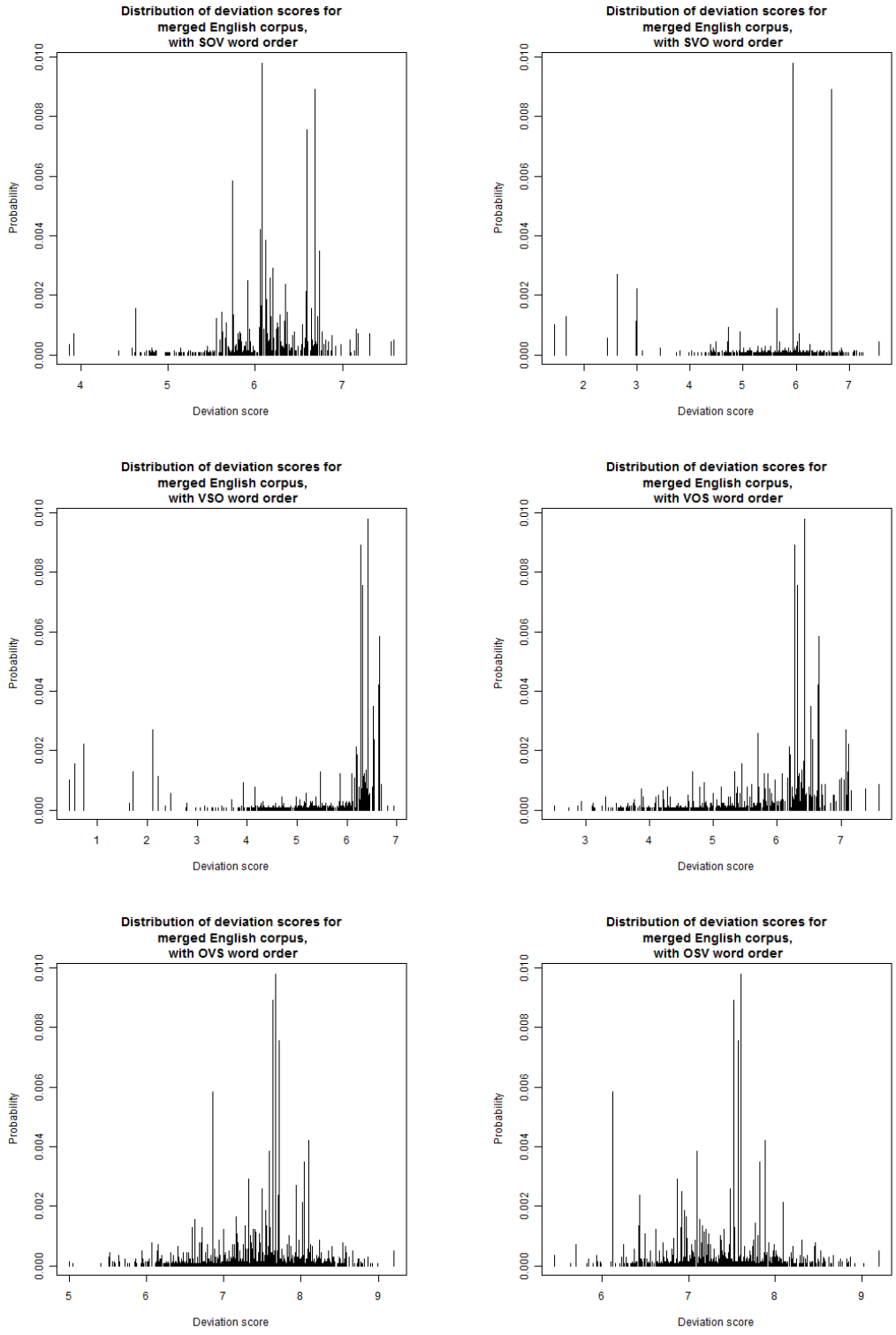


Figure 9.1.11: Distribution of deviation scores for all six basic word orders in the model world instantiated from the merged English corpus.

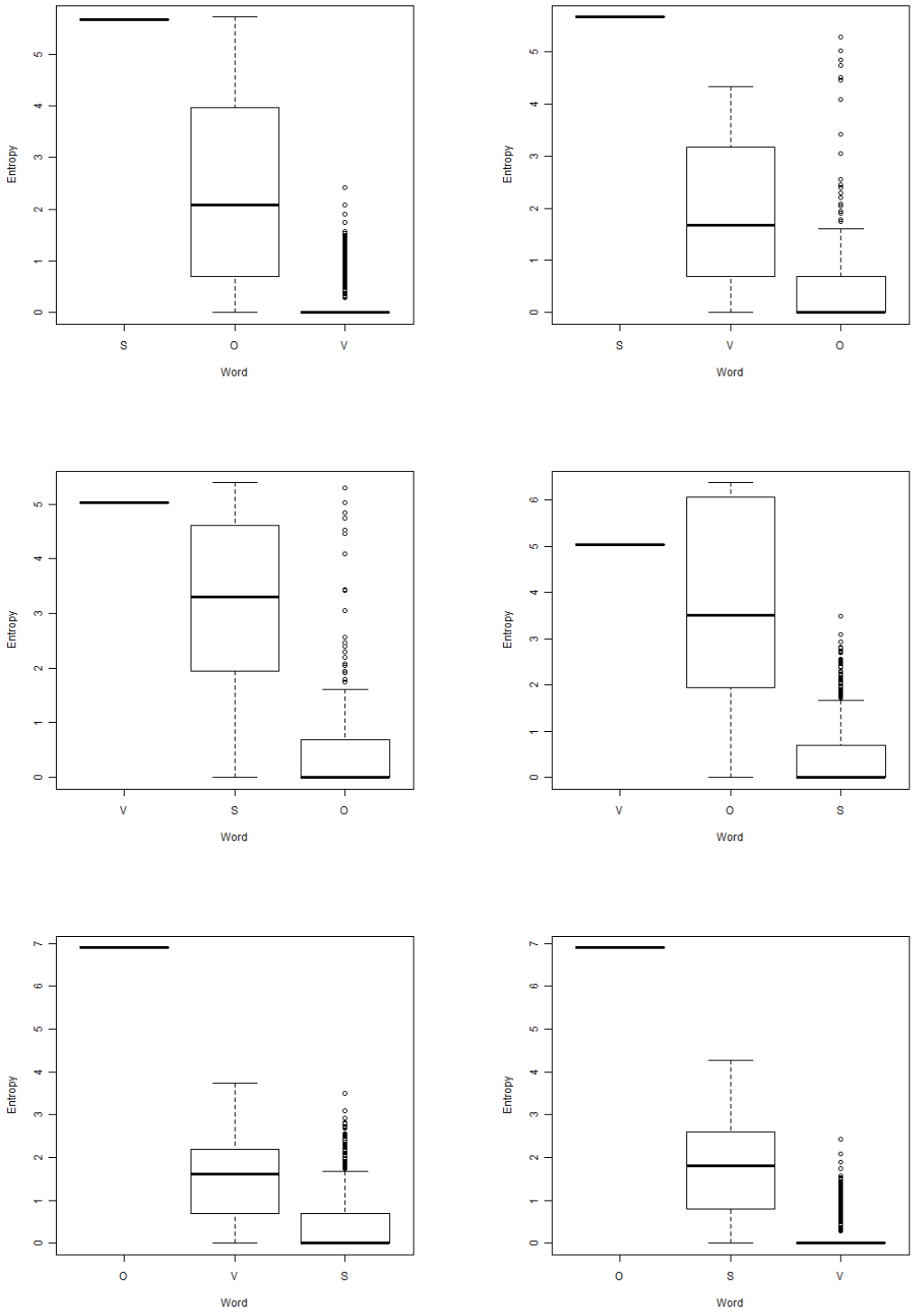


Figure 9.1.12: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from the merged English corpus.

Table 9.2.1: Objects and actions in our experiment's model world. Asterisks denote agent status.

| | |
|---|---|
| Objects | Apple, Bear*, Bed, Belly-Button, Blanket, Bunny*, Cat*, Chair, Cheese, Cookie, Cow*, Cracker, Cup, Diaper, Door, Duck*, Ear, Fish*, Flower, Foot*, Hair, Hand*, Hat, Horse*, Key*, Light, Milk, Mouth*, Nose*, Outside, Person*, Pig*, Spoon*, TV, Telephone, Toe*, Tooth*, Tree, Water |
| Actions | Bite, Drink, Eat, Help, Hug, Kiss, Open, Read, See, Swing |

jects for these verbs, yielding 324 nouns and 81 verbs. The only transformation I made to this list was to replace all nouns that referred to specific people (e.g., "Mommy" or "Grandpa") with a single noun "Person". In order to limit the total number of possible events to a number tractable for parts two and three of the experiment, I then identified the 40 objects and 10 relations[2] uttered by the highest percentage of children below the age of 16 months; these comprise the sets $\mathcal{O}$ and $\mathcal{R}$. The objects and relations are shown in Table 9.2.1.

The 40 objects and 10 relations in our world define a total of 16,000 events, but the overwhelming majority of the events in the world are physically impossible (e.g., (Television, Cat, Drink)) and thus should receive a probability of 0. The goal of the second part of the experiment was to identify these impossible events. The first step was to identify the subset of objects capable of being agents for at least one of the actions in the world. I did this on the basis of inspection and personal judgement, and the objects which I granted action status are indicated with asterisks in Table 9.2.1. I set the probability of events whose actors were not in this selected subset to zero, e.g. all of the events in the eventspace with an agent of "cheese" were given a probability of zero. This left 6,800 events without a probability assigned. Some of these events are still impossible, for instance (Cow, Chair, Open is plainly impossible, even though Cow is capable of acting as an agent. To identify these remaining impossible events, participants in the experiment were asked to judge each of the 6,800 remaining events as being either possible or impossible. At first, two judgements were received for each event. Those events which were judged as impossible by both participants were assigned a probability of zero. The events which received one possible judgement and one impossible judgement were then shown to a third participant, and events which the majority of participants agreed was impossible were given a probability of zero. At the end of this, 2,536 events remained which were possible and required a non-zero probability to be assigned to them. However, it seemed apparent that many participants had interpreted the object Outside as being an adverb in events such as (Bear, Eat, Outside), i.e. this event was interpreted as corresponding to an utterance "the bear ate the honey outside". This lead to many events which should

[2]The ratio of 4 objects for every 1 relation was chosen to reflect the proportion of each reported in (Fenson et al., 1994).

properly have been considered impossible being classed as possible. To correct for this, I set all events involving the object OUTSIDE which did not involve the action SEE (as OUTSIDE can legitimately be a patient of this action) to also be impossible. This further reduced the number of events to 2,352.

The final part of the experiment was designed to facilitate assigning probabilities to these 2,352 events. Participants were shown randomly selected pairs of these events and asked to indicate which of the two they thought was the most likely. The experiment was structured so that each of the events was compared to a total five others, and each of these pairings was judged by three distinct participants. The experiment yielded data in the form of a set of integer values $\{C_{ij}\}$, where $C_{ij}$ counts the number of participants who judged event $i$ to be more probable than event $j$, so that $0 \leq C_{ij} \leq 3$. These values were used to infer a probability distribution $\Phi$, using a model wich I will discuss shortly. First, however, I want to talk about how this experiment was implemented.

This experiment involved 11,839 two alternative forced choice tasks in the second part, and 35,280 2-AFC tasks in the third part, for a total of around 47,000 decisions. In order to implement an experiment of this scale, Amazon.com's "Mechanical Turk" web application was used to distribute the judgement tasks to a large, international pool of participants, who completed the tasks using their web browsers in exchange for small payments of cash or Amazon.com store credit. A total of 8,956 participants provided the judgements which were used in the experiment, presumably but not verifiably representing a broad range of nationalities, ages, levels of education, etc. Care was taken to remove data which was likely to have been produced by "bots", peices of software which complete Mechanical Turk tasks automatically at random in an attempt to quickly acrue payments for their operators. Participants who provided a large number of responses and whose responses did not generally agree with the responses given to identical questions by other participants were considered to be bots and their responses were not used.

The collected values of $\{C_{ij}\}$ were used to infer approximate event probabilities $\Phi$ by applying the Bayesian model presented in (B. Miller, Hemmer, Steyvers, & Lee, 2009), based on early work by Thurstone. The core of this model is a generative process for deriving values $C_{ij}$: each of the events $e_i$ is assumed to have a normal distribution associated with it, with mean $0 \leq \mu_i \leq 1$ and variance $\sigma_i^2$. When a participant is asked which of the events $e_i$ and $e_j$ is the most probable, they sample $s_i \sim N(\mu_i, \sigma_i^2)$ and $s_j \sim N(\mu_j, \sigma_j^2)$ and respond that event $e_i$ is more probable if $s_i > s_j$, and that event $e_j$ is more probable otherwise. Under this generative model, the values of $C_{ij}$ are binomially distributed variables, with $C_{ij} \sim B(3, p_{ij})$, where the "success" probability $p_{ij}$ is determined by $\mu_i, \sigma_i^2, \mu_j$ and $\sigma_j^2$ according to the following equation:

$$p_{ij} = \mathcal{CN}\left((\mu_i - \mu_j)/\sqrt{\sigma_i^2 + \sigma_j^2}\right), \tag{9.2.1}$$

where $\mathcal{CN}$ is the cumulative density function of the standard normal distri-

bution $N(0, 1)$. By placing a prior probability distribution over the set of all parameters, $\mu_i$ and $\sigma_i^2$ for $i = 1, 2, \ldots$, Bayes' rule can be used to compute a posterior probability over parameter values based on the values of $C_{ij}$ recorded by the experiment. I placed a component-wise prior on the parameters such that $P(\mu_i, \sigma_i^2) \propto \exp -\sigma_i^2$, with $P(\mu_i, \sigma_i^2) = 0$ if $\mu_i < 0$ or $\mu_i > 1$.

I performed the Bayesian inference numerically using a Metropolis Hastings algorithm to draw samples from the posterior distribution. The proposal process for the MH algorithm selects a single parameter to change from a uniform distribution over all the parameters, and then proposes a new value for that parameter by sampling from a normal distribution centred on the parameter's current value. The normal proposal distributions are not truncated, with the requirement that $0 \leq \mu_i \leq 1$ is enforced by the prior. To obtain an estimate of the values of $\mu_i$, I took 10 samples from each of 10 randomly initialized chains, for a total of 100 samples, with a lag of 1000 iterations between samples. Each chain was allowed to "burn in" for 15,000 iterations. This value was chosen by examining plots of the log posterior probability versus iterations and observing a plateau after around 15,000 iterations.

At the end of this process I had recovered a value of $\mu_i$ for each of the events in the model world. These values are transformed into an event distribution $E$ via the straightforward process of setting each event's probability to be directly proportional to its value of $\mu_i$. This yields a probability distribution which has quite a different shape from those derived from the corpora. The top left component of Figure 9.2.1 shows a histogram representation for the event distribution corresponding to the merged English corpora. A power low shape to the distribution is immediately evident: there are very many events with the lowest probabilities, a much smaller group which is just a little more probable, and so on with each increase in probability seeing a large drop in the number of events. This shape is shared among the event distributions derived from all the corpora. The top right component of the same figure shows the distribution produced using Bayesian inference on the experiment data. The difference in shape is clear. The rarest and the mose frequent events account for roughly equally sized subsets of the event space. The power law like distribution seen in the corpora distributions feels like the closer match to reality. In order to correct for this, I raise the probability of each event in the experiment's model world to the power of a fixed exponent $n$ and then renormalise. The other four components of Figure 9.2.1 shows the results of this process for $n = 2, 3, 4, 5$. A value of $n = 5$ yields the shape closest to that of the corpus distributions, so the distribution produced by this value is used to compute UID deviation scores.

It is natural to wonder how well the distribution which has been derived from this experiment agrees with our intuition of what a realistic event distribution shoudl look like. To this end, Table 9.2.2 shows the most and least probable completions of severalevent frames according to the inferred distribution. The completions are very much in line with common sense, although some

Table 9.2.2: Most and least probable completions of sentence frames according to experimentally determined event distribution $P$

| Sentence frame | Most probable completion | Least probable completion |
|---|---|---|
| Person Eat _____ | Apple | Door |
| Cat Drink _____ | Milk | Bed |
| Person _____ Cat | Help | Eat |
| _____ Eat Flower | Cow | Tooth |

Table 9.2.3: Global UID word order functionality measures for all six basic word orders in the model world instantiated on the basis of experimental data.

| Word order | Probability of optimality | Mean deviation score |
|---|---|---|
| SOV | 0.000 | 2.211 |
| SVO | 0.494 | 1.455 |
| VSO | 0.476 | 1.394 |
| VOS | 0.030 | 1.905 |
| OVS | 0.000 | 2.273 |
| OSV | 0.000 | 2.321 |

of the least probable completions are in fact physically impossible (e.g. (Cat, Drink, Bed)), suggesting that the filtering in part two of the experiment was not quite perfect in removing these events. Nevertheless, the judgements made in part three of the experiment seem to have successfully assigned these events the lowest probabilities, so that minimal error should be introduced.

Figures 9.2.2 and 9.2.3 show the now standard plots for this model world.

Table 9.2.3 shows the two global UID functionality measures for the experimental model world. If we rank the word orders by their probability of giving the optimal deviation score for a randomly selected event, we get SVO > VSO > VOS > SOV = OVS = OSV. If we rank the word orders by their mean deviation scores, we get VSO > SVO > VSO > SOV > OVS > OSV. For this model world the first word entropies are 2.441 bits for SOV and SVO, 2.035 bits for VSO and VOS, and 3.590 bits for OVS and OSV. Note that in this case initial subjects are less uncertain than initial verbs, which is the pattern displayed by the two CHILDES corpora.

## 9.3 Discussion

In the sections above I have constructed a total of 7 separate model worlds for the purpose of assessing the UID functionality of the six basic word orders. For each model world, I have produced 2 rankings of the word orders, corresponding

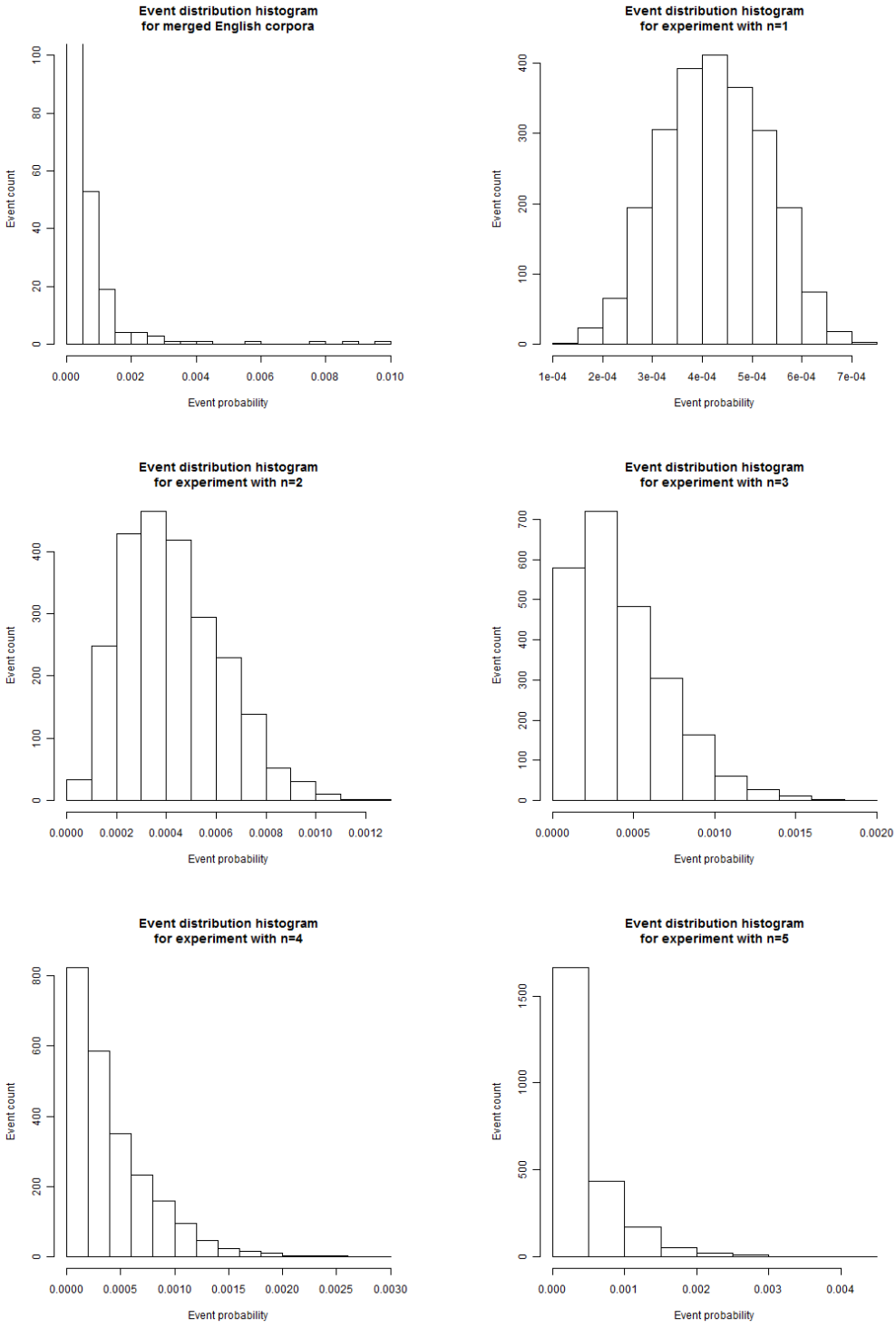


Figure 9.2.1: Results of raising the experiment model world's event distribution to various powers, with merged English corpus distribution for comparison.

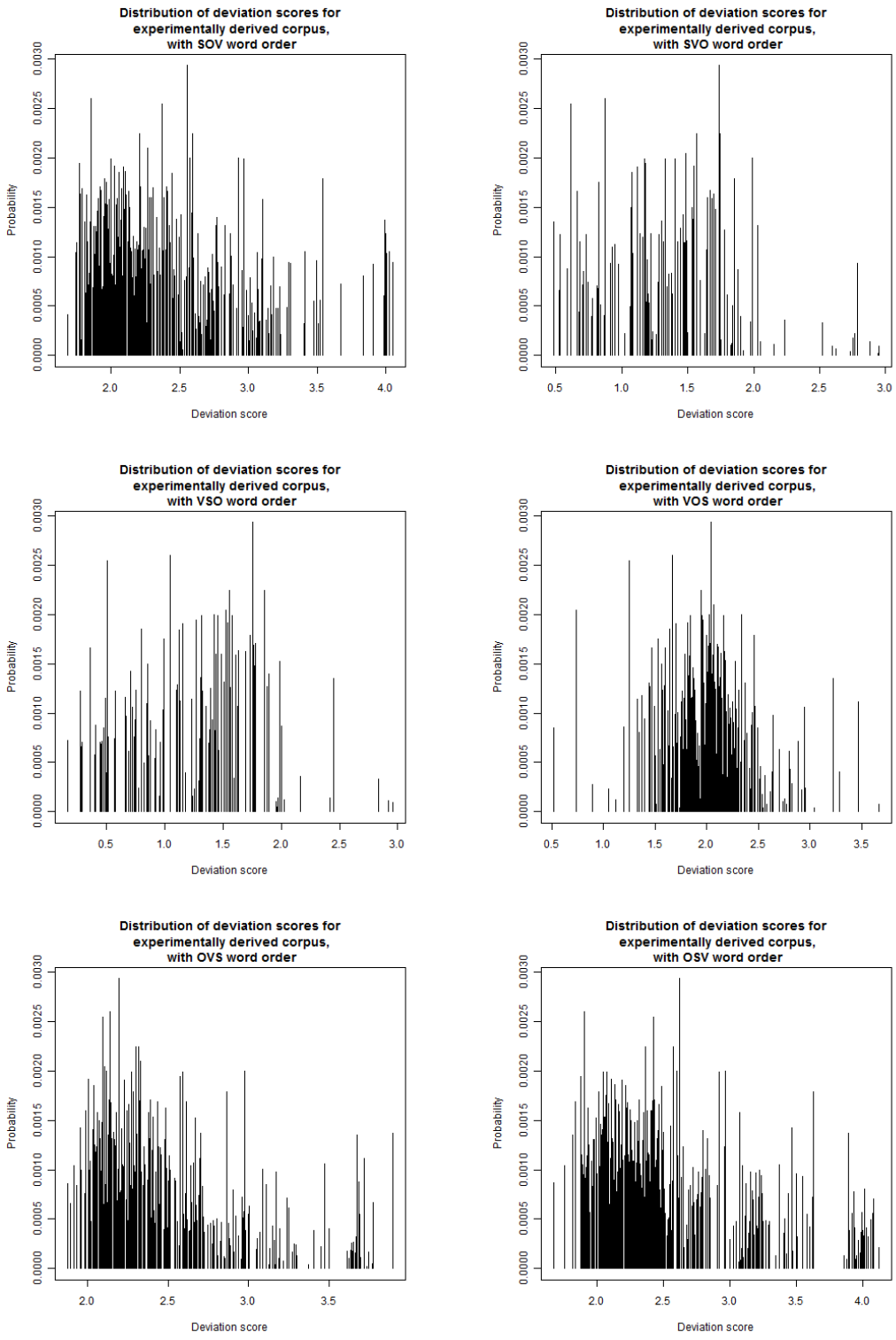


Figure 9.2.2: Distribution of deviation scores for all six basic word orders in the model world instantiated from experimenal data.

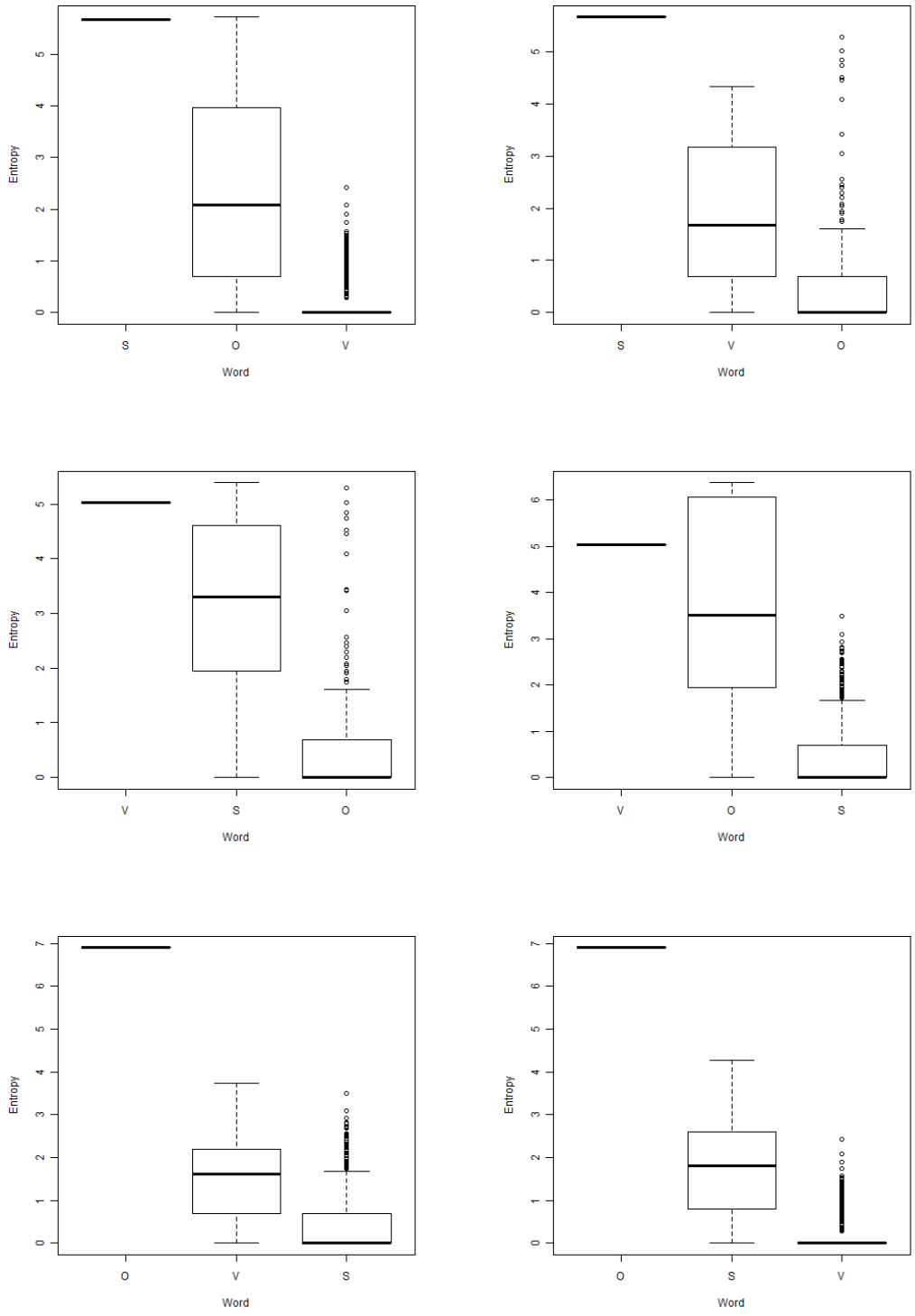


Figure 9.2.3: Per word entropy box and whisker plots for all six basic word orders in the model world instantiated from experimenal data.

to different global measures of UID functionality. This leads to a total of 12 rankings of the six word orders. Obviously, it is difficult to summarise this large a number of rankings straightforwardly, so I shall consider a number of features of the rankigns individually, beginning with those features which display the most consistency across model worlds and global measures.

The most striking feature of the word order rankings derived above is the consistence poor performance of the object-initial word orders, OVS and OSV. These two word orders occupy the lowest two positions (sometimes sharing them with an equal third-worst) in all 12 rankings. Thus, the aspect of UID word order functionality about which we can have the most certainty is that the object-initial word orders are of very low functionality. This is precisely what we require of an account of word order functionality, since OVS and OSV are the least frequent word orders.

One of the key requirements for an account of word order functionality that I established in Chapter 4 was that SVO and VSO be ranked as more functional than SOV. Looking at the 12 rankings given in this Chapter, we can see that SVO is ranked higher than SOV in 11 of these rankings. VSO is ranked higher than SOV in 9 of them. So on the whole, there is good evidence that the UID word order functionality hypothesis ranks SVO and VSO higher than SOV. This is again precisely what we require, as it can explain the long term pattern of drift from SOV to SVO and VSO.

A question that I left open in developing the requirements for the functionality ranking was the relative order of SVO and VSO. Of the 12 rankings I have provided, SVO ranks above VSO in 5 of them, with VSO ranking higher in the remaining 7. However, it is interesting to note how this ranking changes by whether the corpus is of spoken or written English, or experimentally derived. For considering this, I shall disaregard the merged corpus, as it is of mixed type. The two written corpora are the Brown and WSJ corpora, and these both clearly rank VSO above SVO. The three spoken corpora are the English and Japanese CHILDES corpora, and the Switchboard corpus. The first two of these clearly rank SVO above VSO. The Switchboard corpus ranks VSO above SVO, but notice that SVO is ranked much more highly in this case than in the written cases where SVO is ranked highest. This is suggestive of a SVO > VSO preference for spoken language and a VSO > SVO preference for written language. I will discuss why the UID functionality ranking may be different for these two cases in the next chapter. For the experimentally determined model world, the two different global measures of functionality give opposite preferences on this matter.

While it is not in strong agreement with the rankings presented here, later in this thesis I shall take the overall ranking of basic word orders by UID word order functionality to be SVO > VSO > SOV > VOS > OVS > OSV. My justification for this is twofold. First of all, there is some evidence that for spoken language SVO is more functional than VSO, and spoken language is the variety of language relevant for explaining basic word order distribution.

Second of all, as I shall discuss in the next chapter, there is good reason to suspect that the true ranking should be one in which object-final languages are more functional than object-middle languages, and these more functional than object-initial languages. Of the rankings I have presented here, those which rank SVO above VSO seem to fit this pattern best, in that they never allow an object-final to rank lower than third place.

## 9.4 Summary

In this Chapter I have used two approaches to measure the UID functionality of the six different basic word orders. One of these was to use a range of corpora in English and Japanese, spanning written and spoken and child-directed and adult-directed speech, to instantiate model worlds on which to calculate UID deviation scores. The other was to derive an event distribution for a model world from people's judgements of the relative probability of pairs of events. Although the different approaches and corpora, as well as the different global functionality measures, yield different functionality rankings, there are some clear and consistent comminalities among the ranings, and these are in good agreement with the most important requirements that I determined in Chapter 4.

# Chapter 10

# Discussion of UID functionality

## 10.1 A proposal for subexplanation E2

In Chapter 4, I established the following as one of two independent subexplanations required to construct a complete explanation of basic word order frequencies which was compatible with both syncrhonic and diachronic evidence:

**E1** An account of word order functionality which places SOV lower in the functionality ranking than SVO or VSO, but above the other three orders, except perhaps VOS.

Based on the ideas developed and the experimental results presented in the previous chapter, I am now in a position to offer the following proposal for subexplanation E2:

One a fully-developed language faculty had evolved, so that language was no longer slow and improvised but fast and implemented in dedicated neural hardware, one of the strongest functional factors influencing the functionality-driven change of human languages became the requirement to make appropriate trade-offs so as to simultaneously optimise language for efficiecy (i.e. to convey as much information per unit time as possible), reliability (i.e. to be robust against random noise and errors), and comprehensibility (i.e. to be able to be processed in real time with a minimum of information processing effort). Languages which do well in this trade-off will distribute the information in utterances uniformly over time, i.e. they will exhibit uniform information density (UID). The typological parameter of basic word order can have a substantial impact on how uniformly information is distributed across utterances. Based on analysis of corpora and people's judgements of event probabilities as elicited experimentally, the most probable ranking of basic word orders in terms of UID functionality is: SVO > VSO > (SOV, VOS) $\gg$ (OVS, OSV).

## 10.2 Undersanding UID word order functionality

I have derived the ranking of word orders by UID functionality provided above through the analysis of multiple sources of data, but so far it has not been explained any more deeply than by reference to these raw numbers. It would be significantly more satisfying if there was an explanation for the ranking not just in terms UID deviation scores, but also in terms of intuitively meaningful statements about the structure of the event distribution. Afterall, the only reason that the UID deviation scores can give the ranking they do is because of the structure of $\Phi$. Just what is it about the world and the way things happen in it which means that SVO word order conveys information with more uniform density than OSV word order? It may be that many separate aspects of the event distribution structure combine to create this situation, but I feel there is one fairly straightforward aspect which may capture a significant part of the effect.

For the purposes of this immediate dicussion, let us restrict our attention to event distributions in which all events with non-zero probability have precisely the same non-zero probability, i.e. event distributions which are characterised entirely by which events are possible and which are impossible. Obviously event distributions in the real world do not have this property, but this is a useful working assumption for now as it lets us focus on the consequences of structure in which events are possible and impossible.

I will come to argue that a few properties are typical of real world event distributions, and argue that these properties are sufficient to explain the word order functionality rankings found in the previous chapter. The properties are as follows. Firstly, that there are more objects in the world than there are actions. Secondly, that most of the objects in the model world are capable of being the agent of an event are capable of performing multiple actions. This follows from most agents being animate, intentional agents like humans or animals. There may, of course, be occasional agents who only engage in a single action, but we can safely assume that the average agent is be capable of, say, $n_{SV} > 1$ actions. Finally and similarly, the typical action can be applied to a range of patients, say $n_{VO} > 1$. These last two assumptions can be summarised simply by saying that the mappings from agents to compatible actions and from actions to compatible patients are both one-to-many.

I now establish an important consequence of this: that the patient of an event is significantly more informative about the other elements of that event than either the actor or action. Consider first the case of what an action can tell us about the actor. Each actor can perform more than one action, and because the set of actions is smaller than the set of actors, it can be expected that each action will be compatible with more than one actor (this follows from the well known "pigeon hole principle"). On average, each action will have $n_{VS} \simeq n_{SV}|\mathcal{O}|/|\mathcal{A}|$ actors compatible with it. Since $|\mathcal{O}| > |\mathcal{A}$ and $n_{SV} > 1$,

this quantity will be greater than 1. In other words, the mapping from actions to compatible actors is clearly one-to-many just like the mapping from actors to actions.

However, this is not the case with patients. Each action can be performed on more than one patient, but because the set of objects is larger than the set of actions, there will be fewer collisions in the action to patient mapping. The reverse mapping, from patients to actions, is better described as one-to-few than to one-to-many: On average, each patient will have $n_{OV} \simeq n_{VO}|\mathcal{A}|/|\mathcal{O}|$ actions compatible with it. If the ratio of $|\mathcal{O}|$ to $|\mathcal{V}|$ is in the vicinity of $n_{VO}$, this quantity will be approximately one. Thus, knowing the patient of an event and nothing more can significantly reduce the uncertainty of the action of the event.

This is relatively intuitive. Many objects have only a small number of actions which are typically performed of them. For example, if told that the patient of an event is "water", one may conclude with relative confidence that the action is likely to be "drink". Certainly one can be more confident of this than one can be, if told that the action of an event is "drink", that the patient is "water", for it may just as well be "juice", or "coffee" or any other number of drinks. This is an example of an action to patient mapping which is strongly one-to-many with a corresponding patient to action mapping which is one-to-few.

So I have established that patients are significantly more informative than actions. It is also straightforwardly true that they are significantly more informative than actors, since we have assumed from the outset that each actor is compatible with an average of $n_{SV}$ actions, each of which is compatible with an average of $n_{VO}$ patients, so that each actor is compatible with around $n_{SV}n_{VO}$ events.

Figure 10.2.1 contains a diagram showing the essence of this argument.

All of this suggests a "hierarchy of informativeness", with patients (corresponding to objects) being highly informative, and actors (corresponding to subjects) being minimally informative, with actions (verbs) in between.

This hierarchy explains perfectly the structure of the SVO > VSO > SOV > VOS > OSV > OVS ranking. The most highly ranked word orders are those which place the highly informative objects at the end of an utterance, while the lowest ranked orders are those which place objects up front. Word orders with the object in the middle are ranked in the middle. For each position of the object (front, middle, end), there are two compatible word orders, and we can see that the word order which puts the minimally informative subject before the more informative verb is always ranked higher, e.g. SVO > VSO.

Let us consider the extent to which the model worlds analysed in the previous chapter have structures resembling this simplication. Table 10.2.1 below shows some summary statistics of the model worlds' event distributions. Note that every event distribution adheres to the assumption of there being significantly more objects than actions, and of the mean number of actions per agent

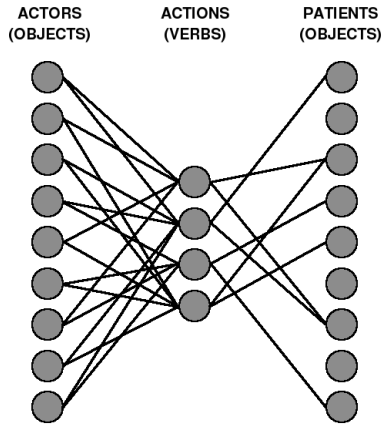


Figure 10.2.1: A diagram showing the essence of my intuitive explanation for why objects convey significantly more information than subjects or verbs. In this model world there are 9 objects and 4 actions. Each actor is compatible with 2 actions and each action is compatible with 2 objects. Because there are more objects than actions, the mapping from actors to actions features many collisions, and the actions to action mapping is one-to-many: each action is compatible with 3, 4 or 5 actors. However, because there are fewer actions than objects, the mapping from actions to patients features significantly fewer collisions, and the patients to actions mapping is approximately one-to-one.

and patients per action both exceeding 1. Also note a significant difference between the spoken and written corpora: spoken corpora tend to have fewer of the objects in their world acting as agents, whereas written corpora are the other way around. This difference may explain the fact that the relative ordering of SVO and VSO seems influenced by whether language is spoken or written. This is an interesting finding which warrants further investigation.

### 10.2.1 Is UID functionality universal?

An implicit assumption in the work of the past two chapters has been that the UID functionality of different word orders is universal, i.e. is the same for all languages. This basically translates into an assumption that when we consider the different environments and cultures in which different languages are situated, even though people may be discussing events involving different objects and actions, the large-scale statistical structure of their event distributions is unchanged: there is effectively one $\Phi$ distribution shared by everybody. It is not immediately clear that this absolutely must be the case. It is conceivable, for instance, that event distributions in Amazonian hunter-gatherer communities is substantially different from that of industrialised Western nations. This could be caused by such things as a larger variety of inanimate (and hence non-agent) objects in industrialised societies and more frequent discussion of non-human animals as patients in hunter-gatherer communities. This raises the intriguing possibility that object-initial languages, which are most common in

Table 10.2.1: Summary statistics of model world event distributions. From left to right, the first row of columns show name of model world, number of objects in the world, number of agents in the world, mean actions per agent and mean patients per action. The second row of columns show the proportion of objects in the world which are actors only, patients only or both.

| Corpus | $\lvert\mathcal{O}\rvert$ | $\lvert\mathcal{A}\rvert$ | Mean actions per agent | Mean patients per action |
|---|---|---|---|---|
| English CHILDES | 264 | 28 | 2.7 | 12.4 |
| Japanese CHILDES | 103 | 64 | 4.7 | 2.0 |
| Brown | 2004 | 720 | 1.7 | 2.9 |
| Switchboard | 686 | 252 | 2.1 | 3.1 |
| WSJ | 3627 | 1220 | 2.6 | 4.9 |
| Corpus | Prop. actors | Prop. patients | Prop. both | |
| English CHILDES | 0.087 | 0.833 | 0.080 | |
| Japanese CHILDES | 0.146 | 0.816 | 0.039 | |
| Brown | 0.435 | 0.411 | 0.154 | |
| Switchboard | 0.400 | 0.468 | 0.133 | |
| WSJ | 0.408 | 0.342 | 0.250 | |

small and isolated tribes, are not actually UID disfunctional languages, but are rational adaptations to unique communicative contexts. It is not immediately clear how this possibility could be tested, but it is worth bearing the possibility in mind.

## 10.3 Future research

There is some scope for future research in directions similar to that of the word developed in the previous two chapters. Most obviously, it would be valuable to repeat the same sorts of calculations for determining UID functionality using ever more and larger corpora, covering as many different languages as possible. There are, however, more interesting extensions as well.

Recall that in defining the generative model for utterances in the model world, I described the word order for each utterance as being drawn from a word order distribution $\Theta$. In the work above I have always set this distribution to a simple point-mass distribution so that only one word order is fixed. It is, of course, quite straightforward to redo all of the work above with more interesting distributions, in which multiple word orders are mixed together in different proportions. Different mixes will of course yield different deviation score distributions. It may be that some mixed distributions can in fact yield better performance than any individual order, for instance if the events which happen to have high deviation scores when described by one word order have low scores when described by another and vice versa, so that the two word orders compliment each others. Alternatively, it would be easy to define models

in which word orders are mixed in a deterministic fashion, such as by assigning two word orders as permissible and having agents always use the optimal word order for each utterance. There are some languages in which two different word orders are used with sufficiently close frequency that it is difficult to pick one as "basic", and some data on the different frequencies of these sorts of mixtures is available in (Gell-Mann & Ruhlen, In press). The sorts of extensions I have just described may be able to shed some light on these data.

It would be very interesting to see the results of experiments designed to test directly the prediction that people should find language comprehension under adverse conditions easier if the language uses more UID functional word orders. Experiments aimed at showing this may involve teaching participants artificial languages with different word orders and investigating how comprehension of the artificial language is changed by ambient noise or increased pronounciation speed. If it turned out, say, that participants could comprehend utterances in an SVO artificial language to a certain degree of accuracy with more noise or at a higher speed than in an SOV artificial language, this would be a strong piece of evidence for the UID word order functionality hypothesis.

# Part IV

# Conclusion

# Chapter 11

# Extending the explanation beyond basic word order

So far this thesis has focused exclusively on the cross-linguistic distribution of basic word order. My explanation for distribution of this single typological parameter has focused on two phenomena: the differential accessibility to processing of the different semantic roles within an event ("'word order in the language of thought"), and the idea of UID word order functionality. An important consideration of both of these phenomena is that their scope naturally extends much more widely than just to basic word order. Languages can also differ, for example, in whether adjectives precede the nouns they modify (as in English: "the black cat") or the nouns preceed the adjectives (as is usually the case in French: "le chat noir"). There is no reason that the same ideas cannot be applied to these word order parameters as well. The idea of word order in the language of thought can, in principle, extend to any semantic element of an utterance which is represented at something like the event level (and it seems reasonable to me to assume that adjectives are). Furthermore, the UID functionality principle will prefer some orderings of *any* constituents of an utterance over other orderings, regardless of whether they exist in non-linguistic mental representations, as long as those constituents are not statistically indepedent. As such, the true scope of the ideas developed in this thesis is the entirety of what is known as *word order typology*.

Therefore, although basic word order is the primary topic of this thesis, it is informative to devote a single chapter to brief, exploratory application of the thesis' two main ideas to general word order typology. Besides the inherent interest of the broader subject, this could serve as a source of additional evidence for UID functionality. It is, afterall, possible that UID functionality has provided the appropriate ranking of basic word by coincidence. As the previous chapters have shown, the appropriate ranking boils down to one in which the object is preferentially placed as far toward the end of the utterance as possible. There may be multiple principles which lead to this kind of ranking; how do we know that UID functionality is necessarily the correct one? If the

predictions of UID functionality are upheld across a range of independent word order parameters, then we may have more confidence that UID functionality is a genuine functional principle.

In this chapter I will briefly review basic word order typology and some principles which have been proposed to explain observations from that field. Since this chapter is intended only as brief "litmus test" of UID functionality in a broader environment, my reviews will be shorter and less detailed than those I presented in Chapter 3 for principles explaining the basic word order frequency rankings. After these reviews I will briefly consider two word order parameters, the relative ordering of nouns and adjectives, and the relative ordering of nouns and determiners. I will ask whether or not the way their values vary cross-linguistically, and particularly the way they interact with the relative order of objects and verbs, can be made sense of using an extended version of my explanation for basic word order distribution.

## 11.1 A brief review of word order typology

### 11.1.1 Other word order parameters and Greenbergian correlations

This thesis so far has been concerned with only a single word order parameter: basic word order The scope of word order typology extends to a number of additional parameters. Some especially important parameters are the relative order of adjectives and nouns (which can be AN, e.g. "the black cat" or NA, e.g. "le chat noir"), the relative order of genetive nouns and the nouns they modify (which can be GN, e.g. "the man's home" or NG, e.g. "the home of the man") and whether a language uses postpositions or prepositions (Po or Pr). Word order typologists are interested in language universals involving these parameters. The existence of these universals has driven a tendency of the field to categorise languages into "types", on the basis of their values of these parameters. My discussion in this section is based on that given by Comrie (1981).

The seminal work in word order typology is generally taken to be a paper by Greenberg (1963). In this work Greenberg proposes a total of 45 universals, on the basis of investigating a rather small sample of 30 languages. Many of the universals concern themselves with of word order. To give a flavour of the work, I reproduce a small subset of Greenberg's word order related universals below. The universals below have been chosen to demonstrate the scope of the universals, in terms of ordering parameters considered as well as the strength and complexity of the universals.

- Universal 1: In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object.

- Universal 3. Languages with dominant VSO order are always prepositional.

- Universal 7: If in a language with dominant SOV order, there is no alternative basic order, or only OSV as the alternative, then all adverbial modifiers of the verb likewise precede the verb.

- Universal 12. If a language has dominant order VSO in declarative sentences, it always puts interrogative words or phrases first in interrogative word questions; if it has dominant order SOV in declarative sentences, there is never such an invariant rule.

- Universal 17. With overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun.

As the samples above demonstrate, most (but not all) of Greenberg's universals are implicational. The universals are generally claimed to be either exceptionless (such as universal 3), or very strong statistical tendencies (such as universal 17, whose phrasing "with overwhelmingly more than chance frequency" is used in many others). Many of the universals are quite complex in their statement. This is a consequence of Greenberg's effort to state universals which were either exceptionless or close to exceptionless, rather than positing simpler universals which are not as strong. Of course, given the small sample of languages on which these universals are based, claims of exceptionlessness can be considered only tentative, a point which Greenberg himself makes quite clear.

Although the 45 universals themselves were the primary focus of Greenberg's paper, there is also some discussion of "big picture" patterns in the cross-linguistic distribution of various parameters. This aspect of Greenberg's work in particular has motivated further speculation. Greenberg considers four word order parameters in particular: basic word order (which he restricts to the values SOV, SVO and VSO, as no others occur in his sample), whether a language uses prepositions (Pr) or postpositions (Po), the relative order of adjectives and nouns (AN or NA) and the relative order of genitives and nouns (GN or NG). This set of four parameters has $3 \times 2 \times 2 \times 2 = 24$ logically possible combinations of values, but the 30 languages of Greenberg's samples attest only four of these, as shown in Table 11.1.1 below:

A few generalisations from this data are immediately clear. First, all the postpositional languages are also GN languages, and with a single exception (Norwegian), all the prepositional languages are also NG languages. In other words, the Po/Pr parameter and the GN/NG parameter are very closely correlated. Second, all of the SOV languages are postpositional (and, hence, GN), and the SVO and VSO languages are mostly prepositional (26 Pr vs 3 Po, and hence mostly NG). This suggests that the relative order of object and verb in basic word order is more important than the position of the subject from a typological point of view, with OV/Po/GN and VO/Pr/NG representing dom-

Table 11.1.1: Frequency of combinations of word order parameters in Greenberg's 30 language sample

| | SOV | SVO | VSO |
|---|---|---|---|
| Po/GN/NA | 5 | 2 | 0 |
| Po/GN/AN | 6 | 1 | 0 |
| Pr/GN/AN | 0 | 1 | 0 |
| Pr/NG/NA | 0 | 4 | 6 |
| Pr/NG/AN | 0 | 5 | 0 |

inant types. Finally, note an overall preference for NA order over AN order, with only 8 out of 30 languages considered showing NA.

Although Greenberg treats the OV/VO division relatively lightly in his work, the idea of this fundamental typological distinction has been picked up by many other linguists and carried significantly further. Lehmann and Vennemann are perhaps the linguists most strongly associated with this position. Through this work the idea has become established that many word order parameters have preferred values which are determined by whether a language belongs to the OV or VO type. These proposed correlations between the OV/VO parameter and other word order parameters have come to bear the label of "Greenbergian correlations", despite actually extending well beyond Greenberg's original proposals.

The OV/VO typology has been subject to two primary kinds of criticism. The first focuses on the validity of collapsing SVO, VSO and VOS into a single VO type, arguing in particular that SVO languages show substantially more variation in their other word order parameters than VSO and VOS. These critics claim that the natural response is to collapse VSO and VOS into one clear type, leave SOV as another and classify SVO languages as belonging to a "mixed type" somewhere inbetween these two ideals . Dryer (1991) closely scrutinises these criticisms and notes that, while they are not entirely without merit, SVO languages still generally reflect the same patterning as VSO and VOS languages (though with some noticable exceptions). He thus concludes that the OV/VO typology is valid overall. The second focus of criticism is the range of word order parameters which have been assumed to correlate with the OV/VO distinction. Dryer (1992) also concludes that many of the claims made in the literature, particularly by Lehmann and Vennemann, are not supported by empirical data. However, Dryer does conclude that at least 17 genuine correlation pairs exist, so that whether or not a language is OV or VO is certainly informative about the rest of its structure. It is worth emphasising that, even allowing for some exceptions in SVO languages and limiting the scope of the OV/VO typology to Dryer's 17 verified correlation pairs, the strongest form of the OV/VO typology is unequivocally false: it is not true that most languages represent exact instantiations of the ideal OV

or VO language. Comrie (1981) notes that in fact less than half the world's languages fit precisely into one of the two prescribed molds. However, it *is* true that most languages deviate relatively slightly from the ideal, so it is fair to think of the ideal OV and VO languages as being centroids of a fuzzy distribution.

Even though the OV/VO typology represents a sort of fuzzy classification rather than a strict division of languages into two baskets, the fact that as many correlations hold as do is striking and demands an explanation. Indeed, a number of explanations have been proposed, and many of them are functional. Could UID word order functionality hypothesis contribute to functional explanation of the Greenbergian word order correlations? This possibility is especially intriguing since it has been shown that the value of the OV/VO parameter played a strong role in determining the UID functionality of a basic word order.

### 11.1.2 Explanations for the Greenbergian correlations

The rich system of implicational universals in word order typology has prompted a number of explanatory principles to be proposed, although relatively few of these have survived close scrutiny over time.

Many proposed explanations for the Greenbergian word order correlations are extremely similar, and boil down to dividing words into two distinct categories and proposing some principle of consistent relative ordering between these categories. The two categories are usually defined semantically and are based on the concept of one kind of word modifying or acting on the other. The earliest mention of this kind of relationship is due to Greenberg (1963) himself, who notes "a general tendency to put modified before modifier" in VO languages, (i.e objects modify verbs, nouns in prepositional phrases modify the prepositions, etc). Vennemann (1973) classifies different types of word as either "operators" or "operands" and argues that there is a preference for languages in which operators are always on the same side of operands. Vennemann terms this the "principle of natural serialisation". J. Hawkins (1983) proposes a slightly modified version of this familiar story, in the form of "cross category harmony", or CCH, which "in a nutshell,...asserts...that there is a quantifiable preference for the ratio of preposed to postposed operators within one phrasal category (i.e. NP, VP/S, AdjP, AdpP) to generalize to the others. Whatever position the operand of one phrasal categority occupies in relation to all its operators will preferably be matched by the position of the operand in each of the other phrasal categories".

Dryer (1992) attempts to unify many of the above explanations by noting that they are all very close to what he labels the "Head Direction Theory" (HDT). Dryer introduces the terminological convention of calling the sentence constituents X and Y verb patterners and object patterners respectively if XY is the dominant order for VO languages and YX the dominant order for

OV languages. In this terminology, the HDT states that verb patterners are heads of phrases while object patterners are the dependents of those heads. A similar hypothesis is the "Head Complement Theory", which states that verb patterners are heads and object patterners are the complements of those heads. Dryer demonstrates conclusively that the HDT and HCT are both empirically inadequate: sometimes they predict OV/VO correlations which are "the wrong way around" with respect to correlations actually shown by the data, and sometimes they predict correlations which do not in fact exist at all.

Note that the explanations belonging to the HDT/HCT families are not very explicitly functional. It seems reasonable that languages are easier to learn if they are systematic, and that languages with consistent relative ordering of two broad categories are more systematic than languages where this ordering can vary freely. This is, however, a fairly weak account of functionality, and ideally independent motivation would be required for systematicity should be most important for one particular way of categorising words. To my knowledge, the first strongly functional approach to the problem of explaining the Greenbergian word order correlations is due to J. A. Hawkins (1990). His theory of "Early Immediate Constituents" (EIC) makes very precise predictions about the extent to which the human syntactic parser should prefer various orderings of constituents. The theory is based on a conception of the human parser inspired by discussions of modular input systems by Fodor (1983). The parser is assumed to construct nodes of a syntactic parse tree immediately and obligatorily as soon as this is permitted by the incoming stream of language. For example, consider the following sentence:

(26) I introduced some friends that John had brought to the party to Mary

In this sentence the parser is taken to construct a VP node immediately upon encountering the word "introduced" (since all verbs must belong to verb phrases). As a result it interprets the immediately following words as being daughters of that VP. Immediately upon encountering "some" the parser constructs an NP node and then attaches it to the VP constructed previously. All the subsequent words are placed inside that NP until the parser encounters "to", whereupon it immediately constructs a PP (and attaches it to the VP, not the NP, in accordance with the grammar rules).

Consider now the alternative ordering of (26) below:

(27) I introduced to Mary some friends that John had brought to the party

When parsing this sentence, the parser constructs the nodes in the order VP, PP, NP instead, such that the complete structure of the syntactic parse tree has been constructed by the time the parser is fed the word "some". This is five words into the sentence as opposed to 12 words into the sentence in the case of (26). Hawkins suggests that the parser prefers orderings of constituents such that, at any given node in a sentence's parse tree, all of that nodes's "immediate constituents" (i.e. immediate descendents) or ICs are recognised as quickly as possible. It is straightforward to construct a numerical measure

of how well this demand is met for any given node. Hawkins defines a node's "recognition domain" to be the set of left-to-right consecutive words under that node which must be parsed before all of the nodes ICs can be recognised. For example, in (27) above, the recognition domain of the VP node is "introduced to Mary some", since at this point the VPs two ICs (PP and NP) have been constructed. Parsing the rest of the sentence consists only of attaching words directly to already constructed syntactic nodes. For any given node, we can compute the ratio of the number of ICs in that domain to the total number of words in the domain, i.e. compute how many words are used on average to permit the recognition of each IC up until the point that all ICs have been recognised. This ratio permits us to rank various orderings of constituents from least to most preferred by the parser. Hawkins uses this techinque to show that the EIC correctly predicts a wide range of phenomena, including the fact that OV languages tend to be postpositional whereas VO languages tend to be prepositional.

In the same paper in which he demonstrates the inadequacy of the HDT and HCT, Dryer (1992) proposes the "Branching Direction Theory" (BDT), which states (roughly) that the OV and VO ideal languages correspond to languages which are consistently left branching and consistently right branching, respectively. The BDT makes predictions that are in better agreement with the data than either the HDT and HCT, and makes predictions which are also very similar to (but not precisely identical to) those made by Hawkins' EIC theory. More recently, Dryer (2008) has revisisted the BDT, focusing on the fact that its predictions depend crucially upon the particular phrase structure which constituents are supposed to hold. He notes that a recent trend of postulating "flatter" structures for certain phrases than was once fashionable appears to reduce BDT's ability to account for many of the universals it was originally held to explain. He goes on to observe that Hawkins' EIC theory is unaffected by this trend. This essentially leaves the EIC as the best present explanation for the Greenbergian word order correlations. Indeed, Newmeyer (1998) holds the EIC up as one of the best examples of a functional explanation of a language universal to have been proposed so far.

It is worth noting that recent work by Dunn, Greenhill, Levinson, and Gray (2011) has been claimed to pose a problem for functional explanations of the Greenbergian word order correlations. Using computational phylogenetic techniques from the field of evolutionary biology, these authors have used well-established language family trees and data on present day word order parameter values for many languages to investigate the extent to which changes in one word order parameter have historically correlated with changes in other word order parameters. Surprisingly, it is claimed that while correlated change does exist, it is largely lineage-specific. That is to say, parameters which are closely correlated in one line of related families may have evolved independently in a distinct family line. On the face of it, this finding is certain contrary to what we would expect if word order correlations had a functional origin. Since func-

tionality criteria are assumed to be equal for all language users, a correlation which is functional anywhere should be functional everywhere. It should be kept in mind, though, that this work is the first of its kind, and that the application of computational phylogenetic methods to historical linguistics is itself a very novel idea. It is unclear how well this challenge to the idea of a functional explanation of Greenbergian word order correlations will stand the test of time. Already, work by Levy and Daumé (2011) has identified shortcomings in the statistical methodology of Dunn et al. (2011), suggesting that larger and more diverse samples of language data, given the same treatment, may yield different results, potentially consistent with a functional view of the phenomenon. Given the tentative nature of this recent result, combined with the lack of an obvious alternative explanation, I shall continue to consider the prospect of developing functional explanations for the Greenbergian correlations.

## 11.2 Going beyond EIC

The EIC hypothesis is a well-motivated functional explanation for many correlations between the relative order of object and verb and other word order parameters. It also achieves a close fit with the data. However, there are word order parameters which it makes no predictions about, such as non-branching modifiers of nouns, including adjectives and determiners. These two parameters have been assumed in the past to pattern straightforwardly with the VO/OV parameter, e.g. it has been assumed that that VO languages are typically NA languages while OV languages are typically AN languages. However, Dryer (1992) has shown that this assumption is not supported. Nevertheless, it does not seem to be the case that these parameters are distributed roughly equally for all values of basic word order, and in the case of adjective and noun ordering, the two parameter values are not equally frequent. So some explanatory work seems necessary.

A possible UID-based explanation suggests itself. It seems intuitive that words which preceed nouns, such as adjectives, will provide some context which is useful for predicting those nouns. As a result, the presence of the adjective might decrease the uncertainty of the noun. Conversely, adjectives which come after nouns themselves might have their uncertainty reduced by the context provided by the noun. The straightforward implication is that the relative order of nouns and adjectives can influence the entropy profiles of utterances. The same is true of determiners, although it is not as intuitive that these convey much information about nouns (a point to which I shall return). As such, the UID hypothesis suggests that some settings of these parameters will be more functional than others. Since the basic word order parameter has a strong influence on the uncertainty of the nouns which adjectives and determiners can modify, it seems reasonable to expect some degree of interaction between these parameters. This is consistent with fact that the ratio of NA to AN and ND

to DN languages is not roughly the same for each basic word order. Since the influence of basic word order on entropy profiles is largely determined by the relative order of V and O, a UID-based interaction between these parameters could perhaps explain how so many linguists came to falsely believe that these parameters were correlated so closely with the VO/OV parameter.

In this section I propose an explanatory hypothesis for the cross-linguistic distribution of word order parameters. This hypothesis is a logical extension of the hypothesis I have used for basic word order, combined with EIC. I wish to emphasise the speculative nature of this hypothesis. I do not mean to claim that it is true, or even that it is probably true, and I do not claim to subject it to extensive comparison with data. My claim is simply that the hypothesis could very well prove to be true, that it would be a significant advance in word order typology if it were true, and that it passes a few very simple initial tests. My goal is to motivate work exploring these ideas more thoroughly. Here as in Chapter 7, I am inclined to quote Givón (1979): "While observed facts and facts deduced from facts are the flesh and bone of scientific inquiry, its heart and soul is creative speculation about the facts". Here goes, then.

The hypothesis I wish to propose consists of multiple parts. These are:

1 Many of the word order parameters which apply to natural languages (such as basic word order, use of prepositions or postpositions or the relative order of nouns and their various modifiers) can be profitably thought of as have corresponding parameters in the language of thought. The values of these parameters are fixed cross-linguistically due to the fact that the human CIS is largely uniform across the species. That is to say, some components of our compositional mental representations of the meanings of sentences are, e.g., more quickly retreivable from memory, can be processed more quickly or are made accessible to processing earlier than other constituents, etc., such that a kind of ordering is imposed.

2 All else being equal (to be qualified below), natural languages have a preference to adopt word order parameter values which match the values of the corresponding LOT parameters over values which conflict with the LOT parameters. Furthemore, the LOT parameter values will be strongly preferred in improvised communication, such that these values are the most probable for very early human languages and thus majority descent from these values is more likely than from other values.

3 Word order parameter values which differ from the corresponding LOT parameters may be adopted in the interests of increasing a language's functionality as determined by Early Immediate Constituents. "All else being equal" above includes, at least, EIC not predicting a preference for one parameter value over another.

4 Word order parameter values which differ from the corresponding LOT parameters may be adopted in the interests of increasing a language's

> functionality as determined by the UID hypothesis. "All else being equal" above includes, at least, UID functionality not predicting a preference for one parameter value over another.

We may summarise this hypothesis as follows. The observed cross-linguistic distribution of different word order parameter values is determined by the interaction of three individual principles, these being: (1) compatibility with the corresponding "word order" parameters in language of thought, (2) ease of syntactic parsing as dictated by the theory of Early Immediate Constituents, and (3) efficiency as a noisy channel code as dictated by the uniform information density hypothesis.

I wish to briefly note a few characteristics of this hypothesis. Firstly, its scope extends, in principle, to all word order parameters, although not all three of the underlying principles will necessarily make predictions about all word order parameters. Secondly, by virtue of the hypothesised likelihood of majority descent from word order values corresponding to the LOT values, it makes both synchronic and diachronic predictions. Thirdly, all three of the principles involved have very clear cognitive and functional groundings. Finally, all three of the principles involved are capable of being formulated sufficiently formally that their preferences for various parameters can be objectively and precisely agreed upon. This last point above is a substantial improvement upon many of the early explanations for the Greenbergian word order correlations, which relied upon sometimes poorly motivated divisions of words into categories such as operator and operand.

The question arises as to the relative importance of these three principles in determining word order parameter settings. The fact that EIC has been developed independently and that the correlations it has been designed to explain are relatively strong suggests that in those cases where EIC dictates a preference, it may be able to overpower the other principles. The question of interest, then, is how the larger hypothesis can explain situations where the EIC has nothing to say. I shall consider two such cases below.

### 11.2.1 Adjective and noun ordering

Let us consider the relative ordering of adjectives and nouns. As mentioned briefly above, it seems intuitive that adjectives which occur before a noun (i.e. prenominal adjectives) have the ability to decrease the uncertainty of the noun. To see this, consider the two sentences below.

(28) I ate the apple.

(29) I ate the green apple.

In particular consider the entropy of the word "apple" in both cases. In the shorter sentence, the only context constraining the word after "the" is that it must be part of a noun phrase describing something edible. In the longer sentence, the word "apple" occurs in a context where it is constrained to be part

of a noun phrase describing something which is both edible *and* green. Since the set of edible green things is a proper subset of the set of edible things, it follows that the entropy of the word "apple" is decreased in the second sentence, compared to the first. However, if English were an NA language, so that we had the sentences:

(30) I ate the apple.

(31) I ate the apple green.

then the word "apple" conveys an equal amount of information in both cases. As such, it is clear that prenominal adjectives necessarily decrease the entropy of the nouns they modify but postnominal adjectives do not. This suggests an information theoretic function of adjectives: they may be used to "smooth out" spikes in entropy caused by unexpected nouns, to borrow language from Levy and Jaeger (2007).

This function of prenominal adjectives does not come "for free", of course. The entropy of "apple" in (29) has been decreased relative to (28), but we must now consider the entropy of the word "green" which has been inserted into it. Just like the word "apple" carries some amount of information in the context of "I ate the", so too does the word "green". The amount of information "green" carries is influenced by how often it is used to describe something which a person eats. "Green" is perhaps not a surprising adjective to find in this context, since many fruit and vegetables are green, so the information it conveys may well be less than that conveyed by "apple" at the same position in the utterance. If this is the case, then the overall change is toward a more uniform distribution of information. However, the opposite may well be the case: in changing "I ate the eggplant" to "I ate the purple eggplant", the information conveyed by "eggplant" is drastically decreased since purple foods are quite rare, but by the very same token the word "purple" is introduced in a context where it is quite unexpected and it is not as apparent that much has been gained in terms of UID. In fact, we need to consider not just how unexpected "green' or "purple" are in the context of "I ate the", but also how unexpected it is that an adjective will be used *at all* in this context. This is an example of the information profile of an utterance being influenced by both semantic and syntactic considerations. Ultimately, the question of how often inserting a prenominal adjective is a functional thing to do in terms of UID is an empirical question which can only be answered with data from corpora or experiments. However, it is certainly true that prenominal adjectives *could* in principle be UID adaptive and that postnominal adjectives can never be UID adaptive.

Some interesting and recent work bearing on the empirical question of whether or not the predictive function of prenominal adjectives can be put to good use is due to Ramscar and Futrell (2011). Ramscar and Futrell observe that some frequently used adjectives in English appear to be redundant, e.g. "cute and small puppy" is a common noun phrase, despite the fact that

the majority of puppies are both small and cute. As a result, the adjectives in that noun phrase do not seem to convey any information which would not be conveyed by the simpler noun phrase "puppy". Observing that prenominal adjectives have a predictive function, (Ramscar & Futrell, 2011) hypothesise that prenominal adjectives will be used in English more frequently to modify infrequent (and hence higher entropy) nouns. Analysing a corpus of English text, they find that this is indeed the case: "more infrequent, more informative nouns are more likely to be preceded by adjectives...and [this] serves to manage noun entropy in context, helping to smooth spikes in entropy, thus allowing for more efficient communication". This work also demonstrates that prenominal adjectives are used in English to assist in the prediction of high frequency nouns when they appear in unusual contexts. This is evidence that, at least in English, prenominal adjectives *are* an efficient way to smooth out what would otherwise be spikes in entropy.

What are the implications of this for word order typology? Part III of this thesis has suggested that in OV word orders, the objects are very high in entropy. This leads to far less uniform information distribution than in VO word orders. Thus, we should expect that the predictive function of prenominal adjectives is most advantageous in OV word orders, and less (though still somewhat) advantageous in VO word orders. Since SVO appears to have more uniform information density than VSO, it follows again that the selective pressure to adopt AN word order will be stronger in VSO than SVO languages.

What of the language of thought? Assuming that adjectives and nouns are ordered in the language of thought, there are two possibilities. If the language of thought has AN word order, then the language of thought and the UID hypothesis are in agreement. Since EIC is indifferent in this matter, we should see an overall strong preference for AN word order in all languages. This is certainly not the case, however, as Dryer (2008) shows only 29% of languages with a dominant ordering of adjective and noun as being AN. This suggests that the language of thought has NA word order. If this is the case, we should expect to see the following cross-linguistic distribution of adjective and noun orders:

- A fairly strong preference for NA word order in SVO languages, where the language of thought is the dominant principle because the predictive function of prenominal adjectives is not especially valuable, since the SVO basic word order already yields quite good UID functionality.

- A less strong but still existing preference for NA word order in VSO languages, since VSO also yields good UID functionality due to being a VO language. However, its UID functionality is not as good as SVO, such that the UID functionality requirement for noun-adjective order "fights harder" against the language of thought.

- A stronger preference for AN word order in SOV languages than SVO or VSO. This is because SOV basic word order leads to quite poor UID

functionality with large entropy spikes at the O, so that using prenominal adjectives to smooth these spikes off becomes more important relative to matching the language of thought's preference.

Consulting Dryer (2008), this seems to be precisely what we find. The frequency of NA word order is 84% for SVO languages, 78% for VSO languages and 57% for SOV. This is consistent with a conflict between LOT word worder and UID functionality. Since the more UID functional setting of AN is still only seen in 43% of SOV languages, where it would be most useful, this distribution seems to suggest that the LOT principle may be stronger than the UID functionality principle - a fact which could perhaps explain why SOV word order is still as frequent as it is today, despite SVO and VSO offering drastically increased UID functionality.

### 11.2.2 Determiner and noun ordering and noun gender

Roughly half of the world's languages feature a system of so-called noun gender. In such a system, nouns are assigned, often quite arbitrarily, to distinct classes, known as genders. When a language has only two or three genders they are often referred to as masculine, feminine and neuter, but many languages have more than 3 genders. Gendered languages typically require agreement between some constituents and nouns, based on the gender. Determiners are a common example of this: the determiner "the" in English, which has no gender system, is replaced by "le" and "la" in French's two gender system, and by "die", "der" and "das" in German's three gender system. Grammatical gender is a curious phenomenon: it seems to have no useful function, given its often arbitrary implementations. This is a particular case of a broader sentiment expressed by Comrie (1981): "while the function of semantic roles and pragmatic roles can be readily understood in terms of the need for language to express semantic relations and apckage them in some way in terms of information flow, it is much less obvious why human language should require syntax (in the linguist's sense of syntax) at all". So far as I know, the question of why any language should have gender at all has not been particularly well explored. In this section I posit the possibility that gender can serve the purpose of UID-functionality in languages where determiners are gender-dependent and preceed nouns.

The intuition here is fairly straightforward. Consider the English sentence beginning "the boy ate the...". The nouns "apple" and "pizza" could both equally well complete this sentence, from the perspective of grammaticality. However, the same is not true of the corresponding French sentence beginning "Le garon a mang la...". "Pizza" could be a grammatical completion here, but "pomme" (apple) could not be, as French grammar would require "le pomme", not "la pomme". Thus, the gender-inflected determiner "la" has conveyed information regarding the final noun. The prenominal gender-inflected determiners are here providing the same sort of predictive function as prenominal adjectives have been shown to do.

According to the idea that OV languages have poor UID functionality as the objects convey a lot of information, one would expect OV languages to have the most to gain from this interaction. This suggests two possible typological consequences: either grammatical gender should be more common in OV languages than VO languages, and/or determiners should preceed nouns more frequently in OV languages than in VO languages.

Consulting Dryer (2008), I see no evidence for the first possibility - that the likelihood of a language having a gender system correlates with the relative order of verb and object. This does not immediately discount the story under consideration. It may be that gender also serves a purpose in VO languages, or that gender systems may be easily acquired by OV languages but are hard to lose once that language changes to VO basic word order. Assuming this is the case, then, the account developed here requires that the second possibility is true - that the relative order of determiners and nouns should change with basic word order in a similar way to the relative order of adjectives and nouns: prenominal determiners should be most frequent in OV languages, less frequent in VSO languages and least frequent in SVO languages.

Dryer (2008) provides data only for the order of nouns and demonstratives (such as "this" and "that"), not the order of nouns and articles (such as "a" and "the"), so this prediction can only be partially tested. However, to the extent that it can be, it appears to be upheld. Dryer gives the frequency of the DA setting to be 71% for SOV, 44% for VSO and 24% for SVO. This is consistent with an order of AD in the language of thought which is over-ridden to progressively greater degrees as the basic word order leads to less uniform information density and hence more pressure to adopt the DA value.

## 11.3 Summary

In this chapter I have presented a speculative hypothesis which extends the ideas I have developed elsewhere in the thesis for basic word order to language typology as a whole. Although much more work must be done before this hypothesis can be taken too seriously, it seems at first glance that it may to be able to explain the distribution of values for the order of nouns relative to adjectives and determiners. The hypothesis also offers an interesting big picture view on language typology: perhaps languages can be divided into two types, (1) those where the basic word order matches the language of thought at the expense of UID functionality, so that many other word order parameters are set to those preferred by UID functionality at the expense of additional LOT compatibility (SOV languages), and (2) those where the basic word order yields good UID functionality, so that other word order parameters are free to "relax" into the values preferred by the LOT. The involvement of noun gender in explaining the relationship between basic word order and order of noun and demonstrative is interesting as it hints at a possible functional explanation for

a seemingly unecessary feature of many languages. Of course, much more work is necessary before this can be considered anything more than speculation.

# Chapter 12

# Conclusion

## 12.1 Summary of proposed explanation

In this section I summarise the full "explanatory story" for basic word order frequencies which I have developed in this thesis.

The story begins, ultimately, in non-linguistic properties of the human mind. I have suggested that there exists in the mind a representation of events (that is, occurences involving an agent, action and patient) which is independent of the various representations that are employed in the course of perceiving those events. Furthermore, I have advanced the notion that this representation is compositional (that is, it is built up of separate representations of the agent, action and patient involved in the event) and that the different components of the representation may in important senses be differentially accessible to cognitive processing. In particular, I have put forward the hypothesis that agents are more accessible than patients and that both are more accessible than actions. In other words, the "word order of the language of thought" may be AgPaAc. I have presented the results of a reaction time experiment which provides some initial evidence in support of this hypothesis. In short, I have suggested that:

**All human beings internally represent events in their mind in a manner which can be meaningfully thought of as having the order: agent, patient, actor.**

As a result of the AgPaAc ordering of mental representations of events, AgPAc/SOV is also the "most natural" order for people to use in various tasks that involve producing physical orderings of representations of the components of an event. These tasks include arranging transparencies, conveying the meaning of an event using hand gestures or describing an event in spoken language. This all implies that the basic word order(s) of the earliest human language(s) was/were SOV (since the agent and patient of a sentence are more often than not also the subject and object of that sentence). This is a functionalist argument in defence of majority descent from SOV; in particular, it is an iconicity argument in favour of SOV (although a very different one compared to "standard" iconicity arguments, due to its deep cognitive grounding in mental

representation rather than arbitrary communicative principles). In short:

**Because humans think in agent, patient, actor order, SOV may have been the most natural word order for early "protolanguages", which relied on direct interaction between the conceptual-intentional system and the sensori-motor system. This implies that many early languages were SOV.**

This connection between mental representation and SOV word order in early language constitutes the subexplanation **E1** from Chapter 4.

As protolanguage gave way to full blown language, two important changes can be presumed to have taken place. Firstly, more special-purpose neural machinery would have been dedicatd to language processing, including the task of mediating between extra-linguistic representations such as the language of thought and the representations involved in constructing sentences. Secondly, language would have come to be used for a diverse range of tasks (including those of significant importance to survival and reproduction) and in a diverse range of environments (including those suboptimal for spoken communication). As such, the importance of external linguistic form closely mirroring the form of thought, would have decreased relative to the importance of language form facilitating reliable communication. That is, there may have been a selective pressure for communication which is rapid, requires minimal cognitive processing on the part of speakers and listeners, and is robust against external noise from the environment. In other words, the optimality criteria are different when one considers the earliest of languages, before the evolution of a specific language faculty and when communication was improvised using domain-general cognitive machinery, and when one considers languages which are routinely used for important tasks, processed by special purpose cognitive machinery. In short:

**Since the origin of language, there has been a systematic drift in word order toward those word orders which best serve the goal of conveying information in a way which is simultaneously rapid, easy to comprehend and resistant to external noise.**

The Uniform Information Density hypothesis sets one "gold standard" for what constitutes efficient communication in the sense alluded to above (though there are, of course, other considerations, such as words sounding sufficiently different that confusion is rare). I have connected the UID hypothesis to word order using mathematical modelling. Data from linguistic corpora and probability judgement experiments show good overall agreement in ranking the UID functionality of the object-initial word orders OVS and OSV lower than that of other word order. More broadly, there seems good reason to believe that VO word orders are more functional that OV word orders. The relative ranking of VO word orders seems to vary somewhat, although it does seem to be the case that SVO and VSO are likely to be more functional that VOS, due to their placing the object further toward the end of the sentence. To summarise:

**Efficient communication, in particular the requirement of uniform**

**information density, favours placing the object after the verb, and preferentially at the end of the sentence. As a result the SVO and VSO word orders are substantially preferred over SOV according to UID functionality. This suggests that the drift alluded to earlier may have mostly been SOV → SVO and SOV → VSO. Drift from any word order toward OVS or OSV should be very rare.**

This connection between word order and information density constitutes the subexplanation **E2** from Chapter 4.

The overall picture of the present day distribution of basic word orders which follows from this explanation is as follows:

**The presently observed distribution of basic word orders, which is strongly dominated by the three word orders SOV, SVO and VSO, is not an equilibrium distribution which directly reflects the relative functionality of the word orders. Rather, it is a snapshot of a dynamic linguistic landscape which shows both some influence of its initial conditions (SOV dominance) and some influence of its functionality-driven dynamics (drift toward SVO and VSO, rarety of OVS and OSV).**

## 12.2 Why are there still so many SOV languages around today?

Perhaps the biggest question surrounding the proposed explanation to which I have not yet given much attention is the following: if SOV is so much less functional than SVO and VSO, why is it still after many thousands of years so prevalent? I can think of two reasonable seeming responses to this.

The first is to simply posit that the drift from SOV to SVO and VSO is very slow and to date the drift toward functionality is only partially complete. Under this view, the present high frequency of SOV languages is only temporary. The cross-linguistic distribution of basic word orders has not yet reached equilibrium, and if we wait for another few thousand years we will see a linguistic landscape dominated by SVO and VSO languages. This is a fairly boring answer, although there is nothing inherently wrong with it. The big shortcoming though is that it is essentially unfalsifiable: we won't know if this answer is correct or not for thousands of years.

The second response is to suppose that there is something "stabilising" SOV languages, which counteracts the UID functionality driven drift away from it. Under this view, the cross-linguistic distribution of basic word orders *has* essentially reached equilibirum, and we should expect SOV, SVO and VSO to be the dominant word orders (though perhaps not in quite the same relative frequencies as we seen them today) for thousands of years to come. What could this stabilising force be? Without introducing any new ideas, we can hypothesise that while the effect of having SOV basic word order on the UID

functionality of SOV languages is quite negative, they still manage to achieve passable overall UID functionality by setting many other word order parameters to UID optimal values. These other parameter settings might compensate for the suboptimal choice of basic word order to such an extent that the pressure to change basic word order is relatively minimal. Unlike the first answer to the question of why SOV languages have persisted as much as they have, this hypothesis could be verified without a very long wait, through work extending our understanding of the UID functionality of various combinations of word order parameters - work I shall describe in much more detai later in this chapter.

## 12.3 Answering Tomlin's questions

In chapter 2, I reproduced eight specific questions which Tomlin (1986) suggested any account of basic word order frequency should endeavour to explain. In this section I revisit those questions and give the answers provided by the explanation developed in this thesis.

**Why do subject-initial languages outnumber verb-initial languages**? There are two separate reasons for this, one to explain the case of SOV and another to explain the case of SVO. The latter case has a more straightforward explanation: SVO languages outnumber verb-initial languages due to higher UID functionality. In the former, SOV languages outnumber VSO and VOS languages due to majority descent from SOV. This, in turn, may be a consequence of the architecture of the human conceptual-intentional system (CIS), namely the fact that AgPaAc is the "word order of the language of thought". Note that this explanation suggests that, were it possible to prevent any influence due to non-functional considerations, we should expect that SOV languages may be less frequent than VSO and possibly VOS languages.

**Why do object-initial languages occur so much less frequently than do subject-initial or verb-initial languages**? Object-initial languages occur with very low frequency because they have highly non-uniform information density. The reason for this is that, according to Chapter 10, objects are significantly more informative than either subjects or verbs. As such, placing an object at the beginning of a sentence results in an information profile where the majority of the information in a sentence is presented "up front", with the remainder of the sentence being relatively uninformative. This leads to a very uneven distribution of information and hence poor UID functionality.

**Why do VSO languages outnumber VOS**? VSO languages outnumber VOS languages due to higher UID functionality.

**Why do VOS and OVS languages outnumber OSV**? VOS languages outnumber OSV languages due to higher UID functionality, as a consequence of the general property that VO languages are substantially more UID functional than OV languages (for the same reasons discussed earlier).

**Why do SVO languages outnumber VSO?** SVO languages outnumber

VSO languages due to higher UID functionality.

**Why are VOS and OVS languages of approximately equal frequency?** This is perhaps the question of Tomlin's which my account is least able to answer. VOS should certainly be more UID functional than OVS due to the object following the verb, although VOS is the least functional of the VO word orders. As far as shortcomings go, I think this one is fairly minor: VOS and OVS languages are both very rare, and the difference in frequency between them is much less pronounced than most other features of the basic word order distribution.

**why do SOV languages outnumber VSO languages** and **why are SOV and SVO languages of approximately equal frequency**? The answer to this question depends on how we account for the high frequency of SVO languages despite their relatively low UID functionality, an issue I described above. On the one hand, it is possible that there is no especially deep or insightful reason for either of these facts. They may simply be a consequence of the particular moment in time at which basic word order frequencies have been observed, which is a moment when the cross-linguistic distribution has not yet reached equilibrium. Under this interpreation, it would be reasonable to expect that in the future SVO languages will be more frequent than SOV languages. On the other hand, it may be that SOV languages are stabilised somehow, possibly by UID-optimal settings of other word order parameters, and that this has stopped their frequency declining below that of the more functional VSO and SVO languages.

## 12.4 Assessment of proposed explanation

The explanation that I have developed in this thesis for basic word order frequencies is certainly not simple by the standards of other explanations that have been offered for the same phenomenon. While I have done my best to motivate all the various interacting aspects of the explanation in the preceeding chapters, I expect that some people may be uncomfortable with the various inferential jumps which have been made throughout. As such, I now want to spend some time highlighting what I consider to be some of the reasons why the explanation should be taken seriously, above and beyond the evidence presented in Chapters 6 and 9.

### 12.4.1 Explanatory adequacy

First and foremost, we have to consider the issue of explanatory adequacy. If we accept the arguments I developed in Chapter 4 - and these are not entirely new arguments but simply logical extensions of arguments which have already appeared multiple times in the literature - then it is a straightforward fact that none of the previously proposed explanations for basic word order distribution

are simultaneously consistent with the existing synchronic *and* diachronic evidence. If we are interested in explanations which do meet both these criteria, then the explanation I have developed here is, at least for now, the only game in town. This does not, of course, mean that it should be accepted without scrutiny, but to some extent it is a defence against the fact that it is a substantially more complicated explanation than previous ones. Certainly good explanations should be as simple as is necessary, but they should not be simpler than is useful.

### 12.4.2 Formal precision

One advantage of the explanation offered in this thesis over alternatives is the degree of formalisation and precision which it enables. Ultimately, it rests on two quantitative concepts: reaction time in the case of the SOVLOT hypothesis, and entropy in the case of the UID word order functionality hypothesis. These are both quantities with clear and objective definitions which can be measured independent of any particular language. Once a probabilistic language model is defined, there is no ambiguity as to what the functionalities of the different word orders are according to that model. This contrasts withs many of the concepts involved in other explanations of basic word order. Often it can be a subjective matter of opinion as to which of two noun phrases is the most thematic or topical, heaviest, etc. It is easier to assess the agreement between theory and data when the theory is formally defined and the concepts involved can be measured objectively and precisely. Not only that, but we can be more confident about the level of agreement which is found, as there is minimal room for spurious agreement introduced through subjective decisions.

### 12.4.3 Utilisation of pre-existing ideas and lack of problem specificity

Two additional properties of this explanation are fairly closely related.

The first is the fact that the explanation is constructed out of pre-existing ideas. As a result major components of the theory have already been justified on independent grounds. If I had synthesised the UID hypothesis for the first time herein and applied it to the problem of basic word order, although the same functionality ranking would have been derived, the validity of its use to explain basic word order frequencies would have been more suspect. In such a situation it could well be argued that the ranking produced by the UID hypothesis meets the newly established criteria simply by accident, and human languages are actually not similar to ideal codes. However, the UID hypothesis has already been tested independently multiple times and found to explain multiple different phenomena. It is therefore less likely that it provides an appropriate ranking of word orders by accident. This work in fact strengthens

the case for the idea that the UID hypothesis is true and applies to a large range of linguistic phenomena.

The other related property is that it is not necessarily specific to the problem of basic word order. Although this thesis only develops in necessary detail the application of the UID hypothesis to basic word order distribution, I have argued (I hope convincingly) that there is some promise for it applying to word order typology more broadly, and this can only increase the value of the theory. Given two explanations which do equally good jobs of explaining the distribution of basic word orders, we should prefer one which can also provide an explanation for other phenomena over one which cannot.

Both of these properties can perhaps be summarised simultaneously by stating that the explanation here is well integrated into overall linguistic and cognitive theory. It is not a case of several new ideas being developed for the purpose of explaining a few individual points of data. Rather, it is a further refinement of pre-existing perspectives on the mind, language and the relationship between them. This leads to a big picture view in which a large number of independent problems are explained through the consistent application of a small number of fundamental principles.

## 12.5 Future research

It is clear that there is tremendous scope for future research above and beyond the immediate extensions to my work discussed at the ends of Parts II and III.

### 12.5.1 Better mapping the language of thought

One of the main suggestions advanced in thie thesis is that the language of thought can profitably be thought of as having a "word order", in the sense of some components of mental representations being more accessible than others to memorisation and processing. Although Chapter 6 presented some empirical evidence in support of this idea, there is a wide range of work both experimental and theoretical which is necessary to further explore this idea.

On the experimental side, the basic paradigm is well exemplified by the experiment I described in Chapter 6. There would be value in conducting multiple experiments of this kind using larger sets of participants, with participants better balanced across a wider range of native language word orders, testing performance on a wider range of non-linguistic memory and processing tasks and using training stimuli in a variety of sensory modalities. As Chapter 6 discussed, strong support for the SOVLOT hypothesis can come only from finding consistent accessibility rankings in a wide range of separate tasks. Furthermore, the nature of the stimuli and the tasks should be expanded so that we can seek evidence not only of differential accessibility between agents, patients and actions but between other hypothesised semantic components of

mental representations (e.g. adjectives and adverbs). The more detailed a picture we can build up of the compositional structure of mental representations of the meanings of utterances, the better we can investigate extending the basic explanatory tools derived here to word order typology in general.

On the theoretical side, there is the substantial task of developing formal models of the cognitive processing underlying the memory tasks used in these experiments. This task is obviously of interest to cognitive psychology in general, even when not considered in the context of language, but the possibility also exists that this sort of modelling may lead the way to psycholinguistic models shedding some light on how differential accessibility in the language of thought leads to word order biases in improvised communication tasks. Stochastic models of the sort developed by Townsend and Ashby (1983) seem like an appropriate starting point for this work.

### 12.5.2 Investigating the UID functionality of additional word order parameters

For the UID word order functionality hypothesis, the most obvious direction for future work is to take motivation from the early work in Chapter 11 and to extend the work to encompass additional word order parameters beyond basic word order. This, combined with greater knowledge about word order in the language of thought, would facilitate better testing of the very general theory of word order typology sketched out in Chapter 11. The primary challenge here is the development of probabilistic language models sufficient for estimating the information profiles of more complicated sentences that include, for example, adjectives, adverbs, prepositions, oblique noun phrases, embedded sentences, etc. It seems unlikely that sufficiently realistic probability distributions over these kinds of large and complicated sentences could be constructed from corpora alone: the data are likely to be too sparse. The most promising approach would probably be to combine the use of very large corpora with the use of algorithms for generalising the sparse distributions inferred from those corpora to more complete distributions on the basis of either regularities in the sparse distribution or additional non-linguistic information. For instance, suppose that the verb "eat" and the noun "cabbage" both appear frequently in the corpus, but never in the same sentence. This would mean there is no direct evidence that, e.g, "the girl ate the cabbage" is a fairly unsurprising sentence. A method based purely on corpus analysis would assign this sentence a very low or zero probability, which is not what we want. However, a well-designed algorithm may be able to infer that "the girl ate the cabbage" is unsurprising by reasoning as follows:

1 The words "fruit", "pasta" and "steak" appear in the corpus as objects of the verb "eat" often,

2 These words are often modified in the corpus by certain adjective such

as "fresh" or "delicious",

3 Words which are never modified by these adjectives rarely or never appear in the corpus as objects of "eat",

4 "Cabbage" is often modified by these adjectives.

Alternatively, we may be able to provide the algorithm with non-linguistic data such as pairings of attributes and objects, such that it knows that "cabbage" is a FOOD. While these techniques are a likely computational requirement of testing the applicability of LOT and UID to word order typology in general, they are also interesting from a psychological perspective. Presumably the probabilistic models which underly human language processing are not instantiated purely on the basis of linguistic material heard over the course of a person's life, but rather are also influenced by all manner of extra-linguistic experience. Thus in developing algorithms to facilitate constructing probability distributions to explore the UID implications of different combinations of word order parameters, we can simultaneously address a psychological problem of interest if we strive to keep the algorithms cognitively plausible. Existing models which may be a useful starting point for this work are Kemp, Griffiths, and Tenenbaum (2004) and Johnson, Demuth, Frank, and Jones (2010).

### 12.5.3 Simulating language change according to the explanation

The work described above is necessary but perhaps not sufficient to allow a thorough empirical testing of the UID word order typology hypothesis. My earlier discussion of Nettle's basic framework for explaining language diversity pointed out that an essential ingredient in functionalist explanations is the fact that the many different parameters which can contribute to the functionality of a language provide a "virtual ecosystem". Thus, rather than an overall trend toward a single functionally optimal language, the passage of time can result in the diversification of languages as individual languages evolve toward various local maxima in the functionality landscape. Because of the complicated process by which languages change in this situation, it may not be the case that the relative frequencies of different word order types directly reflects their functionality.

In order to investigate how well the predictions of the UID word order typology hypothesis, match the typological data, it may be necessary to conduct simulations of languages diverging from a common ancestor (in which all word order parameters are set in accordance with the language of thought). These would not necessarily need to be faithful simulations of the actual process by which linguistic differentiation occurs, as local optima would still be revealed using standard optimisation algorithms. However, there is an additional motivation for performing these simulations, which is to address an important

component of functional explanations of language universals that I have so far overlooked. This is the issue of solving the *problem of linkage*.

Kirby (1999) defines the problem of linkage as follows: "given a set of observed constraints on cross-linguistic variation, and a corresponding pattern of functional preference, an explanation of this fit will solve the problem: how does the latter give rise to the former?". Ultimately, a language universal is not completely explained by showing that the relevant distribution of linguistic features in some way agrees with some account of functionality. The question still remains: how do languages come to change preferentialsly in the direction of the functional ideal? This aspect of functional explanations of language change was glossed over in Nettle's framework as I discussed it in Chapter 3. How, precisely, does functional selection amplify linguistic variations thatincrease functionality?

In the case of biological evolution, by which Nettle's account is inspired, this process is quite straightforward: genetic variation which increases reproductive fitness, by definition, causes the organism posessing the variation to reproduce more on average than other organisms. There seems to be no straightforward equivalent of this in the case of language. It is not at all clear that if a language user speaks with some linguistic variation which increases overall language functionality that they will necessarily have more children than other members of their language community; even if they did, it would not necessarily follow that their children would learn the variation from them. Children of first-generation immigrants, for example, typically develop the "native" accent of their new home's language, rather than their parents'. Nor is it necessarily the case that children of the following generation will somehow "just know" to pay more attention to the variant speaker's language than to that of other adults when learning the community's language.

Amplification of linguistic variation may follow fairly automatically if the variation increases the learnability of the language in some direct way, such as by being more compatible with an innate learning bias, however this cannot be taken as a given in many situations, such as functional variations which decrease the likelihood of ambiguity. For each separate aspect of functionality, some causal mechanism linking differential functionality with differential cross-linguistic frequency must be provided, i.e. the problem of linkage must be solved for each separate component of functionality, and this is something which has historically received very little attention (this fact is presumably the origin of Manning and Parker's unusual dismissal of functionalism, as discussed in Chapter 3.

A traditional answer to the problem of linkage, which Kirby and Hurford (1997) calls "phylogenetic functionalism" is based on the (widespread but not uncontroversial) assumption the language acquisition process is heavily constrained by a genetically coded language acquisition device (LAD). If we also assume that people who speak languages which are more functional have higher reproductive fitness, then the problem of linkage is solved for that particular

aspect of functionality: biological natural selection acts on the genetic specification of the LAD in such a way that over time language acquisition becomes strongly biased in such a way that only highly functional languages can be easily learned. This is essentially the view presented in the well known work of Pinker and Bloom (1990).

However, there is some question as to whether or not phylogenetic functionalism is always possible, or whether it will be prevented because the pressure to be functional has to compete with the pressure to learn an intelligible and useable language. Consider a hypothetical time when the LAD was hardcoded for a strong bias toward SOV and no other word order. Now suppose a mutant is born whose LAD is heavily biased toward SVO. In principle this mutant would be able to learn a more functional language (assuming the UID functionality ranking is correct). However in practice he may either simply fail to learn his community's SOV language or learn a kind of SOV/SVO hybrid which is still technically more functional but is of relatively little benefit to the mutant because the rest of his community struggles to understand his occasional use of SVO. In these situations the mutant is actually less biologically fit than other members of his community, or at least certainly is not *more* fit as phylogenetic functionalism requires. See Christiansen and Chater (2008) and Chater, Reali, and Christiansen (2009) for discussion of this limitation to phylogenetic functionalism.

Kirby and Hurford (1997) offers a novel solution to the problem of linkage in the form of what he calls "glossogenetic functionalism". Under this account, languages change in the direction of functionality via a process of "linguistic selection", which is entirely non-genetic in nature. Only after a functional linguistic variation has spread throughout the community does natural selection on the LAD introduce a genetic bias toward functionalism, by selecting against mutations which introduce a bias in the direction of less-functional variations. The process of linguistic selection occurs on a much faster timescale than natural selection, and occurs through a filtering of the "arena of use". Children learn their native language not directly on the basis of the previous generation's internalised linguistic competence, but on the basis of a set of utterances created using that competence. Exactly which utterances make it into that set, and hence which aspects of the previous generation's linguistic competence are unambiguously demonstrated by those utterances, depends upon various factors. For instance, grammatical rules which only apply in rare situations are less likely to make it into the set of training utterances for any one child than rules which apply universally. Utterances which are difficult for children to correctly parse may also have little influence on the language which those children learn. Thus, language variations can be amplified or supressed depending on the extent to which utterances which exemplify the variation make it into the arena of use. Under this view, strictly speaking, languages are evolving not to become more functional, but to ensure that their various features survive the repeated process of transmission to future generations. Of course, often be-

coming more functional in some sense is one way for a language to help ensure this.

Kirby has published a number of simulations of linguistic selection, in which simulated populations of agents learn artificial languages and transmit these languages down through successive generations of agents. Simulations of this sort allow us to investigate how languages can be expected to change in response to the various selection pressures that filter utterances entering the arena of use. Examples of these simulations can be found in Kirby and Hurford (1997); Kirby (2000), among many others. Future work could produce simulations of this type in which the utterances appearing in the arena of use are filtered either according to (1) their ability to be accurately perceived under noisy conditions or (2) their ability to be parsed with less than some threshold quantity of effort, where parsing effort is determined by a resource allocation theory of processing difficulty. In these simulations, utterances should make it into the arena of use in proportion to the UID functionality of their word order. If the simulations are initialised so that the agents initially speak primarily in SOV, these simulations would represent an instantiation of the explanation developed herein.

In addition to demonstrating that the problem of linkage can be solved for the explanation, this would also permit a better assessment of the explanation's fit to data. For one thing, in these simulations there would be no need to arbitrarily choose a particular global measure of UID functionality, such as mean deviation score: the details of the simulated process of linguistic selection would implifictly define an appropriate global measure, and this is a more principled measured to use than any of those considered so far. Furthermore, this sort of simulation allows a more fine-grained test of how well the explanation matches the data. These simulations make more detailed predictions than the general prediction I have made so far that languages will drift from SOV to VSO and SVO word order: they predict the relative speeds at which these changes will occur, so that we can use them to ask questions such as "by the time SOV languages have dropped in frequency to account for about 0.45 of languages, what proportion of languages should we expect to have VSO and SVO word orders?".

## 12.6 Summary

In this thesis I have motivated the need to replace pre-existing functional explanations of the different cross-linguistic frequencies of basic word orders with new explanations that are in agreement with both diachronic and synchronic evidence. I have developed one candidate for such an explanation, composed of two parts.

One part focuses on properties of non-linguistic representations and computational processes and how these properties could potentially influence word

order in improvised communication and simliar tasks. I suggest that systematic differences in the accessibility of the representations of agents, patients and actions in a compositional representation system could explain a strong dominance of SOV word order in ancient languages, as well as several other results. The SOV dominance in ancient languages may also explain the fact that this word order is the most dominant today.

The other part of the candidate explanation focuses on how different basic word orders result in different distributions of information throughout a sentence. The Uniform Information Density hypothesis, which has found a wide body of experimental support in recent years, imposes a functionality ranking on word orders, according to how well they represent efficient trade offs between noise resistance and time efficiency. This functionality ranking may account for the diachronic trend for change away from SOV to VSO and SVO, explaining why these word orders are the other two commonly occuring orders today.

In addition to being consistent with syncrhonic and diachronic data, the explanation advanced in this thesis has other advantages over previous explanations. Both of its parts have clear cognitive groundings and depend upon objectively defined quantities which can be precisely measured and verified. Many of the assumptions on which it is based, such as the existence of a compositional "language of thought" and the fact that human languages resembled information theoretically optimal codes in some respects, were posited independently. Finally, the explanation is not limited in its applicability to basic word order. It is straightforward to hypothesise an extension of the explanation which applies to word order typology in general, and some preliminary exploration of this hypothesis has been promising. The ideas developed in this thesis suggest a wide range of future work, of both a theoretical and experimental nature, and as such have the potential to be of considerable interest to the fields of psychology and linguistics alike.

# References


Ambrose, S. H. (1998). Late Pleistocene human population botlenecks, volcanic winter, and differentiation of modern humans. *Journal of Human Evolution*, *34*, 623–651.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates.

Aylett, M. (1999). Stochastic Suprasegmentals: Relationships between Redundancy, Prosodic Structure and Syllabic Duration. In *Proceedings of the XIVth International Congress of Phonetic Sciences.*

Aylett, M., & Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*, 31–56.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bell, A., Jurafsky, D., Lussier, E. Fosler, Girand, C., Gregory, M., & Gildea, D. (2003). Effects of disfluences, predicatbility, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America*, *113*, 1001–1024.

Bloom, L. (1973). *One word at a time: the use of single word utterances before syntax*. Mouton.

Bod, R., Hay, J., & Jannedy, S. (Eds.). (2003). *Probabilistic Linguistics*. MIT Press.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Campbell, L. (2001). Beyond the Comparative Method. In B. J. Blake, K. Burridge, & J. Taylor (Eds.), *Historical linguistics 2001: 15th international conference on historical linguistics.*

Chater, N., Reali, F., & Christiansen, M. H. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1015–1020.

Christiansen, M. H., & Chater, N. (2008). Language as shaped by the brain. *Behavioural and Brain Sciences*, *31*, 489–558.

Comrie, B. (1981). *Language universals and linguistic typologoy*. Basil Blackwell Publisher Limited.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

Diehl, L. (1975). Space Case: Some principles and their implications concerning linear order in natural languages. *Working Papers of the Summer Institute of Linguistics, University of North Dakota Session 19*, 93–150.

Donders, F. C. (1969). On the speed of mental processes. *Acta Psychchgica*, *30*, 412–431.

Dryer, M. S. (1989). Large Linguistic Areas and Language Sampling. *Studies in Language*, *13*, 257–292.

Dryer, M. S. (1991). SVO Languages and the OV/VO Typology. *Journal of Linguistics*, *27*, 443–482.

Dryer, M. S. (1992). The Greenbergian Word Order Correlations. *Language*, *68*, 81–138.

Dryer, M. S. (2008). Order of Subject, Object and Verb. In M. Haspelmath, M. S. Dryer, D. Gil, & B. Comrie (Eds.), *The World Atlast of Language Structures Online.* Munich: Max Planck Digital Library. Available from `http://wals.info/feature/81`

Dunn, M., Greenhill, S. J., Levinson, S. C., & Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, *473*, 79–82.

Erlich, H. A., Bergström, T. F., Stoneking, M., & Gyllensten, U. (1996). HLA Sequence Polymorphism and the Origins of Humans. *Science*, *274*, 1552–1553.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive sience. *Behavioural and Brain Sciences*, *32*, 429-492.

Feinstein, A. (1954). A New basic theorem of information theory. *IEEE Transactions on Information Theory*, *4*, 2–22.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*.

Fodor, J. A. (1975). *The Language of Thought*. Harvard University Press.

Fodor, J. A. (1983). *Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Fodor, J. A. (1998). *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. MIT Press.

Fodor, J. A. (2008). *LOT2: The Language of Thought Revisisted*. Clarendon Press.

Frank, A. F., & Jaeger, T. F. (2008). Speaking Rationally: Uniform Information Density as an Optimal Strategy for Language Production. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 933–938).

Gallager, R. G. (1968). *Information theory and reliable communication*. Wiley.

Gell-Mann, M., & Ruhlen, M. (In press). The origin and evolution of word

order. *Proceedings of the National Academy of Sciences of the United States of America*.

Genzel, D., & Charniak, E. (2002, July). Entropy Rate Constancy in Text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 199–206).

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, *10*, 65–72.

Givón, T. (1979). *On Understanding Grammar*. Academic Press.

Greenberg, J. H. (1963). Universals of Language. In (pp. 73–113). MIT Press.

Haigh, J., & Smith, J. M. (1972). Population size and protein variation in man. *Genetical Research*, *19*, 73–89.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Naacl '01: Second meeting of the north american chapter of the association for computational linguistics on language technologies 2001* (pp. 1–8).

Harpending, H. C., Sherry, S. T., Rogers, A. R., & Stoneking, M. (1993). The Genetic Structure of Ancient Human Populations. *Current Anthropology*, *34*, 483–496.

Harris, A. C., & Campbell, L. (1995). *Historical syntax is cross-linguistic perspective*. Cambridge University Press.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*, 1569–1579.

Hawkins, J. (1983). *Word Order Universals*. Academic Press.

Hawkins, J. A. (1990). A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, *21*, 223–261.

Hurford, J. (1990). Logical Issues in Language Acquisition. In I. Roca (Ed.), (pp. 85–136). Foris Publications.

Hyman, L. M. (1984). Explanations for Language Universals. In B. Butterworth, B. Comrie, & D. Östen (Eds.), (pp. 67–85). Moutin.

Jackendoff, R. (1992). *Languages of the Mind: Essays on Mental Representation*. MIT Press.

Jackendoff, R. (2007). *Language, Conscousness, Culture*. MIT Press.

Jaeger, T. F. (2006). *Redundancy and syntactic reduction in spontaneous speech*. Unpublished doctoral dissertation, Stanford University.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Johnson, M., Demuth, K., Frank, M., & Jones, B. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems 23*.

Keller, F. (2004). The Entropy Rate Principle as a Predictor of Processing Effort: An Evaluation against Eye-tracking Data. In *Proceedings of the*

*conference on empirical methods in natural language processing* (pp. 317–324).

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). *Discovering latent classes in relational data* (AI Memo No. 2004-019). Massachusetts Institute of Technology - Computer Science and Artificial Intelligence Laboratory.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*

Kirby, S. (1999). *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford University Press.

Kirby, S. (2000). The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form. In C. Knight (Ed.), (pp. 303–323). Cambridge University Press.

Kirby, S., & Hurford, J. (1997). Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands & I. Harvey (Eds.), *ECAL97* (pp. 493–502). MIT Press.

Krauss, M. E. (2007). Keynote – Mass language extinction and documentation: the race against time. In O. Miyaoka, O. Sakiyama, & M. E. Krauss (Eds.), (pp. 3–24). Oxford University Press.

Krupa, V. (1982). Syntactic Typology and Linearization. *Language*, *58*(3), 639–645.

Langus, A., & Nespor, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology*, *60*(4), 291–318.

Levy, R. (2005). *Probabilistic Models of Word Order and Syntactic Discontinuity*. PhD thesis, Stanford University.

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *5th international conference on language resources and evaluation.*

Levy, R., & Daumé, H. I. (2011). Computational methods are invaluable for typology, but the models must match the questions: Commentary on Dunn et al. (2011). *Linguistic Typology*. (To appear)

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems* (pp. 849–856).

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the World (online version)* (16th ed.). SIL International. Available from `http://www.ethnologue.com/`

Li, C. N. (Ed.). (1977). *Mechanisms of Syntactic Change*. University of Texas Press.

Light, T. (1979). Word order and word order change in Mandarin. *Journal of Chinese Linguistics*, *7*, 149–180.

Lupyan, M. H., G. & Christiansen. (2002). Case, Word Order and Language Learnability: Insights from Connectionist Modeling. In *Proceedings of*

*the 24th annual conference of the cognitive science society* (pp. 596–601). Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B. (2000). *The CHILDES project : Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Mallinson, G., & Blake, B. J. (1981). *Language Typology*. North-Holland Publishing Company.

Manning, A. D., & Parker, F. (1989). The SOV > ... > OSV Frequency Hierarchy. *Languge Sciences*, *11*(1), 43–65.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Mayr, E. (1942). *Systematics and the Origin of Species*. Columbia University Press.

Meadow, S. Goldin, & Mylander, C. (1983). Gestural communication in deaf children: The non-effects of parental input on early language development. *Science*, *221*, 372–374.

Meadow, S. Goldin, & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *91*, 279–281.

Meadow, S. Goldin, So, W. Chee, Ozyurek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 9613–9168.

Miller, B., Hemmer, P., Steyvers, M., & Lee, M. (2009). The wisdom of crowds in rank ordering problems. In A. Howesa, D. Peebles, & R. Cooper (Eds.), *9th International Conference on Cognitive Modeling.*

Miller, G. A., & Chomsky, N. (1963). Handbook of Mathematical Psychology, volume II. In D. R. Luce, R. R. Bush, & E. Galanter (Eds.), (chap. Finitary models of language users). John Wiley.

Nettle, D. (1999). *Linguistic Diversity*. Oxford University Press.

Newmeyer, F. J. (1992). Iconicity and Generative Grammar. *Language*, *68*, 756–796.

Newmeyer, F. J. (1998). *Language Form and Language Function*. MIT Press.

Newmeyer, F. J. (2000). On the Reconstruction of 'Proto-World' Word Order. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.), *The evolutionary emergence of language* (pp. 372–388). Cambridge University Press.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the Stimulus? A rational approach. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 663–668).

Petraglia, M., Korisettar, R., Boivin, N., Clarkson, C., Ditchfield, P., Jones, S., et al. (2007). Middle Paleolitich Assemblages from the Indian Subcontinent Before and After the Toba Super-Eruption. *Science*, *317*, 114–116.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America*.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The Communicative Lexicon Hypothesis. In *Proceedings of the 31th Annual Conference of the Cognitive Science Society.*

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioural and Brain Sciences*, *13*, 707–784.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*, 130–137.

Ramscar, M., & Futrell, R. (2011, July). The predictive function of prenominal adjectives. In *Workshop on information-theoretic approaches to linguistics.*

Ratcliff, R. (1993). Methods for Dealing With Reaction Time Outliers. *Psychological Bulletin*, *114*, 510–532.

Regier, T., Kay, P., Gilbert, A., & Ivry, R. (2010). Words and the Mind: How Words Capture Human Experience. In B. Malt & P. Wolff (Eds.), (pp. 165–182). Oxford University Press.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 1436–1441.

Rogers, A. R., & Harpending, H. (1992). Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences. *Molecular Biology and Evolution*, *9*, 552–569.

Rogers, A. R., & Jorde, L. B. (1995). Genetic Evidence on Modern Human Origins. *Human Biology*, *67*, 1–36.

Ruhlen, M. (1975). *A Guide to the Languages of the World.*

Ruhlen, M. (1994). *On the origin of languages : studies in linguistic taxonomy.* Stanford University Press.

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 2661–2665.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Shepard, R., & Metzler, J. (1971). Mental rotation of three dimensional objects. *Science*, *171*, 701–703.

Sherry, S. T., Harpending, H. C., Batzer, M. A., & Stoneking, M. (1997). Alu Evolution in Human Populations: Using the Coalescent to Estimate Effective Population Size. *Genetics*, *147*, 1977–1982.

Song Jae Jung. (1991). On Tomlin, and Manning and Parker on Basic Word Order. *Language Sciences*, *13*, 89–97.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*, 652–654.

Stowe, L. Gershkoff, & Meadow, S. Goldin. (2002). Is there a natural order for expressing semantic relations? *Cognitive Psychology*, *45*, 375–412.

Sun, C.-F., & Givón, T. (1985). On the so-called SOV word order in Mandarin Chinese: a quantified text study and its implications. *Language*, *61*,

329–351.

Takahata, N., Satta, Y., & Klein, J. (1995). Divergence time and population size in the lineage leading to modern humans. *Theoretical Population Biology*, *48*, 198–221.

Tomlin, R. S. (1986). *Basic Word Order: Functional Principles*. Croom Helm.

Tomlin, R. S., & Kellog, W. A. (1986). *Theme and attention orientation in procedural discourse.* (Unpublished paper. Department of Linguistics and the Cognitive Science Program, University of Oregon and IBM T.J. Watson Research Center)

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge University Press.

Vennemann, T. (1973). Explanation in Syntax. *Syntax and Semantics*, *2*, 1–50.

Voegelin, C. F., & Voegelin, F. M. (1977). *Classification and index of the world's languages*. Elsevier.

Wang, W. S.-Y., & Minnet, J. W. (2005). The invasion of language: emergence, change and death. *Trends in Ecology and Evolution*, *20*, 263–269.