

THE ROLE OF PROCESSING COMPLEXITY IN WORD ORDER
VARIATION AND CHANGE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF LINGUISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Harry Tily
August 2010

© 2010 by Harry Joel Tily. All Rights Reserved.

Re-distributed by Stanford University under license with the author.



This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.

<http://creativecommons.org/licenses/by-nc/3.0/us/>

This dissertation is online at: <http://purl.stanford.edu/sr587rm2997>

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Thomas Wasow, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Arto Anttila

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Daniel Jurafsky

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Roger Levy

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumpert, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

All normal humans have the same basic cognitive capacity for language. Nevertheless, the world's languages differ in the kind and number of grammatical options they give their speakers to express themselves with. Sometimes, a language's grammatical constructions may differ in how easy they are for comprehenders to process or how readily speakers will choose them. It has been observed that languages which allow more difficult constructions also tend to allow easier ones, and when a language only allows one option, it tends to allow the easiest to process (Hawkins, 1994, 2004). This correlation is intuitive: languages tend to give their speakers options that they find easy to use. However, the causal process that underlies it is not well understood. How did the world's languages come to have this convenient property? In this dissertation, I discuss a family of evolutionary models of language change in which processing-efficient variants tend to be selected more frequently, and hence over time have the potential to displace less efficient variants, pushing them out of the language. I begin by showing that a psycholinguistic theory, *dependency length minimization* (see e.g. Gibson, 2000, i.a.), accounts for word ordering preferences in data taken from Old and Middle English just as it does in Present Day English. I then discuss computer simulations of a model of language change which implements this bias, predicting observed word order changes in English. Finally, I present experimental studies of online comprehension in Japanese which not only display evidence for the dependency length bias, but also suggest that comprehenders encode it as part of their knowledge about language, using it to help understand the sentences they receive from their peers.

To my parents, Sally and Rob

Acknowledgments

A dissertation is a lot like a tumor. Mine was first diagnosed by my adviser, Tom, who pointed out that some of the topics we had been chatting about during our meetings were beginning to sound dangerously like a dissertation project. Of course, by that point, it had probably been taking form for months or years unnoticed, thanks to earlier exposure to ideas and influences from many others around the department and elsewhere. But it was really only in the last year or so that I recognized how much this dissertation project has grown — seemingly of its own accord — to the point that it has shaped the way I think about almost all aspects of language. Now that I have a final document in front of me, excised and ready to submit, I can much more easily recognize those influences and see how they have grown together. I have many people to thank for my tumor.

First and foremost is Tom Wasow, an adviser for whom the word “adviser” isn’t an adequate job title. I’ve received insight, aid, and encouragement from Tom, during the writing of this dissertation and before, that went far beyond anything you could call “advice”. I’m honored to have been one of Tom’s students. I’ve also benefited enormously from the input of the other members of my dissertation committee. Roger Levy was originally lured into coming on board to help with a research project that I ended up not including in the dissertation, but instead he managed to give insights and guidance right across the very broad range of topics I foolishly tried to cram into this document. Elizabeth Traugott, who unfortunately couldn’t be around during the final revisions and defense, humored some of my strange ideas about language change with interest and open mindedness, while always being able to point me in the right direction when I began to veer into the untenable or simply wrong. I regret

only beginning to talk to Arto Anttila quite late in the process, since I've found the ideas I gained by doing so eye-opening. Finally, I probably enjoyed my meetings with Dan Jurafsky more than any other academic conversations I've had: Dan is the Brita filter of academic mentors, taking murky and unpalatable half-ideas and making them transparent and appealing almost in real-time.

The six years I spent at Stanford were the most intellectually stimulating I have experienced. The atmosphere in the Linguistics Department is scholarly, but fiercely collaborative and supportive, to the point that it sometime feels as though other students and faculty are competing with each other to help, teach, and inspire you the most. My academic path has been shaped enormously by that environment. I'm especially grateful to Joan Bresnan, whose inspiring classes and seminars turned into research projects that gave me the first taste of the empirical linguistics I would end up advocating. I was also lucky to have Florian Jaeger around for the first few years as a trailblazing *senpai*, and I think any one of our long and fruitful discussions could have worked out as an alternative dissertation project. I've been the willing recipient of his generous advice and ideas ever since. The same holds for many other fellow students here: for instance, Uriel Cohen Priva, Marie-Catherine de Marneffe, Laura Whitton, Laura Staum Casasanto and Philip Hofmeister, just to list a few with impressive-sounding last names. In fact, since so many other students have made my PhD experience so much better, I could just have cut-and-pasted a list from the department website here. But you know who you are. Thank you all!

Perhaps more than any of my friends at Stanford, I will miss my former roommates, Inbal Arnon, Lis Norcliffe, and Scott Grimm; as well as my neighbor, Neal Snider. Between cohabitation, commuting, and collaboration, I ended up sharing most of my life with you for a few wonderful years. And I'd do it all again. Some day in the near future we will get together in San Francisco, cook shakshouka, and reminisce.

Finally at Stanford, I owe a huge debt to the department's administrative staff. In particular, Melanie, Gretchen and Alyssa patiently managed to fix every one of my attempts to fail out of the program by getting most bureaucratic and department requirements entirely wrong, and no doubt averted many more similar catastrophes

behind the scenes. Thank you!

Beyond Stanford, I have been lucky enough to visit and work with some of the smartest and most interesting people I can imagine. In Paris, Luc Steels and the members of his group (particularly Remy van Trijp, Wouter Van den Broeck, and Martin Loetzsch) made me welcome in a new and beautiful city, and taught me to think about my research in a very different way. I hope that influence is apparent in this dissertation, as I think it will be in all my work hereafter. At MIT, I was amazed at the brilliant ideas that were thrown around like confetti every day in Ted Gibson's lab. The collaborations which I began there with Ted, as well as Ev Fedorenko, Steve Piantadosi, and Mike Frank have become so important to me that they are now basically what I do, and for that I'm very grateful. A much earlier influence, and perhaps the fundamental reason I followed the path I have ended up on, is that of Antje Meyer, who "adopted" me during my undergrad degree. I got my first taste of a real lab, and of any real collaborative academic environment from Antje and her lab members. Thank you!

Just as much as the academic influences, my friends and family are responsible for me having survived grad school. In San Francisco, I'd like to thank Neal and Natalie, Robert, and Fabian for the many good times. Also Melissa, Alexis and Tessa, and all former, current, and (why not) future residents of the infamous house on Church St, where I could and frequently did forget I was a grad student and remember I was a person. In Tokyo, my life has been much better for Vicky Muehleisen, Nobuko Suzuki, John Brooke, Kuki Koibuchi, and Greg Poitevin. I also want to recognize the profound influence of San Francisco itself, whose coffee shops, bars, views, and crazy people are all I can imagine anyone ever wanting to be surrounded by. Similarly, I'd like to thank Palo Alto for being so unapologetically, painfully unpleasant to live in that it forced me to move to the city.

Finally — and most of all — thank you Nathan, for your support and love. I promise I will never write another dissertation.

Contents

| | |
|---|-----------|
| Abstract | iv |
| Acknowledgments | vi |
| 1 Introduction: The order of meaningful elements | 1 |
| 1.1 Absolute and statistical universals | 3 |
| 1.2 A functional evolutionary account of typological patterns | 6 |
| 1.3 The contents of this dissertation | 12 |
| 2 Weight and word order in OE and ME | 14 |
| 2.1 Weight and word order | 14 |
| 2.2 Verb-object order in Old and Middle English | 17 |
| 2.3 Analysis | 24 |
| 2.3.1 Data preparation | 24 |
| 2.3.2 Statistical analysis | 29 |
| 2.3.3 Results | 33 |
| 2.4 Discussion | 36 |
| 3 Dependency length minimization | 40 |
| 3.1 Serialization of non-ordered meanings | 40 |
| 3.1.1 Conceptual accessibility and availability-based production . . | 42 |
| 3.1.2 Dependencies between words | 48 |
| 3.1.3 Prosodic theories | 55 |
| 3.2 Language-wide dependency lengths | 57 |

| | | |
|----------|---|------------|
| 3.2.1 | Dependency lengths in Historical English | 60 |
| 3.2.2 | Left dislocation | 62 |
| 3.2.3 | Measuring the change in dependency lengths | 63 |
| 3.2.4 | Word order freedom and morphology | 68 |
| 3.3 | Summary | 70 |
| 4 | Weight as a factor in word order change | 71 |
| 4.1 | Processing effects in word order change | 71 |
| 4.1.1 | Variation and change | 71 |
| 4.2 | Evolutionary simulations | 74 |
| 4.2.1 | Processing in simulated variation and change | 76 |
| 4.3 | A model of OV loss by processing bias | 80 |
| 4.3.1 | A multi-cue variational learner | 80 |
| 4.3.2 | A biased multi-cue variational learner | 86 |
| 4.4 | Problems with the theory | 89 |
| 4.4.1 | Is there really a universal preference for SVO? | 89 |
| 4.4.2 | Are language learners really this bad? | 95 |
| 5 | Inference driven comprehension | 97 |
| 5.1 | Word order and case cues in processing | 98 |
| 5.1.1 | Processing differences by language type | 98 |
| 5.1.2 | The utility of word order variation | 102 |
| 5.2 | Japanese word order and case | 106 |
| 5.3 | Comprehension as Bayesian inference about production choice | 111 |
| 5.3.1 | Study 1: Case and word order in comprehension | 118 |
| 5.3.2 | Study 2: Interaction of NP type with other cues | 124 |
| 5.4 | “Explaining away” of mutually informative cues | 130 |
| 5.4.1 | Study 3: Length as an information source in comprehension | 136 |
| 5.5 | Discussion | 144 |
| 6 | Discussion and extensions | 146 |
| 6.1 | Dependency minimization in English | 146 |

| | | |
|-----|---|-----|
| 6.2 | Should processing effects be treated as priors or biases? | 148 |
| 6.3 | Language evolution and the relationship between processing and gram- mar | 158 |
| 6.4 | Conclusion | 162 |

List of Tables

| | | |
|-----|---|----|
| 2.1 | Total number of clauses with each word order in dataset | 27 |
| 2.2 | Predictors removed during model comparison | 34 |
| 2.3 | Final model for VO/OV order (positive outcome is VO) | 35 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | Proportion of main clauses in each text with VO order | 29 |
| 2.2 | The estimated influence of the length and pronominality of accusative/ direct objects on order over time (solid lines); predictions for a 2 word dative/indirect object over time (dashed line). Other predictors are at baseline level. | 39 |
| 3.1 | The logically possible orders of a three-word dependency structure. . | 49 |
| 3.2 | Mean per-sentence dependency lengths from Gildea and Temperley (2007). Numbers show length ratios relative to the theoretically pos- sible minimum. | 59 |
| 3.3 | Mean total dependency length for each manuscript (raw values) . . . | 66 |
| 3.4 | Mean total dependency length for each manuscript (relative to theo- retical optimum) | 67 |
| 4.1 | Unbiased simulation, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations. | 84 |
| 4.2 | Unbiased simulation, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations. | 85 |
| 4.3 | Biased simulation: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations. | 87 |

| | | |
|-----|---|-----|
| 4.4 | Biased simulation (OV start state): Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations. | 89 |
| 5.1 | Sketch of graphical model for speaker's choice between SOV and OSV orders in Japanese. | 114 |
| 5.2 | Reading times at each region in study 1. | 122 |
| 5.3 | Reading times at embedded verb in study 1. Bars show standard error of the mean. | 123 |
| 5.4 | Reading times at each region in study 2 for direct (<i>wo</i> -marked) object conditions. | 127 |
| 5.5 | Reading times at each region in study 2 for indirect (<i>ni</i> -marked) object conditions. | 128 |
| 5.6 | Reading times at embedded verb in study 2. Bars show standard error of the mean. | 129 |
| 5.7 | Reading times at each region in study 3 for short-short conditions. . . | 140 |
| 5.8 | Reading times at each region in study 3 for long-short conditions. . . | 141 |
| 5.9 | Reading times at adverb and embedded verb in study 3. Bars show standard error of the mean. | 142 |
| 6.1 | Prior probability over the object length coefficient for weak prior (dotted, sd=4) and strong prior (solid, sd=1) simulations | 152 |
| 6.2 | MAP simulation with weak object length prior, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates. | 153 |
| 6.3 | MAP simulation with weak object length prior, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates. | 154 |

| | | |
|-----|---|-----|
| 6.4 | MAP simulation with strong object length prior, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates. | 155 |
| 6.5 | MAP simulation with strong object length prior, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates. | 156 |

Chapter 1

Introduction: The order of meaningful elements

The ordering of words and phrases is one of the most apparent ways in which languages can differ from one another, and is one of the central objects of interest for language typologists. Greenberg’s (1963) seminal publication, which fuelled much modern interest in the question of “linguistic universals”, was named *Some universals of grammar with particular reference to the order of meaningful elements*. Among the variations in order that have been studied, the relative position of the subject, object, and verb have attracted the most attention, with 15 of Greenberg’s 45 universals making mention of the relative order of those three elements specifically. Part of the reason for this focus is simple practicality: compared with many other features of grammar, linguists can agree reasonably well on a set of properties that will identify a subject, object, and verb in the diverse languages of the world, and their relative order can be straightforwardly recorded.¹ But the very fact that reasonably consistent categories of subject, verb, and object can be identified in most of the world’s languages suggests those categories themselves play some fundamental role in the way people of all language backgrounds structure their communication. Equally, the fact that

¹This is not to say that there is no disagreement over the identity of the subject and object in some languages, or even the degree to which these are universal categories (see in particular the literature on so-called Ergative-Absolutive languages, e.g. Anderson (1976)).

certain orders of those elements are far more frequent than others cross-linguistically suggests that the ordering of meaningful elements is constrained or biased in a way that all people share. Descriptive generalizations like “subjects almost always precede objects”, or “languages with basic order verb-subject-object tend to have prepositions before the noun rather than postpositions after” may reflect deeper properties of the human mind.

Superficially, it might seem remarkable that something as directly observable and unmistakable as word order could undergo change at all. While *sounds* exist in a gradient space of possible articulations, and therefore can be misheard or misarticulated with some probability, the order of words is a discrete fact and it seems unlikely that one order of words could be consistently “misheard” as another. In fact, Li (1977) calls word order change “the most drastic and complex category of syntactic change”, remarking that word order is relatively stable compared to other categories of language change, being sometimes even used as a diagnostic when determining how closely two languages are related. However, languages often give their speakers the option of choosing between one of several orders in production. Latin, for instance, has extremely free word order; Old English and Modern German often allow more orderings of the basic sentence constituents than Present Day English does. This said, even Present Day English allows its speakers a certain amount of word order freedom through the use of constructions like the following, in which the members of each pair are approximately meaning-equivalent, but the order of some of the words reversed:

- (1) a. the building’s roof
b. the roof of the building
- (2) a. the dog bit the man
b. the man was bitten by the dog
- (3) a. bring in the suspect
b. bring the suspect in

- (4) a. give some toys to the children
b. give the children some toys

This existence of this kind of variation between orders allows the potential for language change, if we see language change as a gradual replacement of one form with another (rather than merely the instant when a form is invented or lost). Labov (1982) takes this competition between forms to be the central object of interest when investigating how languages change, writing “change implies variation; change *is* variation”. In English, both members of each pair above are currently used with relatively high frequency. However, if some process were to cause speakers to prefer one of the forms over the competing way of expressing the same meaning, then over time the alternative might be lost. The question of interest in this dissertation is whether there are factors that consistently cause speakers to favor certain orders of words over the alternatives. If this is the case, we would expect to observe constructions with the favored ordering of words more frequently — even universally — across human languages.

1.1 Absolute and statistical universals

The observation that certain properties recur again and again in the world’s languages is not controversial, but there is ongoing debate about the extent to which any such patterns are universal, and about what causes them. An influential program in theoretical linguistics since the Chomskyan revolution (e.g. Chomsky, 1965, 1980) has explored the idea that many commonalities in languages arise because humans are endowed with a genetically specified “language faculty”, some set of cognitive abilities which are highly adapted for — and specific to — language. Under this account, all languages are necessarily governed by some small set of shared rules and structures: a so-called “Universal Grammar” (UG). Therefore, it should be unsurprising that all languages should display shared characteristics. In fact, Chomsky’s original version of the UG theory holds that the precise contents of this faculty are not as important as the fact that it is shared by all humans, making language learning much easier

than would be the case if languages could take any imaginable form. Accordingly, the universal features of language it gives rise to could be arbitrary and potentially quite baroque, and hence relatively easy to spot. However, despite the large volume of linguistic research that has assumed UG, there is very little agreement about exactly what cognitive faculties might be specifically adapted for language (see e.g. the debate between Chomsky et al. (2002) and Pinker and Jackendoff (2005); cf Bickerton (2009)) or about what the linguistic manifestations of those natively endowed abilities might be (see e.g. Newmeyer, 2005; Scholz and Pullum, 2006; Haspelmath, 2008b; Evans and Levinson, 2009). As a result, many linguists now reject the idea that linguistic universals reflect arbitrary constraints. Instead, they have explored the idea that the languages that exist show similarities because extant languages are among those which are in some respect the most efficient for human communication or use.

A growing number of linguists working under the assumption of universal grammar accept some variety of this proposal. In formulations like those made by Pinker and colleagues (e.g. Pinker and Bloom, 1990; Pinker and Jackendoff, 2005), it is UG itself which has evolved for communication. This means that the shared characteristics of all languages might be identifiably “efficient” for communication in some sense, or at least might have been efficient at some point in the past. Other statements deny that UG itself shows hallmarks of evolutionary selection for communicative efficiency, but instead that UG only accounts for a smaller set of “core” universal properties, while the majority of phenomena studied by typologists occur for reasons that are independent of the innate grammatical machinery. For Newmeyer (2005), the language faculty is “autonomous” in that it defines the set of possible languages without regard to what language users might actually find easy or useful to use. Nevertheless, since these “possible” languages differ in how efficiently they can be used, some are more “probable” than others, and hence occur more frequently. With slightly different argumentation, Kiparsky (2008) separates typological patterns into two classes: the “true” universals, which result from core UG constraints, and the mere “typological generalizations” which describe the end result of diachronic processes which are particularly frequent for various reasons (not necessarily due to considerations of

efficiency). Naturally, positions like those taken by Newmeyer or Kiparsky restrict the explanatory scope of UG substantially relative to the positions of Chomsky or Pinker. In either case, though, UG is taken to be genetically hard-coded and shared by all humans, so the phenomena it entails should be observed in all languages.

However, in recent years typologists have called into the question whether there are in fact any properties that are truly shared by all languages. Dryer (1998) discusses “statistical” universals, generalizations that hold for many or most languages, but not necessarily all. These contrast with the “absolute” universals predicted by the hard-and-fast constraints of traditional UG. Dryer concludes that the majority of universals for which there are empirical support are of the statistical kind: strong but not exceptionless tendencies. Dryer also rejects claims that absolute universals are methodologically preferable, or more informative about the nature of language. Modern statistical techniques allow hypothesized universals to be tested even if they are tendencies rather than absolute rules, and in practice, absolute universals need to be stated so weakly in order to include all languages that they end up telling us very little about language at all. More recently, Evans and Levinson (2009) have surveyed a set of properties that have been claimed to be shared by all the world’s languages. For all the facts they consider they find that there are exceptions. The CV (consonant-vowel) syllable structure is only *nearly* always found in languages, since at least one language has no syllable-initial consonants. Syntactic part-of-speech classes such as noun, adjective, and adverb are not found in every language, and some languages may even lack distinctions as apparently basic as noun versus verb. There is little evidence for an independently justified set of “core” semantic concepts, either: some languages lack equivalents of ideas as basic to an English speaker as *leg* or *if*. Even superficially trivial sounding universals like “all languages have vowels” are invalidated by the existence of sign languages. However, while none of these putative universals turn out to be hard and fast rules, all reflect real *tendencies* across languages. Languages can, in principle have only V and VC syllables, or can be signed rather than spoken, but in practice such languages tend not to emerge. If the picture emerging from typology is correct, there is very little that can be said to be truly universal, and therefore very little work for a universal grammar to do. In contrast, there is a rich

tradition of literature, chiefly in the functionalist tradition, which attempts to explain the tendencies for certain linguistic forms to reoccur repeatedly, doing so by appealing to their naturalness or efficiency for human communication.

1.2 A functional evolutionary account of typological patterns

Like biological changes in species, language change can be understood as an evolutionary process. In the familiar case of the evolution of biological organisms, change over generations can be broken down into an iterated generation-by-generation process, where each generation is made up of individual organisms. Analyzing this process within an evolutionary framework is possible thanks to three crucial properties. First, the organisms are *replicators*: they can reproduce, creating self-similar copies. Second, the genotype of these individuals may differ minutely from their ancestors thanks to genetic *combination* and random *mutation*. Third, each individual may have a greater or lesser chance of surviving and breeding depending on their *fitness*, a measurement of their ability to deal efficiently with the environment: to obtain food, find suitable mates, etc. In exactly the same way, language change can be broken down into iterated generation-by-generation process, except that instead of organisms, we have the linguistic knowledge of individual human language users. Since humans acquire language by listening to any copying other users' language, the individual words and constructions that make up this linguistic knowledge are replicators (Croft (2000) coins the term *lingueme*, in analogy to Dawkin's (1986) cultural replicators, *memes*). Each person's linguistic knowledge differs slightly, since it is acquired by learning from language produced by a different set of individuals (combination) and may be subject to "mutation" thanks to chance differences in the input, mislearning, or deliberate innovation. The third property, that of differential fitness, is what makes it interesting to study language change in an evolutionary framework. UG constraints are absolute — in the biological metaphor, they might correspond to limits on possible organisms, statements like "all organisms must form a contiguous

mass”, or “no organism can be smaller than the smallest physically possible cell”. On the other hand, factors that probabilistically influence which “linguemes” are more likely to propagate to the next generation correspond directly to factors that influence how fit an organism is, such as how efficiently it can find food or defend itself. Measures of how likely a certain linguistic form is to be used therefore translate directly into the probability of its survival over generations, and consequently should be able to predict the kind of statistical universals described by Dryer (1998) or Evans and Levinson (2009). This view of language change suggests that languages are “phenomena of the third kind” in the sense of Keller (1994); macro-level patterns that have evolved over time due to the interaction of individuals who share common goals and abilities (see also Hopper, 1987; Croft, 2006). Linguists working in several subfields have proposed models of various kinds of language change as an evolutionary process (see e.g. Briscoe, 1998; Blevins, 2003; Wedel, 2003; Croft, 2006; Carstairs-McCarthy, 2007; Clark et al., 2008; Rosenbach, 2008; papers in Christiansen and Kirby, 2003; chapter 4 here).

Within the functionalist typological literature, general cognitive principles are posited to play a key role in language typology and change. For instance, a great deal of research has been done on the topic of *iconicity*, the observation that “the structure of language reflects in some way the structure of experience” Croft (2002, p102).² Naturally, the human cognitive apparatus determines how we see the world, and so language reflects this apparatus, preferring to encode properties that the mind finds salient, encoding properties that cluster perceptually as tighter linguistic units (e.g. Givón’s (2001, p64) proximity principle), and representing more complex mental

²Here, experience is understood as experience with the non-linguistic world, rather than specifically linguistic experience. Theories based on specifically linguistic experience are also influential in psycholinguistics, and will be discussed in chapter 5. For the purposes of predicting statistical universals in typology, the influence of purely linguistic experience is a difficult quantity to evaluate, since using the frequency of linguistic forms within particular languages to predict the frequency of those forms crosslinguistically might lead to circularity. While there is a large literature on the role of frequency and probability in language change and typology (e.g. Bybee and Thompson, 1997; Bybee and Hopper, 2001; Diessel, 2007; Haspelmath, 2008a, i.a.), it tends to be concerned with the reduction, unmarkedness and grammaticalization of frequent and probable forms, without addressing the question of whether the frequency of linguistic forms can be decoupled from the frequency of the meanings they express.

concepts with more complex linguistic structures. Although such claims are clearly based in an intuitive understanding of human psychology, they are typically made at a relatively abstract level, which has perhaps limited the amount of attention and investigation they have received by laboratory psychologists. Nevertheless, a certain amount of empirical work has tested whether these posited functionalist principles have observable behavioral correlates. For instance, Sridhar (1989) proposes a list of concrete hypotheses which specify which kinds of properties the mind finds salient (e.g. continuums of temporality, activity vs stativity, and spatial organization) and shows that when required to describe visually presented scenes, speakers of multiple languages preferentially verbalize the most salient properties (see also Osgood, 1971). Croft (2010) has extended this methodology, linking the types of situations and entities that tend to be verbalized to the kinds of specific grammatical constructions that tend to be used to encode them crosslinguistically. This research program makes an explicit link between general cognitive processes and statistical universals. It illustrates that certain cognitive principles bias the choices that people make about what to verbalize first (for instance, choosing to mention human, agentive, and physically close referents early), and explores the result that those biases may have on the shape of languages' grammars after sufficient time (for instance, a grammatical requirement to place agents or grammatical subjects early in the sentence).

The cognitive constraints explored in the functionalist typological literature are claimed to correlate with typological facts, but the causal mechanism that links them is typically not discussed. At a fairly high level of abstraction, though, it's easy to see how an evolutionary conception of language change could give a suitable framework. The "fitness" of a linguistic form is partially predicted by the extent to which they adhere to cognitive biases. In a language which allows both subject-object and object-subject orders, speakers may choose to use the subject-object order more often because this order adheres to a cognitive bias towards mentioning more salient referents first. This increases the subject-object order's "fitness", because more productions means a higher probability of being acquired and used by the next generation of language users. Over generations, then, the grammar might evolve to favor

subject-object orders, perhaps even pushing the object-subject order out of the grammar altogether. Of course, this account relies crucially on the higher frequency of a certain form in production leading to preferential learning in at least some situations. In the language acquisition literature, there is ample evidence that input frequency does play a central role in acquisition (e.g. Tomasello, 2003; Lieven et al., 2003): more frequent lexical items and instances of constructions are learned and fully mastered earlier, with any generalization to lower frequency instances of the same constructions taking much more time. In fact, language change may occur even within the lifespan of individual speakers if they re-organize their usage patterns in response to previous communicative episodes. Jäger and Rosenbach (2008) argue that priming or alignment of forms of the sort observed in the short term (e.g. Bock, 1986; Bock and Griffin, 2000; Pickering and Garrod, 2004) may be responsible for longer term language change: repeatedly producing or hearing a certain word or structure makes a speaker more likely to use it again.³

Unlike the function typological literature, the majority of psychological and psycholinguistic research focuses on much more fine-grained phenomena than general cognitive principles like differential salience. Nevertheless, there is no reason why psycholinguistic theories which have been developed over the last few decades to account for behavioral findings observed in the lab could not be used to derive predictions about the evolution of linguistic constructions. Psycholinguistic theories can be used to derive measures of *processing complexity*, a notion of difficulty or dispreference which is reflected in behavioral measures like reading times/accuracy in comprehension, or choice between constructions in production. In chapter 5, I discuss some psycholinguistic theories that are relevant to word order specifically. For a summary of psycholinguistic theories that might prove relevant to our understanding of language change and typology more generally, see Jaeger and Tily (in press); also Hawkins (2007).

Until very recently, comparatively little work has attempted to incorporate modern psycholinguistic results into a theory of language universals. The largest contribution

³The central point made in Jäger and Rosenbach’s article is that the priming theory predicts so-called “unidirectional” language change, a hallmark of grammaticalization, but that is not directly relevant to the discussion here.

is probably that of Jack Hawkins, whose research relates the distribution of forms appearing across languages to their processing complexity. This relationship is stated clearly in Hawkins' *Performance-Grammar Correspondence Hypothesis* (PGCH):

Grammars have conventionalized syntactic structures in proportion to their degree of preference in performance, as evidenced by patterns of selection in corpora and by ease of processing in psycholinguistic experiments. (Hawkins 2004, p3)

The PGCH specifically predicts that where some language allows two forms but users find one harder to process than the other, other languages will only allow the easier of the two. For example, Keenan and Comrie (1977) note that all languages seem to have some “cutoff” on the hierarchy of grammatical relations in (5) beyond which extractions out of a relative clause are not grammatical. Examples for each position are given for English in (6), which allows all four.

(5) subject > direct object > indirect object/oblique > genitive

- (6) a. The reporter that interviewed the senator
 b. The senator that the reporter interviewed
 c. The senator that the reporter gave the gift to
 d. The reporter whose article the senator read

Unlike English, Turkish disallows genitive relative clauses; Greek also IO/obliques; and Arabic allows only subject extraction. In the disallowed cases in these languages, direct translations of the sentences in (6) are often possible if a resumptive pronoun is added (e.g. “... *that the reporter gave the gift to him*”). Hawkins (1999:258; 2004:180-186) explains this in processing terms: in languages which allow extraction from different positions like English, comprehenders find object-extracted relatives (ORCs; 6b) harder than subject-extracted relatives (SRCs; 6a), as evidenced by reading-time studies, comprehension accuracy, and ERP studies of brain activity,

in multiple languages (e.g. Wanner and Maratsos, 1978; Ford, 1983; King and Just, 1991). Although the comprehension difficulty of other positions has not been tested in as much depth, Keenan and S. Hawkins (1987) did find that comprehension question accuracy in English diminishes monotonically further down the hierarchy. In production too, there is evidence that speakers avoid forms lower on the hierarchy when possible. Keenan (1975) found the frequency of relative clauses of each type in a written corpus to decrease down the hierarchy, and more recent corpus studies have been largely in accord. Roland et al. (2007) find SRCs more common than ORCs in all the written corpora they examine, though not all their spoken corpora. Gennari and MacDonald (2008) find that in a constrained completion task, subjects frequently express meanings like that in (6b) with a passive (i.e. “*the senator that was interviewed by the reporter*”), avoiding the ORC construction, and with certain verb and object types, ORCs are only used very rarely. The PGCH suggests that while languages may differ with respect to the amount of processing complexity they will tolerate, any language allowing a certain construction will also allow all those that are less complex, and any language disallowing that construction will disallow all those that are more complex.

Like the functionalist literature discussed above, the PGCH notes a *correlation* between processing and grammar without specifying a causal relationship between the two. Again, evolution is a natural framework in which to model the causal processes that lead to this correlation. Speakers are more likely to use linguistic forms that they find easy to process, and therefore over generations languages may evolve towards the more processing-efficient. In this way, forms with higher processing complexity could be gradually pushed out from the language over generations. Of course, the hypothesis that languages evolve towards the optimal opens up a Pandora’s box of questions: for instance, why do languages allow inefficient linguistic forms at all, and indeed why have they not all converged on a single solution? It is likely that competing pressures determine which forms are most efficient in any particular situation, and that there is no single “best” solution which is optimal in all ways. A definitive answer will only become apparent with further crosslinguistic research on the interacting processing factors that determine online language use at the individual level, and

with diachronic and typological work testing the predictions of the resulting theories against documented trajectories of change and distributions of linguistic forms across the world’s languages. This dissertation is intended as one small contribution in that direction.

1.3 The contents of this dissertation

In chapter 2, I discuss the widely studied phenomenon of *End Weight* – the tendency for languages to place “heavier” or longer phrases at the end of their sentences. This tendency has been observed cross linguistically, and here I show evidence for it in diachronic data as well, with corpus work showing that the length of object phrases in Old and Middle English affected the placement of the object. English was changing in its word order and in the amount of word order flexibility it gave its users, and the different possible positions in which the object could appear reflected a variation between different basic word orders, chiefly, SOV and SVO.

In chapter 3, I consider some psycholinguistic theories that might account for the influence of phrase weight on placement within the sentence, concluding that so-called “dependency minimization” theories offer plausible explanation. Language users appear to prefer to place words that are dependent on each other for interpretation close together, perhaps to allow for structures to be processed efficiently online without having to continually recall words that were passed a long time ago in the sentence. This would result in a preference for SVO, since there both the subject and object are adjacent to the verb they rely on for interpretation. If this dependency minimization pressure was actually able to bias the word order changes that were ongoing in Old and Middle English, we might expect that the lengths between dependent words would shorten over time. Using corpus data, I show that this was the case.

In chapter 4, I propose some ways in which a psycholinguistic bias like the dependency minimization theory might causally influence word order in a language which is undergoing change. Language learners manage to reproduce the constructions their language gives them with approximately the same frequency that they hear them

used in their environment. If, however, processing biases affect the frequency with which each construction is used, speakers in their environment will use the constructions favored by the bias more frequently than they otherwise would, which could lead learners to overestimate the frequency with which they themselves should use it. Essentially, learners would need to “filter out” any disruption in the frequency with which they hear each construction that might be due to factors like processing bias. If they do not, then over generations the distribution over linguistic constructions would gradually shift to favor those favored by the bias.

In chapter 5, I question whether or not learners can determine when a certain word order has been chosen because of a processing bias, like the pressure to place heavier material at the end of sentences in English. I investigate Japanese, a language which allows SOV and OSV order. Japanese displays the reverse weight tendencies to English, favoring OSV more commonly when the object is long. In a number of online experiments, I show that adult comprehenders are implicitly aware of this bias, and in fact seem to use it to predict the upcoming structure online. I present a novel model of language comprehension in which comprehension is understood as rational Bayesian inference about production, incorporating both the speaker’s intent and other biases in making its predictions. This model accounts for the Japanese facts. However, it calls into question the assumption underlying the models of change discussed in chapter 4 which relies on learners failing to account for processing biases.

In chapter 6, I summarize the major findings and discuss the conclusions that can be drawn from this work, as well as considering some potential extensions.

Chapter 2

Weight and word order in Old and Middle English

In this chapter I describe the major case study of this dissertation, investigating the role of constituent length/weight as a predictor of word order in the history of English.

2.1 Weight and word order

A century ago, Behaghel (1909, 1932) observed what he called *Das Gesetz der wachsenden Glieder* (“the Law of Increasing Elements”), a striking tendency for speakers to place their words and phrases in order of length where possible, starting with the shortest and increasing thereafter. Behaghel took his evidence from examples in texts written in several languages, and also from a simple experiment anticipating the “constrained production” task used more widely in recent psycholinguistics. Participants were given written phrases in a random order, and asked to construct a sentence from them. He found that where the resulting sentences had a co-ordination of two phrases, the shorter would precede the longer: *Gold und edles Geschmeide* rather than *edles Geschmeide und Geld* (“money and precious jewels”). Behaghel’s explanation for these patterns would be called a “processing” account today: his intuition was that longer phrases were more difficult for the speaker, and leaving them until later would buy valuable time to process them.

Man wird nicht nur die länger dauernde Arbeit auf den Zeitraum verlegen, wo man den Abschluß leichter hinausschieben kann; man wird auch, wenn man sich Zeit lassen kann, die Arbeit gründlicher tun, mehr ins Einzelne gehen, oder, sprachlich ausgedrückt: man wird nicht nur für den umfangreicheren Ausdruck die spätere Stelle wählen, sondern auch für die spätere Stelle den umfangreicheren Ausdruck sich zubereiten.

—Behaghel, 1909, p138

People do not merely put off more time-consuming work to a later period because they can avoid doing it sooner. By leaving themselves enough time, they can do the work more thoroughly and go into more detail. Linguistically speaking, they are not merely postponing the production of the most elaborate expression, they are preparing that expression until the latest possible position.

This intuition is quite powerful, and underlies psycholinguistic theories that are influential today. These will be discussed in Chapter 3.

Similar patterns are noted in Quirk et al.'s (1972) grammar of English: English phrases are often “postponed” until towards the end of the sentence when they are long. This observation seems to hold for both entire arguments of a verb such as the “heavy NP shift” in (7a) and their modifiers, as in the “complement clause extraposition” of (7b).

- (7) a. Ellen told to Ross *the rumour that he was secretly engaged to the Marchioness*
 b. A rumour circulated widely *that he was secretly engaged to the Marchioness*

Quirk and colleagues describe this influence with a general “Principle of End Weight”: phrases are presented in order of increasing weight, where “weight” is understood as some measure of complexity or length. Even though we often think of English as a “fixed word order” language, the possibility of word order variations like these means that weight in fact plays a major role in determining speakers’ choices. Across a range of constructions in English, Wasow’s corpus-based research (Wasow, 1997, 2002) finds a great deal of support for this principle. This is true in constructions

where word order is relatively free, such as the order of object and particle in (8): speakers strongly prefer to place longer objects after the short particle, making (8c) much more frequent than (8d).

- (8) a. the detective brought in [the suspect]
 b. the detective brought [the suspect] in
 c. the detective brought in [the man who was accused of having stolen the automobile]
 d. the detective brought [the man who was accused of having stolen the automobile] in

It is also true in constructions where one order is much more common than the other: the canonical order (9a,b) is often flouted if doing so preserves end weight (9c) but rarely otherwise, making (9d) rare or marginal.

- (9) a. John took [his friends] [into account]
 b. John took [only his own personal acquaintances] [into account]
 c. John took [into account] [only his own personal acquaintances]
 d. John took [into account] [his friends]

Although there have been suggestions that it is not merely length in words that influences placement but some more sophisticated measure of phrasal weight or complexity (see e.g. Chomsky, 1975, p477), empirical investigations have found that the different proposed measures are correlated to the point that they are interchangeable for any practical purpose (Wasow, 1997, 2002; Szmrecsányi, 2004).

In addition to the corpus studies mentioned above, phrase lengths have been shown in lab experiments to be a good predictor of which of two alternants a speaker will choose in production. Stallings et al. (1998) showed that when asked to reproduce an NP-PP sentence similar to those in (9) that they had previously heard, speakers were more likely to shift the NP to the end of the sentence if it was long. Arnold et al.

(2000) showed a similar effect for both heavy NP shift sentences and the ditransitive alternation in corpus data. In a lab experiment also reported in that article, participants interacted with each other in a cooperative task which entailed them producing large numbers of dative sentences like “*give it the small green crayon*”. The results show that again, participants prefer to place the shorter constituent closer to the verb (avoiding, for example, “*give the small green crayon to it*”).

Across multiple constructions in multiple languages, then, there is a noticeable tendency for longer phrases to appear later in the sentence than shorter ones. This tendency arguably reflects some kind of processing bias due to the cognitive apparatus used by human language users, and I will discuss this point in detail in the next chapter. In the rest of this chapter, I will test whether a similar effect obtains in the textual record of earlier stages of English.

2.2 Verb-object order in Old and Middle English

English has received a great deal of attention in historical linguistics. No doubt this is partly due to its relationship to Present Day English, the most studied modern language, but it has also received attention thanks to the wide ranging fundamental changes it underwent over the course of a few hundred years, and because a relatively large number of original documents survive. A distinction is usually made between Old English (or Anglo-Saxon), as spoken up until around 1100, and Middle English, spoken from around that time until the late 1400s. Around the time of the Battle of Hastings (1066), English was undergoing substantial grammatical changes, including the loss of case and gender inflections on nouns, significant vocabulary changes, and restructuring of the article system, that justifies a post-hoc separation into Old and Middle English at around this time. However, the Norman invasion itself was probably not responsible for many of these changes, as there were very few Norman speaking immigrants outside of the court, and for most English speakers life probably continued relatively uninterrupted despite the change of rulership. The Norman invasion did, unfortunately, result in English being replaced by Norman (and Latin) as the prestige written language, resulting in an enormous decrease in the number of texts composed

in English for some time after the invasion. No court documents were written in Middle English until 1258, and English did not become the standard written language of government again until the end of the 14th century.

The generative literature has been particularly concerned with the syntactic analysis of Old English word order, and even now there is some controversy about how best to describe the order facts. Much like modern German and Dutch, Old English (OE) appears to have a “verb-second” (V2) order in main clauses, such that exactly one constituent appears before the verb (van Kemenade, 1987; Pintzuk, 1999; Fischer et al., 2000). This tendency continues well into the Middle English (ME) period, where the variety of preverbal phrase types gradually decreases until the only constituent to frequently appear before the noun is the subject. However, unlike modern German, even in the earliest recorded OE texts, the verb does not always appear in second position. Additionally, subordinate clauses display a great deal of variation: while German and Dutch place the verb last in a subordinate clause, OE has VO order in subordinates roughly as often as it has OV (Koopman, 2005). OE, and to a slightly lesser extent, ME, allowed clauses to be produced in a multitude of different word orders. In particular, I will focus on the relative order of the verb and object: throughout OE and ME, both VO and OV orders are relatively common. Over the course of several hundred years, however, OV was gradually replaced with the VO order of the modern language.¹

The majority of linguistic research on English word order change assumes that there is a shift over time between two discrete and independent grammars, which occurs because a certain grammar may not generate only VO or only OV sentences, but potentially both, by different means. For instance, a primarily OV grammar may be able to generate VO sentences by optional rules like “object extraposition” or “verb seconding”. This means that certain actual English sentences can be “ambiguous” between two such grammars. In contrast, the study reported here looks directly at the

¹One broadly accepted view within historical syntax is that there was a relatively fast *parametric* change between two “underlying” orders (Lightfoot, 1991), which appears gradual due to subsequent syntactic operations changing the order of words in the string actually produced. However, the evidence for rapid change in even this underlying order has been called into question (see Pintzuk, 1999; Allen, 2000; Kroch and Taylor, 2000b). In any case, for reasons described next I only consider the “surface” order of the phrases, which underwent gradual change.

order in which the verb and object appear, without considering any more abstract representation which might underlie the order. To avoid confusion, I will briefly outline the justification for this and the way in which my terminology will differ from some previous literature.

In a derivational theory of syntax, clauses are “base generated” with some underlying structure: for the verb and object, the underlying order of OE is typically understood to be VO for later texts, and before then either uniform OV (see van Kemenade, 1987), or variable OV/VO (Pintzuk, 1999). Subsequent transformation or dislocation operations, however, can move words and yield a substantially different “surface” order. The V2 phenomenon described above is typically analyzed as movement of the verb to second position from elsewhere in the sentence (e.g. Pintzuk, 1999); another displacement that has been posited in historical English is the movement of an object phrase from its original position to the end of the clause (“extraposition”: see van Kemenade, 1987; Pintzuk and Kroch, 1989).

Because of these transformations, the labels VO and OV are sometimes applied in the syntax literature to clauses in which the actual order of words is in fact the opposite. Here, I will only consider the order in which words are actually produced, using terms like VO and OV to refer to the position of words in a sentence as spoken or written. The reason for this is that my focus is on the effect that properties of the human cognitive architecture used to produce and comprehend language have on form of the languages we speak. Generative grammars in general, and transformational grammars in particular, are not designed with cognitive realism in mind: Chomsky (1965, p9) states quite succinctly that “... a generative grammar is not a model for a speaker or hearer.” (see Chapter 1). Therefore, we would need independent evidence before assuming that the structures they posit are those actually processed by language users. Early psycholinguistics in fact did attempt to find processing correlates of the structures and operations assumed in contemporary syntactic theory (see summary in Chomsky, 1968). Miller and colleagues (e.g. Miller and Chomsky, 1963; Miller and McKean, 1964) presented evidence that sentences containing transformations (similar to those which might change the underlying OV order to surface VO) incurred processing difficulty and were recalled less well. However, these results turned out be

be confounded by semantic differences between the sentences, and within a decade, it was widely accepted that no evidence for transformations playing a role in human language use had been uncovered (see Fodor et al., 1974; Reber, 1987).

Although only the order of words as actually produced is relevant for determining their processing complexity, it is still important to understand which word orders were among the grammatical options the language gave its speakers. As mentioned above, non-V2 clauses in OE are common, with most but not all falling into three classes.

First, verb-initial main clauses are common. Though over time this order comes to be reserved for its modern usage (basically only in questions and a few other constructions), verb-initial (V1) clauses in OE and closely related Old Germanic languages were particularly favored to express certain discourse functions, as discussed by Petrova (2006) and Petrova and Solf (2008). In particular, V1 clauses were used in certain cases where no given referents were being mentioned. Such uses include “presentational” sentences where new referents are being introduced (expletive *there*, as in “*there arrived a man...*” did not become available until at least the end of the OE period); similarly, the first clause of a new episode in a discourse was often V1, as is the case in (10). Additionally, V1 was also used for a sequence of clauses which describe closely related or temporally sequential events.

- (10) **noalde** eorla hleo ... cwealm-cuman
not.wanted noblemen-GEN protector murderous visitor-ACC
cwicne forlætan
alive let.go
‘*The protector of the warriors did not wish to let the murderer go alive*’
(*Beowulf*, likely c750, quoted in Hinterhölzl and Petrova, 2010)

Second, V2 order is variable after certain “scene setting” adverbs such as *pa* (“then”): there, the verb commonly appears in second or third position, and occasionally later, as in (11):

- (11) [pa] [py ylcan gere] [onfaran winter] [pa Deniscan pe
 [then] [the same year] [before winter] [the Danes that
 on Meresige sæton] **tugon** hira scipu up on Temese
 on Merseyside sat] **pulled** their ships up on the Thames
 (*Anglo-Saxon Chronicle: Parker MS*, c890, quoted in Kroch et al., 2000)

Third, in OE, pronouns typically appear before the verb even when there is another preverbal constituent, whether the pronoun is subject (12a) or object (12b). These sentences can still be analyzed as V2 if the pronoun is treated as a clitic that “does not count” as taking up the second position (van Kemenade, 1987; Pintzuk, 1999). However, pronouns are not *always* preverbal, even in relatively early texts (13).

- (12) a. [purh his wisdom] [he] **geworhte** ealle ping
 [through his wisdom] [he] **worked** all things
 (*Ælfric’s Homilies*, c1000, cited in Koopman (1997))
- b. [ah] [twegen culfran briddas] [him] **genihtsumedan**
 [but] [two pigeon birds] [him] **suffice**
 & twegen turturan gemæccan
 and two turtle-doves a pair
 (*Blickling Homilies*, c900)
- (13) donne **ærnap** hy ealle toward dæm feo
 then **gallop** they all toward that treasure
 (*Orosius*, c925, quoted in Traugott, 2007)

In summary, the V2 construction is clearly present in both Old and Middle English, but it is not used universally even in the main clause: Haeberli (2002) estimates that a syntactic V2 rule may have applied in about 70% of OE and early ME sentences. Moreover, the exceptions are a diverse group: only the more frequent orders are mentioned above, but all of the logically possible orders of subject, verb, and object are used (as will be demonstrated below).

It is also important to keep in mind that the possible word orders of Old English were not necessarily meaning-equivalent to each other, but could have signalled

particular discourse functions. As described above, V1 was commonly used when introducing new referents or in sequences of events, and certain scene-setting adverbs seem to co-occur with word order variation. The V2 template itself has been suggested to be associated with a separation of topic from focus in other Germanic languages and arguably Old English itself (e.g. van Kemenade, 1987; Kroch et al., 2000; Biberauer and Roberts, 2005; Hinterhölzl and Petrova, 2010, i.a.): the single preverbal constituent is taken as the topic (what the sentence is “about”), while the verb and postverbal arguments provide the comment (the additional information the clause provides about its topic). Hinterhölzl and Petrova (2010) argue that although OE has this basic distinction, it patterns slightly differently from Old High Germanic and related Germanic languages in that not only the “aboutness topic” but *any and all* backgrounded material appears preverbally. In the majority of sentences, at most one backgrounded element is mentioned, so the verb often appears in second position. However, this theory also accounts for the many exceptions to V2, including the fact that pronouns often appear preverbally even in addition to other constituents, since pronouns typically express given information. Therefore, Hinterhölzl and Petrova give an alternative (or additional) explanation to the suggestion that pronouns are clitics. A related hypothesis that could also explain the *reason* V2 might be associated with a topic-comment structure is advanced by Speyer (2008). Speyer suggests that there is a general pressure to avoid adjacent focused² or prominent elements, much like the pressure to avoid adjacent stresses in prosody. Since verbs themselves are rarely focused, placing a verb immediately after a focused topic NP complies with this pressure. However, pronouns are also rarely focused, so they can be placed adjacent to the initial topic, accounting for many of the exceptions to V2 described above. Speyer’s theory assumes that syntactic operations resulting in V2 order are lost for independent reasons, and that this results in the partial decline of object topicalization observed in Old English, since without the option of inserting a verb after an initial topic object, topicalization would often result in a focused subject and object being adjacent. I will return to this theory in chapter 3.

²Here, the term *focus* is used to indicate an explicit contrast with other potential referents or events.

Evidently, a full understanding of when each word order is used will probably require looking at multiple simultaneous factors such as discourse status, argument pronominality, adverb types, and so on. The present study will test whether phrase length is among these factors.

Previous research on historical English has uncovered numerous correlations between constituent length and placement. Kohonen (1978), for instance, suggests that length played a role in the placement of several types of constituent which retained relatively free word order in early ME. Seoane (2006) additionally showed that the passive construction, which exists in OE but increases drastically in frequency in late ME and Early Modern English, is used instead of an active construction more frequently in cases where the active would involve putting a long phrase before a short one. This finding is directly predicted by the principle of End Weight. Traugott (2007) finds that the majority of left-dislocation sentences in Old English involve a long and complicated object: as in the Present-Day English (henceforth PDE) sentences shown in (9), an infrequent word order is more common when it involves a particularly long constituent.

The most relevant finding is work by Pintzuk and Taylor (2006), who provide a statistical analysis of OE and ME data, finding that length as well as certain other properties of the object influence its position: longer objects are more often postverbal. They attribute the effect of length to processing factors: “length [...] may reflect a processing constraint against center-embedded material” (p252). However, Pintzuk & Taylor are not primarily interested in the role of weight or in processing constraints; rather, they are investigating the syntactic generalizations that can be made about the language after certain processing effects have been factored out.

Below, I present a model which investigates many potential predictors of word order in OE and ME simultaneously. I focus on the potential role of dependency length minimization as an influence on speakers’ choices. I also test whether the influence of any of these variables on word order changes over time.

2.3 Analysis

2.3.1 Data preparation

The data analyzed here is taken from two corpora: the York-Toronto-Helsinki Parsed Corpus of Old English (YCOE: Taylor et al., 2003) and the Penn Parsed Corpus of Middle English 2 (PPCME2: Kroch and Taylor, 2000a). The two corpora consist of texts from abutting time periods which have been parsed, yielding a collection of sentences annotated like (14), from late OE.

| | | | | | | | |
|---|---------|-----------|-----------|-----------|-------------------|---------------|------------------|
| (14) | [IP-MAT | [ADVP-TMP | [ADV-T | da]] | <i>there</i> | | |
| | | [VBD | mæssede] | | <i>celebrated</i> | | |
| | | [NP-NOM | [MAN-N | man]] | <i>one</i> | | |
| | | [NP-DAT | [D-D | dam] | <i>the</i> | | |
| | | | [N-D | cynge]] | <i>king</i> | | |
| | | [PP | [P | æt] | <i>at</i> | | |
| | | | [NP | [NP-GEN | [NR | Sancte] | <i>Saint</i> |
| | | | | | [NR-G | Clementes]] | <i>Clement's</i> |
| | | | | [N | cyrcean]]]] | <i>church</i> | |
| <i>“There people celebrated the king, at St Clement’s church”</i> | | | | | | | |
| <i>(Vision of Leofric, c1150)</i> | | | | | | | |

The YCOE and PPCME2 were produced using similar annotation and coding standards, which minimizes the problems that can arise when combining data from multiple sources. Even so, the two corpora were tailored for two stages of English which had quite different grammatical properties, and so there are some differences in the annotations chosen. For instance, the YCOE annotates noun phrases for case (nominative, accusative, etc) while the PPCME2 instead includes grammatical function information (subject, direct object, indirect object, etc). This is motivated by the decay of case marking over the course of time: by the 13th century, the case paradigm had largely collapsed, and the grammatical function of a NP was indicated by an adposition or by word order, or was structurally ambiguous. See Allen (1999) for discussion of changes in the English case system. Of course, case was lost gradually,

and the sharp discontinuity between the two corpora therefore introduces artifacts into the data. In particular, the collapse of the dative and accusative case was underway at the start of the ME period. NPs in the OE data which are morphologically ambiguous between the two cases are therefore annotated as such, even though their grammatical function may be evident and therefore included in the ME data. Since no other annotation is available, here I equate the OE accusative/dative distinction with the ME direct/indirect object distinction, and include case-ambiguous OE NPs in the more frequent accusative/direct category. This is not ideal, but better than ignoring the difference between object types altogether.

All of the texts included in this data predate the use of Gutenberg’s printing press, which was only invented around 1440. Texts were copied manually, and this fact has a great influence on the amount and type of data we have today. First, it means that both the concept of *text* and *manuscript* are relevant: originals could be copied many times, and copies themselves copied, with no guarantee of accurate reproduction. Copied manuscripts may therefore reflect both the language of the original author and that of the copyist. Second, the overwhelming majority of writing was religious in nature: until the 13th century, almost all writing and copying was performed by Christian monks, and even after that, the secular scribes who increasingly replaced them as the major producers of books chiefly addressed the demand for religious texts. The time and labor involved in copying a text meant that only writing seen as very important was reproduced, so beside the bulk of religious texts there are some histories and handbooks, but very little early secular poetry or prose. Third, a great deal of the extant texts are translations from Latin, and so even the English “originals” may show an influence of the Latin word order. It isn’t possible to exclude all these texts, since this would lead to the loss of a great deal of data, particularly in the earlier periods where there is already sparsity.

From a statistical point of view, it is important to model as accurately as possible the dependence and independence between the individual data points being analyzed, the clauses. For instance, clauses taken from a single text are not independent observations: they share properties such as the fact that they were written in the same year and by the same author, and therefore might be expected to behave more similarly

than a set of clauses taken from different texts. We can model this nonindependence by including variables such as the year of composition, the author, and even the specific text. With large datasets like the one I describe here, it is good practice to include any available control variables that are theoretically justified: each of these may account for variation in the outcome variable (here word order) that might otherwise have been falsely attributed to the predictive variables of interest (here object weight). One serious problem for this particular dataset is the severe nonindependence introduced when there are several manuscripts corresponding to a single text. Ideally, we would model this dependency by including for each manuscript a per-clause control variable indicating the word-order in the original text. This would only leave deviations from the original word order introduced by later copyists to be accounted for by variables like manuscript date and object weight. Unfortunately, that solution is impractical for two reasons: first, an identifiable original manuscript often (typically) is not available; and second, the class of regression models I use here would not be able to account for *both* originals and copies with this control variable as it would be undefined for originals. Instead, I take the expedient solution of identifying all sets of manuscripts with a single source text, and deleting all but one of them at random. This results in a substantial loss of data, but avoids the serious problem of “double counting” clauses that are copies of each other.

Manuscripts differ somewhat in the amount and accuracy of information we have about them. Later works tend to be easier to date accurately, although certain earlier works — such as a great number of laws, wills and charters — are explicitly dated. I used dates given in the Helsinki corpus (Kytö, 1996), in Ker’s (1957) Catalogue, and in Sawyer’s (1968) Anglo-Saxon Charters to establish dates for the composition of each text and copying of each manuscript. Although in some cases these dates are specific years, in many cases there is simply some interval of time within which the text or manuscript is believed to have been created, often as wide an interval as a century. In these cases, I took the midpoint of the range.³

³Pintzuk and Taylor (2004) argue against this, saying that “assigning the midpoint of a 100-year period as the date of composition for a quantitative analysis seems meaningless.” It is true and should be stressed that the less accurate the date estimate, the less useful and interpretable the resulting analysis. However, there is no good reason to throw away data for those texts that *can* be more

For the purposes of this chapter, I restrict the object of study to word order in matrix clauses (rather than clauses embedded inside other clauses, such as “I think that [John lives here]”) and non-co-ordinated verbs (as opposed to co-ordinations like “eat and drink”). I exclude clauses with an auxiliary (including (pre-) modals used with a second verb), since the dependency structure between subject and main verb is arguably different there.⁴ The variable of interest here is the order in which the verb and its dependents are placed, so I only include clauses with an overt subject and object. Both accusative/direct and dative/indirect objects are included, but clauses with *both* a direct and indirect object are excluded.

The number of clauses in the data falling into each possible word order is given in Table 2.1.

| | SO | | OS | |
|----|----------------|---------------|---------------|--------------|
| | SOV | | OSV | OVS |
| OV | 3576 (65%) | | 1135 (21%) | 765 (14%) |
| VO | SVO | VSO | VOS | |
| | 10292 (74%) | 2904 (21%) | 664 (5%) | |

Table 2.1: Total number of clauses with each word order in dataset

These summaries show that all of the possible orders of the verb and two arguments occur in the data.⁵ The three SO word orders together account for 87% of this data. Typologically, these are the base orders of almost all the world’s languages, with SVO and SOV alone accounting for something like 90% (Tomlin (1986) gives 45,

accurately dated. The alternative approach of binning text dates — Pintzuk & Taylor in fact use 200 year periods for OE — reduces to fundamentally the same solution of assigning some value for an unknown date, and only differs in that the values chosen are arbitrary century boundaries rather than a best estimate, making it strictly *more* “meaningless”. More sophisticated techniques could preserve and model the different uncertainty associated with dating each text — explicitly representing the different measurement error associated with each datapoint — but this I leave to later work.

⁴The precise dependency relation depends crucially on whether we assume that dependencies capture purely semantic or also syntactic relationships between words. I will return to this very briefly in section 4.2.1.

⁵It is possible, of course, that some of the more esoteric exceptional word orders are simply transliterations from Latin, which has extremely free word order.

42, and 9% for SVO, SOV, and VSO respectively). The relatively high frequency of SOV and OSV confirms that V2 is not a categorical constraint, though of these 4,711 clauses, 3,986 have a pronominal subject or object (or both). This leaves at least 725 (15%) that cannot be explained by analyzing pronouns as clitics on the verb, even ignoring the possibility of preverbal adverbials or other constituents. Of the 3,568 VSO and VOS clauses, 3,194 begin with some adverbial or other constituent, leaving 346 (10%) that are truly main verb-initial.

If we ignore the temporal dimension and just look at word order across the entire dataset, significant variation is evident. As described above, I focus here on the order in which the object and main verb appear: additionally, I will only consider sentences with an SO order, excluding the relatively infrequent OS orders. Therefore, the VO/OV alternation amounts to the choice between putting the object at the end of the clause versus putting it in the middle.⁶

The plot in Figure 2.1 gives an initial overview of the change in frequency of VO and OV orders. Points on this plot represent individual texts, and their relative size represents of the number of clauses in each that fit the criteria above. Some notable texts are labeled. The y-axis indicates the proportion of those clauses that are are VO, and the dotted line shows the trend of change in proportion over time.⁷ Finally, for manuscripts which are not originals, a horizontal arrow shows the period of time that elapsed between the approximate original text date and the approximate date of copying.

The figure shows that some periods contribute much less data than others. In particular, social changes resulting from the Norman invasion of 1066 led to a sharp decline in the output of written English, as scribes reverted to writing in Latin, or occasionally French. There are almost no newly composed English texts for a century after the invasion. It also appears that there is much more word order variation in OE than ME, though this should be taken with a grain of salt since there are many more short OE texts (mostly wills and charters) making per-text estimates less accurate.

⁶In other work, I am using more sophisticated modelling techniques to look at the influence of length and other factors on *all* logically possible word order outcomes. Here, I stick to standard, widely available statistical packages, which means using only a two-way outcome.

⁷This is a *loess smoother*, a statistical method for approximating an arbitrary curve.

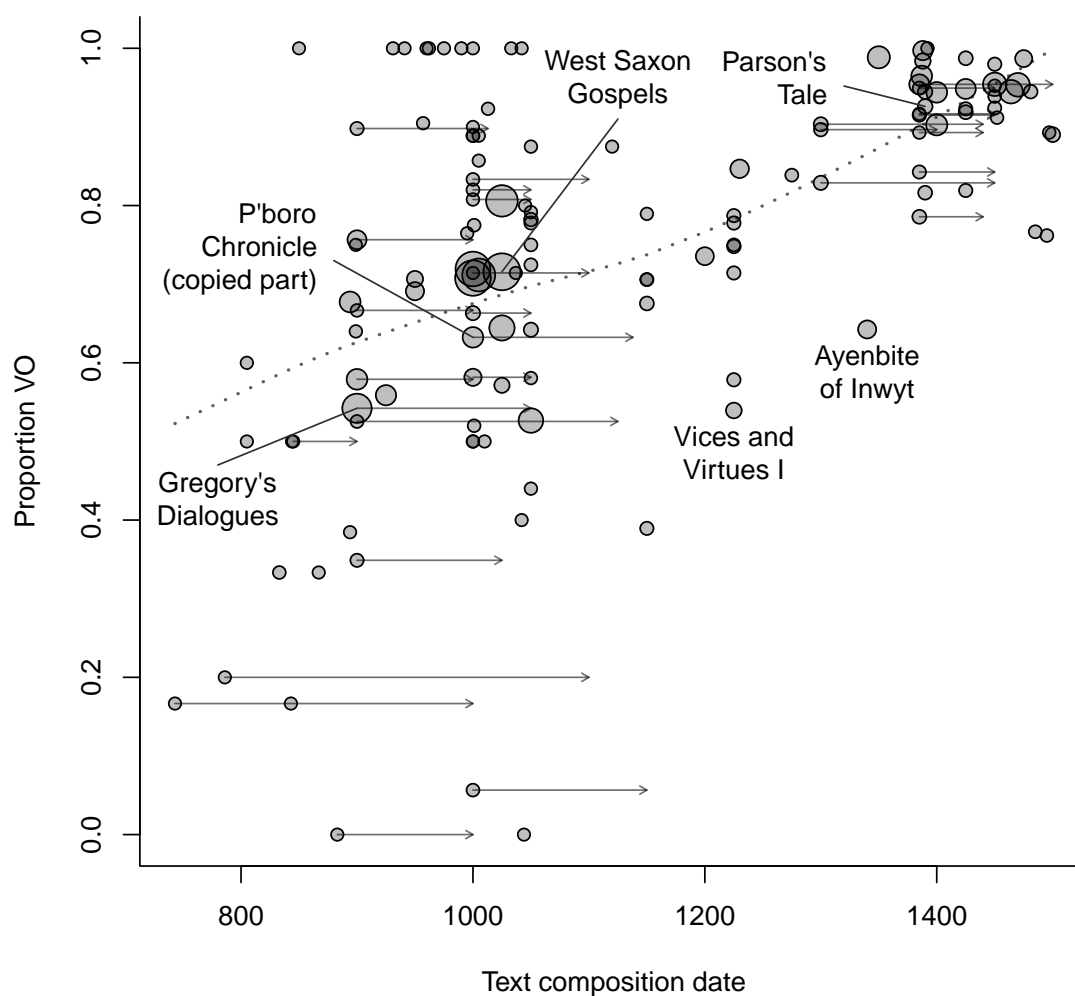


Figure 2.1: Proportion of main clauses in each text with VO order

2.3.2 Statistical analysis

To model the importance of object length and other factors on the choice of word order, I use a multilevel logistic regression (see Gelman and Hill, 2006; Jaeger, 2008) with VO order as the outcome variable.

Pintzuk and Taylor (2006) used a VARBRUL style multivariate analysis to show that longer objects are more likely to be placed after the verb. In that analysis, both date and object length were binned and entered as discrete (and non-ordered) factors. Their results suggest that it would be possible to produce a single model of the effect of phrase length on word order choices for the entire period, including year and length as covariates. Multivariate analysis, which models multiple possible influences simultaneously is to be preferred over simple summaries for each influencing variable. Any one variable may be correlated with another (and hence appear to have an influence when it does not, or vice versa) or may have an effect which is small enough that it does not become apparent until other effects are factored out. However, a multilevel logistic regression model is more suitable for this kind of analysis, and has many advantages over VARBRUL. First, it is better to use continuous covariates for variables like date and length that are naturally continuous: binning reduces statistical power (see e.g. Cohen, 1983) and obfuscates smooth trends across time. In fact, treating each period separately prevents the direct investigation of possible differences in the influence of certain variables over the timecourse of the entire dataset. Second, multilevel regression allows sources of variation at different levels of analysis to be taken into account, which allows us to factor out variation due to observations coming from (e.g.) different texts and regions. For general arguments motivating multilevel analysis see Gelman and Hill (2006) and Jaeger (2008); for a specific comparison of traditional VARBRUL and multilevel analysis see Johnson (2009).

Below I summarize the predictors to be entered into an initial model as fixed effects. The first are the variables investigated by Pintzuk & Taylor:

- **Text Date**, as a year.
- **Object length**, in words.
- **Object type/case**, either direct/accusative or indirect/dative.
- **Negated object**, a binary indicator of whether the object is negative: either inherently (e.g., PDE *none*) or by being negatively quantified (e.g. PDE *no*).
- **Quantified object**, a binary indicator of the presence of a non-negative quantifier such as *every* in the object.

- **RC modifying object**, a binary indicator of whether the object NP is modified by a relative clause.

The following fixed effects are added in this study:

- **Subject length**, in words.
- **RC modifying subject**, a binary indicator of whether the subject NP is modified by a relative clause.
- **Pronominal subject**, binary.
- **Pronominal object**, binary.
- **Manuscript lag**, the number of years between the date of the original text and the date of the manuscript.⁸ It would not be a good idea to enter the raw manuscript date along with the text date in the model: the two are highly correlated, and for a majority of data the same value.
- **Latin original**, a binary indicator of whether the text is known to be a translation from a Latin original.
- **Verb POS**, the part of speech of the verb, from the set *base* (infinitive, imperative), *present tense* and *past tense*.

The following factors are entered as random effects; that is, they might be expected to capture individual deviations from the general population-wide trends, but we have no particular theoretical interest in the specific nature of those deviations (see explanations in Gelman and Hill, 2006).⁹

- **Text**, the identity of the text (note that only one manuscript is included per text, so this could equivalently be considered as the identity of the manuscript).

⁸This intuitively interpretable measure is comparable to what would be obtained by the standard decorrelation technique of residualizing.

⁹Arguably, some of these variables might have been entered as fixed effects, since they exhaust the set of possible values (e.g., each text belongs to one of a finite set of genre classifications). However, I am not interested in the effects of these variables in particular, and I have no strong theoretical reason to think any of the variables where this is the case should influence the outcome, while I knew that other existing fixed effect variables that correlate with these would have an effect of primary interest (in particular, date). Therefore it is better to minimize the number of degrees of freedom “wasted” on these secondary controls by treating them as random effects.

- **Genre**, an indicator of text type from the set *science*, *law*, *religious*, and *other*.
- **Dialect**, an indicator of whether the text comes from the *West Saxon*, *Kentish*, or *Anglian* OE dialects or the *Southern*, *West Midlands*, *East Midlands* or *Northern* ME dialects (see below).

To investigate the possible changes in the importance of the variables over time, I also enter interactions between text Date and each of the following: Object Length, Subject Length, Object Pronominality, Subject Pronominality, Object Type/Case, and Dialect. Similarly, I included random slopes for Text Date to allow the potential for different levels of Genre and Dialect to display different rates of change.

The Dialect variable is problematic, and not just because of the intrinsic difficulty of determining where to impose dialect boundaries on the fluid space of regional variation. Although there is a reasonably good consensus on the division of extant OE texts into dialectic categories of West Saxon, Kentish, Mercian and Northumbrian, there is not a continuous record of text witnesses for each; rather, certain regions dominate in different years. All of the earliest extant manuscripts are West Saxon, and there are very few Kentish or Northumbrian documents at all. This means that the dialect variable is confounded with the date of the texts. The problem is compounded by the fact that dialect boundaries in the ME period — by which time there had been a great deal of levelling of the differences between regional varieties — are controversial, and do not map cleanly onto the OE dialects. See Görlach (1997), Hogg (2006) for details. In order to avoid this problem, I experimented by grouping the more northern and eastern dialects apart from the southern and western ones, yielding a binary variable which spans the entire dataset. However, this variable evidently conflates the differences between dialects, so I also tried a coding with different levels for the OE and ME texts based on the Helsinki corpus classifications. This includes the four ME dialects listed above, and for OE collapses together Mercian and Northumbrian texts as well as texts of unclear Mercian/West Saxon influence as the region “Anglian”. In fact, neither coding improved model quality significantly, and whichever variable was used, it was removed during model comparison. The results reported below used the seven-way classification.

The Genre variable was designed to minimize similar problems of confounds with date: again, because different types of text were produced during different periods, a finer distinction would have made it impossible to tell differences between genres from differences due to change over time. I constructed four categories that have reasonable support across the entire dataset: *science*, texts which detail medicinal, astronomical and other technical knowledge; *law*, including charters, wills, and ecclesiastic laws; *religious*, consisting of biblical text and apocrypha, sermons, homilies, and similar material; and *other*, a catch-all category containing philosophy, fiction, biographies, and history.

To a lesser extent, the Text variable is also conflated with date: in fact, it is nested within it, as each text has only one date. However, the large number of texts means that the model fitting procedure should be able to correctly partition variation into between-text and between-year effects, so this does not pose a problem.¹⁰

2.3.3 Results

A model was built by first entering all of the effects into a regression, and then deleting variables stepwise that do not significantly improve the fit to the data by chi-square model comparison. The “drop 1” procedure was used: first a full model was fitted containing all of the predictors described above. Then, each predictor was tested as a candidate for removal: if model comparison indicated the model with that predictor was not superior at $p < .1$, it was removed. This procedure was repeated to arrive at the final model. The predictors removed are listed in order in Table (2.2). The final model had all VIFs < 1.5 , indicating no obvious collinearity problem.

Continuous variables were standardized by centering and dividing by two standard deviations, as suggested by Gelman (2008). Centering means that the intercept represents the log odds of VO order for an “average” clause with mean object and subject length, from the middle of the time period, etc. Dividing makes the regression coefficients directly comparable as “effect sizes”, revealing which predictors have

¹⁰Accordingly, later inspection of the model’s per-text random intercept estimates (Gelman & Hill’s (2006:251–278) α_j , sometimes called BLUPs) show no correlation with date, and no differences in distribution across time.

| | df | χ^2 | p_{χ^2} |
|-----------------------------|----|----------|--------------|
| Quantified object | 1 | <.001 | 0.98 |
| RC modifying object | 1 | 0.022 | 0.88 |
| Manuscript lag | 1 | 0.22 | 0.64 |
| RC modifying subject | 1 | 0.46 | 0.50 |
| Subject length * Text date | 1 | 0.84 | 0.36 |
| Text date * Dialect | 1 | 0.85 | 0.36 |
| Dialect | 1 | 0.74 | 0.39 |
| Latin original | 1 | 0.010 | 0.92 |
| Text date * Pronoun subject | 1 | 0.025 | 0.87 |
| Text date * Object length | 1 | 0.26 | 0.61 |

Table 2.2: Predictors removed during model comparison

a greater effect on the word order outcome.

In Table (2.3), I present the coefficients (β) for all fixed effects in the final model, and plot them with error bars indicating the standard error of those estimates. Intuitively, variables with effects further from the zero value have a larger influence on the outcome, with variables favoring VO having a positive effect size and those favoring OV negative. The standard deviation of the remaining random effect, Text, is also listed and plotted on the same scale. Individual texts may differ from the global baseline level of VO rate, and the plots indicate the distribution of per-text base rates around that global baseline.

The variable of most interest is of course the length of the object, and its effect patterns in the expected direction: longer objects have a higher probability of being placed postverbally, controlling for all the other variables in the model. This is in accordance with the predictions of the Principle of End Weight. This can be seen more clearly in Figure 2.2, which shows the model's predictions for objects of different lengths over time: throughout the data, longer objects have a higher probability of being realized postverbally, though by the ME period VO order has almost reached ceiling.. This figure also illustrates the effect of the object case/type interaction with text date: in early texts, dative or indirect objects are more likely to be preverbal, while in later texts the reverse is true. Unexpectedly, the model also suggests that shorter and pronominal subjects lead to more VO sentences. However, these effects

| | β | p_z | |
|-----------------------------|---------|--------|--------------|
| Intercept | 1.9 | <.001 | -1.5 0.0 1.5 |
| Dative/indirect | 0.031 | 0.58 | |
| Negated object | -0.33 | 0.011 | |
| Pronoun object | -1.2 | <.001 | |
| Object length | 0.92 | <.001 | |
| Subject length | -0.087 | 0.0012 | |
| Text date | 1.5 | <.001 | |
| Text date * Dative/indirect | 1.4 | <.001 | |
| Text date * Pronoun object | 0.43 | <.001 | |
| Pronoun subject | -0.089 | 0.0033 | |
| Verb POS=base | 0.72 | <.001 | |
| Verb POS=past | 0.16 | <.001 | |
| | sd | | |
| Intercept Text | 0.84 | | |

Table 2.3: Final model for VO/OV order (positive outcome is VO)

are extremely small, and may be spurious.

The significant effect of object pronominality shows that pronouns display exceptional placement even controlling for their short length. In OE, pronominal subjects and objects were both usually preverbal, and the tendency to keep pronouns before the verb survived well into the ME period. Remnants of this tendency exist even today in fossilized language, such as standard Christian wedding vows, which include phrases like “I *thee wed*”, and “until death do *us part*”. As mentioned above, the distributional behavior of OE pronouns has led some syntacticians to analyze them as bound morphemes: clitics, or “weak” pronouns (van Kemenade, 1987; cf Koopman, 1997); others have suggested that pronouns appear preverbally simply because they are typically unfocused or given information (Speyer, 2008; Hinterhölzl and Petrova, 2010). Either way, the model shows that pronominal objects tend to be placed preverbally as expected.

The fact that the Dialect variable does not improve the model probably reflects the inadequacy of this measure rather than being evidence that there were not dialect differences in word order. As described above, the dialect coding I chose avoids

confounds with date and genre at the cost of conflating together potentially real differences between speakers of different dialects, perhaps leading to non-significance. The non-significant genre variable can be explained similarly. Coding for Latin originals also fails to improve the model. This is not surprising: as Latin has free word order, there is no particular reason why translating from Latin should systematically bias the scribe in one direction or the other; rather, we might simply expect “more noise” due to influence from the original word order, which may be OV or VO in each clause. More advanced modelling techniques could investigate this possibility.

As found by Pintzuk & Taylor, negated objects tend to appear preverbally. In contrast to their findings, however, neither quantification nor the presence of a relative clause are found to be significant predictors of word order. In the case of the RC, this is probably due to a different coding system: Pintzuk & Taylor binned together long objects with objects of any length having relative clauses. This analysis controls for both separately, but suggests that treating NPs with relative clauses just like any other long NP is valid: relative clauses do not appear influence word order above the additional length they add. This supports the idea that weight effects are well captured by simple phrase length rather than any more nuanced metric of syntactic complexity (Wasow, 1997; Szmrecsányi, 2004). There isn’t an obvious explanation for quantification not being significant here, except to say that it is possible that this effect only manifests itself when there is an auxiliary verb, or that whatever influence object negativity had in the earlier analysis may have in fact been due to one of the other variables controlled for simultaneously here. Finally, the type of verb has a significant effect: although there is little difference between present and past tense inflected forms, base forms (imperatives and infinitives) favor VO.

2.4 Discussion

The fact that object length influences word order choice in historical English matches the observations made by Wasow and colleagues in Present Day English in support of the Principle of End Weight.

While the regression model presented here is a production model in the sense

that it predicts the probability with which a writer should produce a VO or OV ordered clause, it is unlikely to be a good model of the actual syntactic choices that a language user would go through when producing Old or Middle English. In particular, the model collapses over all possible constructions that result in a surface VO order, and all that result in a surface OV order.¹¹ For the purposes of the psycholinguistic theories of interest here, this is legitimate: the End Weight principle, for example, predicts that heavier phrases should be placed later in the *linear string of words* where possible, regardless of which syntactic structures might have to be chosen to arrive at any given linear ordering.

However, equating all structures with the same order of object and verb obscures potential differences that prove to be relevant both in the choice of word orders and in the change towards VO. Regardless of which syntactic theory is chosen to describe the data, it is likely that the V2 construction influenced the change from OV to VO. In OE, OVS sentences were more common when the object was given, but the reanalysis of the preverbal position as a marker of syntactic rather than information structural status led to the loss of this order, and perhaps to the introduction of alternative constructions to express a nonsubject topic (see Los, 2009). Of course, this is not the whole picture: all Germanic languages seem to have had V2 at some point, and as many of these have remained stable as have switched. The V2 construction may therefore be an *enabling* cause for change in that it provides a frequent SVO structure, but cannot be a *sufficient* cause, since a language-wide shift does not always result.

Unfortunately, information structure is not annotated in this data. It is likely that object length is related to givenness, if old or more accessible material can be expressed with shorter referring expressions. Thus, shorter objects might be predicted to appear preverbally in OE. Including pronominality in the model should control for this possibility to a large extent, as given referents are typically pronominal. Even so, without explicitly controlling for accessibility or givenness, it's not clear

¹¹This said, excluding OS sentences has the side-effect of excluding sentences that might be analyzed as the result of an object topicalization rule in which the object is moved to a position before the subject. This means that the VO/OV alternation under study here is not directly related to object topicalization, a process which has been argued to decrease in frequency from Old to Early Middle English (Speyer, 2008).

how much of the residual effect is “purely” due to object length, and how much due to something else with which it is correlated. Hawkins’ (1994) theory suggests that weight is the primary determinant of word order, and that other putative influences like information status in fact only affect order choices through their correlation with phrase complexity or length. However, there is evidence for both weight and accessibility/givenness having independent roles: given two NPs of similar weight, speakers of present-day English are more likely to choose a word order that puts given information first; but when both NPs are given, they are also more likely to put the heavier one last. Therefore, it is unlikely that apparent effects of givenness can be reduced to simple weight (see e.g. Arnold et al., 2000; Wasow, 2002). As I will discuss in Chapter 3, it is also unlikely that weight can be reduced to givenness, since in some languages the two appear to have opposite effects on ordering.

In summary, the results here confirm that relationships between weight and word order like those discussed by Wasow (2002) can be identified even in the textual record of a long dead language. It seems possible that universal processing preferences underlie these trends; I will next turn to a discussion of what those processing preferences may be.

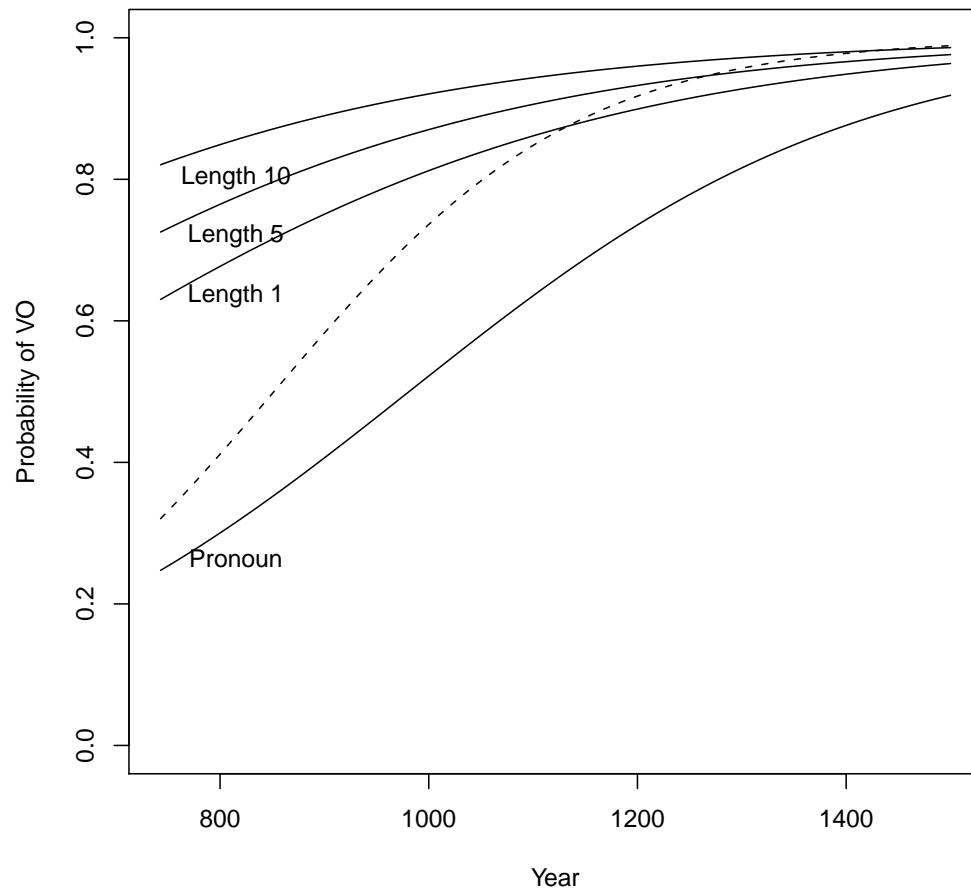


Figure 2.2: The estimated influence of the length and pronominality of accusative/direct objects on order over time (solid lines); predictions for a 2 word dative/indirect object over time (dashed line). Other predictors are at baseline level.

Chapter 3

Dependency length minimization

In Chapter 2, I presented evidence that the order in which the verb and object appeared in a clause of Old or Middle English was influenced by the length of the object. When writers expressed a clause with a longer object, they were proportionally more likely to place the object after the verb rather than before. In this chapter, I describe some psycholinguistic theories which predict certain word orders to be more prevalent than others, focusing on *dependency minimization*. I consider the evidence for a general pressure towards minimization of linguistic dependencies in the history of the English language.

3.1 Serialization of non-ordered meanings

All human languages, spoken and signed, impose a serial ordering on the symbols that make up their sentences: disregarding relatively minor “out-of-channel” signals like prosody, the sounds or gestures that are the building blocks of language are produced one-by-one. This fact can be explained as simple convenience. In production, more complicated articulatory and gestural systems would be needed to produce multiple symbols in parallel, and parallel transmission would place extra demands on the comprehender’s attention that might lead to comprehension difficulty or mistakes. Comprehenders could, in theory, wait until an entire sentence has been heard and stored in some memory buffer before starting to process it, allowing all the words to be

processed in parallel even if heard or read serially. However, this would result in extra memory demands, brief periods of frantic processing interspersed by idle waiting, and delayed comprehension. In fact, psycholinguists are in more-or-less complete agreement that comprehenders incorporate incoming information incrementally as it is heard or read, without waiting for a complete sentence (see e.g. Tanenhaus et al., 1995). Therefore, both production and comprehension operate on a serial string of words.

But this convenient seriality is only a property of the language systems used to express meanings, not of the meanings themselves. There is nothing in the meaning of the sentence “*Andrew likes cake*” to suggest that the meaning of *Andrew* should come before or after that of *likes* or *cake*. Given a three word sentence like this, there are in fact six logically possible ways the words could be ordered. For this sentence, standard English only allows one of the possible word orders, SVO (OSV is also allowed in many dialects with when it signals a particular discourse function, as in “*now chocolate cake, Andrew likes.*”) Other languages differ in having some limited flexibility (say, allowing only SOV and OSV) or allowing all six. The same is true for the order of words in other functional categories: languages differ in whether they have prepositions or postpositions or allow either, for instance. According to the Performance-Grammar Correspondence Hypothesis discussed in the Introduction, however, there should be some systematic pattern within this cross-linguistic variation. Specifically, an order that is fixed as the only possibility in a certain language should (a) be among the possibilities available to languages with freer ordering constraints, and (b) be among those which are on average less complex to process in those languages that have it as an option.

This raises the question of why there might be differences in the processing complexity of different word orders at all. In this section I discuss some prominent psycholinguistic theories which make concrete suggestions.

3.1.1 Conceptual accessibility and availability-based production

One influential proposal for a theory about which word orders are favored by the human language user is that of *availability-based* production. This theory suggests that speakers tend to order phrases first if the concepts they describe are somehow easier, simpler, or otherwise more readily available or accessible to the processor. At the heart of this theory is the intuition expressed by Behaghel (1909) and discussed in Chapter 2 that a speaker has more time to plan words or phrases that occur later in the sentence. Put differently, the speaker may try to produce linguistic units as soon as possible after they have been planned, and therefore the syntax chosen will reflect the order in which information about phrases being planned becomes available (Prat-Sala and Branigan, 2000; Ferreira and Dell, 2000). Ferreira and Dell show that speakers do not seem to be concerned whether the omission of certain optional words like the complementizer *that* would lead to ambiguity, but instead choose whether to omit based on whether the linguistic material that follows is repeated or explicitly probed, and hence more likely to be immediately available to the speaker. In this situation, it is a discourse factor that determines how available a linguistic expression is: repeated or given phrases are more available, so optional material is not inserted before them. Similar factors influence word order: cued, repeated or given phrases tend to be ordered earlier where the syntax allows (Bock and Irwin, 1980; Bock, 1986; Tomlin, 1995, e.g.). Such “temporary” influences on availability are labeled *derived* accessibility (Prat-Sala and Branigan, 2000), since they are derived from context, and are known to influence the form of a referring expression (e.g. omission, pronoun, lexical NP) as well as its position in the sentence (Ariel, 1990, 2001; Gundel et al., 1993). However, these properties are just a subset of those now known to influence the form and position of linguistic expression. Unchanging properties of the referents and expressions involved, known as *inherent* accessibility, have similar effects. Indeed, Jaeger and colleagues (Jaeger and Wasow, 2006; Jaeger, 2006, 2010) show that optional words like those studied by Ferreira and Dell are more likely to be included when a multitude of different properties of the following expressions render

it less salient, predictable, or otherwise accessible.

Inherent and derived accessibility are often gathered together under the umbrella term *conceptual accessibility*, a construct that was first investigated in detail by Bock and colleagues in the early eighties (e.g. Bock and Irwin, 1980; Bock, 1982; Bock and Warren, 1985). Bock and Warren (1985) collected together previous findings that a number of different properties of noun phrases tend to lead to them being expressed as subjects in English: animacy, concreteness, imageability, definiteness, salience, givenness, and the speaker's interest in and empathy with the referent. They group all of these properties as facets of *conceptual accessibility*, which they suggest can influence the speaker's choice of which referent to make the grammatical subject in any given clause. However, Bock & Warren suggest that conceptual accessibility *only* influences the choice of which grammatical function (subject, direct object, and so on) to assign to each referent, and does not have a direct effect on word order independently of the ordering of these grammatical functions. Their evidence for this comes from a set of recall experiments: participants saw a set of sentences in which there were two critical nouns, one more imageable (and hence more conceptually accessible) than the other. If more accessible concepts were more likely to be placed earlier in the sentence regardless of grammatical function, then when participants were later asked to recall the sentences they had seen, Bock and Warren expected more misrememberings that moved the imageable noun to the front when it actually had been seen later. If more accessible concepts are more likely to be made subjects, then they expected that effect when participants had to remember whether they had seen an active or passive sentence (15). If accessibility also influences the assignment of direct and indirect object, it might also influence recall of the dative alternants in (16). If accessibility in fact affects word order directly, it should influence those two conditions and also simple co-ordinations which do not involve a change in grammatical function (17). In each case, half the items in each sentence type had the high imageability noun first, and half second (for instance, the simple active in (15a) was matched by another active where an imageable word like “doctor” was the grammatical object).

(15) a. the doctor administered the shock

b. the shock was administered by the doctor

(16) a. the old hermit left the property to the university

b. the old hermit left the university the property

(17) a. The lost hiker fought time and winter

b. The lost hiker fought winter and time

Accordingly, participants misremembered whether a sentence had appeared as an active or passive, or as an NP NP or NP PP dative, more often when doing so moved the imageable NP earlier in the sentence. Essentially, participants “reformatted” the sentences into an accessible-before-inaccessible template. Nevertheless, there was no effect within the co-ordination sentences. Bock and Warren took this to show that accessibility influences a speaker’s choice in grammatical function assignment, but not in word order more generally.

Under this account, length could be one of the properties that makes up conceptual accessibility, but only if it does not directly influence order except through grammatical function choice. However, this would be incompatible with the early results of Behaghel discussed in Chapter 2. In a fairly similar task to Bock and Warren’s, he found that the length of phrases in a German co-ordination did affect order production: longer — and therefore by hypothesis less accessible — phrases appeared later. More recent work by Temperley (2005) found that co-ordinations in Present Day English corpus data show a tendency to place the shorter noun phrase first. Additionally, Benor and Levy (2006) showed that in similar data, not only length but also frequency — another component of accessibility — influences order.

In fact, a great deal of recent experimental work on accessibility has brought into question Bock and Warren’s suggestion that accessibility does not influence word order except via grammatical function. Most clearly, in languages in which the order of subject and object is not fixed, accessibility influences the order of the noun phrases independently of grammatical function. Branigan and Feleki (1999) had participants

recall Greek sentences that were either the canonical SVO word order or the grammatical but less frequent OVS order. Participants were more likely to mis-remember a sentence that had been presented in one word order as if it had been presented in the other order in cases where the misrecalled word order placed the animate NP first. As in Bock and Warren (1985), participants “reformatted” the sentences to fit an animate-first template. This finding was replicated for the SOV and OSV orders of Japanese by Tanaka et al. (2005) (reported in Branigan et al., 2008). Therefore it remains possible that length could be a component of accessibility.

Given what has been said so far, it could be that not all of the factors grouped together as accessibility have independent effects on ordering, but are merely correlated with one or a small number that do. Length, for instance, could appear to affect word order merely because it is correlated with a component of accessibility such as givenness or animacy. However, recent work suggests otherwise. The English possessive alternation (i.e., the choice between “*Constantine’s conversion*” and “*the conversion of Constantine*”), for instance, appears to be influenced by both animacy and phrase length independently. Rosenbach (2005) used a corpus analysis and an experiment involving a forced choice between these two constructions given in context to investigate the factors determining construction choice. Although animate referents tend to be shorter, both the animacy and the length of the possessor NP were found to have independent effects on the choice of outcome, with shorter and animate possessors favoring the *’s* construction (see also Jäger and Rosenbach, 2006). Note that this again is a word order alternation, since the construction chosen determines the order of the two NPs: animates and shorter NPs tend to be placed first. Kempen & Harbusch’s (2004) corpus study of German subject/object/indirect object order revealed similarly independent effects of animacy and definiteness. Bresnan et al. (2007) used a heavily annotated corpus to determine the properties that influence speaker choice in the English dative alternation (e.g. “*give* recipient theme” vs “*give* theme *to* recipient”). They found independent effects of concreteness, animacy, definiteness and length of the arguments, amongst other effects.

Given the evidence discussed above, it seems possible that the phrasal length effects discussed by Wasow (2002) and in Chapter 2 could be due to accessibility.

Longer constituents are more complex, and therefore tend to be “put off” until later in the sentence more often. However, there is now ample evidence from head-final languages that phrase length and accessibility can be dissociated. This is possible because the relationship between conceptual accessibility and word order seems to hold independently of the basic word order of the language (see Jaeger and Norcliffe, 2009; Hawkins, 2004). Human and accessible concepts tend to be placed earlier in the sentence in both verb-medial languages like SVO English (Bock et al., 1992; Ferreira, 1994; McDonald et al., 1993) or Spanish¹ (Prat-Sala, 1997) and verb-final Japanese (Branigan et al., 2008). Likewise, Japanese and Korean (which is also verb-final) both follow the same given-before-new ordering as head-initial languages (Choi, 1999; Yamashita, 2002). In German, which — like Middle English — has SVO as the most frequent main clause order but also allows OVS, OVS is more common with animate objects and inanimate subjects (van Nice and Dietrich, 2003; Bader and Häussler, 2010). Similar effects hold in German’s verb-final subordinate clauses (Kempen and Harbusch, 2004).

Weight in head-final languages, however, has a qualitatively different effect. In uniformly head-final languages like Japanese, speakers seem to prefer to place heavier phrases close to the beginning of the sentence (Hawkins, 1994; Yamashita, 2002; Yamashita and Chang, 2001). Although Yamashita’s (2002) Japanese corpus study found only 19 sentences (less than 1%) to have a direct or indirect object preceding the subject, 14 out of these 19 had long objects (non-nominal complements are included under this definition of “object”). These objects were either sentential complements as in (18a), or a noun modified by a relative clause as in (18b). These examples are edited slightly for brevity. Although OS sentences like (18) are apparently rare, when they do occur it is almost always with a long object.

- (18) a. [zinsei ni oite muda na koto ha nani hitotu mo nai to]
 [in life meaningless things-TOP even one not.exist-QUOT]
 [watasi ha] omou
 [I-TOP] think

¹Verb- and object-initial main clauses are also possible in Spanish, but an SVO template appears to account for more than 90% of transitive main clauses (Clements, 2006).

I think there is nothing at all in life which is meaningless

- b. [watasi ga oboete iru hanasi wo] [kare ha]
 [I-SUBJ remember story-OBJ [he-TOP]
 sukkari wasurete simatte iru kamosirenai
 completely forgotten maybe
He may have completely forgotten the story that I remember

Using a constrained sentence construction task in which participants constructed a Japanese sentence from phrases presented out of order on a screen, Yamashita and Chang (2001) also showed that participants were more likely to produce an OSV sentence in the lab when the object was longer. In Chapter 5, I also present a further replication of this finding using a sentence recall paradigm like the studies by Bock and Warren (1985), Branigan and Feleki (1999) and Tanaka et al. (2005) described above.

These facts are clearly at odds with the predictions of the End Weight principle, and with the idea that the influence of length on order can be purely derived from accessibility. Although Japanese speakers prefer to place human referents earlier in the sentence, they place shorter NPs *later* where possible. Therefore, although phrase length may be correlated with conceptual accessibility, it does not always pattern with other accessibility factors, and therefore cannot be reduced to accessibility alone. This is not to say that accessibility effects are not responsible for significant typological generalizations about possible grammars. For instance, since subjects are much more often human and more often salient or given than objects, this is plausibly part of an explanation for Greenberg's (1963) Universal 1: when subject-object order is fixed in a language, subjects almost always precede objects.² However, more is needed to account for phrase length effects if the theory is to generalize to head-final languages. I will now turn to a second psycholinguistic theory which can account for findings like those.

²Assuming that objects tend to be longer than subjects (a natural consequence of the prevalence of pronominal subjects in languages which do not allow subject omission), and due to the preference for longer phrases to be ordered before shorter ones in verb-final languages, it could also explain the fact that SOV languages commonly have OSV as an alternative word order, while primarily SVO languages less commonly allow OVS.

3.1.2 Dependencies between words

Above I discussed the fact that the order of words does not typically reflect a property of the meaning expressed, since many meanings that speakers express are not intrinsically ordered in any obvious way. However, this should not be taken to suggest that the meaning of the sentence has no logical bearing on word order at all. Rather, there are systematic relationships between certain words which are independent of the order in which those words are placed, and these need to be reconstructed from the serial string of words which is received as input in order to arrive at a meaning (see discussion in Pinker and Bloom, 1990; Dell et al., 1997; Grodner and Gibson, 2005). Therefore, it might be expected that the meaning relationships between words will influence order in terms of the *adjacency* or *relative distance* between them, even if they do not predict a certain order. Although the details of what these “relationships” constitute vary markedly between theoretical frameworks, grammarians and linguists from all theoretical perspectives are at least in agreement that certain words combine with, or depend on some other words in the sentence more directly than the remainder. This can be seen most clearly in dependency grammar formalisms stemming from work by Tesnière (1959) (e.g. Hudson, 1984; Sgall et al., 1986; Sleator and Temperley, 1993; Järvinen and Tapanainen, 1998). Dependency grammars explicitly represent the pairs of words which *depend on each other for interpretation*. In the simple sentence “*Andrew likes cake*”, *Andrew* has a dependency on *likes* because Andrew is the subject of *likes* (if dependencies are considered to be syntactic) or its experiencer (if dependencies are considered to be semantic). Likewise, *cake* is the object or theme of *likes*. The words *Andrew* and *cake* are not directly dependent on each other, and are only related indirectly through the verb. Given three words related in this way, then, the six logical ways in which a language could order them result in the dependency graphs shown in Figure 3.1.

Critically, the total length of the two dependencies can vary depending on the order of the words: four of the possible orders above include a long dependency, while the other two orders (the X V X column) only have dependencies between adjacent words. Of course, this is true for more complex sentences as well. A great deal of

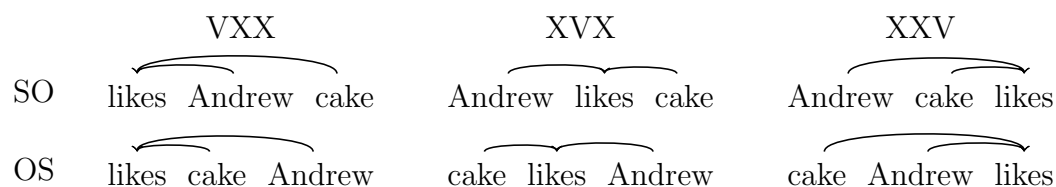


Figure 3.1: The logically possible orders of a three-word dependency structure.

previous research in psycholinguistics has demonstrated a link between longer dependency lengths and increased processing complexity. In comprehension, processing complexity manifests itself as slowed comprehension or failure to understand, and in production it manifests itself as avoidance of difficult structures, and in some work (though perhaps none dealing with dependency lengths in particular) as speech errors.

The first formal models of human parsing to capture the idea that long distance dependencies incur processing difficulty appeared a full half-century ago. Yngve (1960) proposed a model which equated human memory resources with the depth of a parser stack containing words being held for later processing. One of the facts that motivated this model is the increased difficulty of center-embedded sentences like (19a) over the meaning-equivalent (19b):

- (19) a. This is [the malt that [the rat that [the cat] killed] ate].
- b. This is [the malt that was eaten by [the rat that was killed by [the cat]]].

Yngve's memory-based explanation rested on the fact that as the comprehender progresses serially through the sentence in (19a), items corresponding to *malt* and *rat* must be held on the stack until the verbs they depend on (*ate* and *killed* respectively) are met. In (19b), each noun is positioned close to its verb. Yngve hypothesized that the maximum number of items that must be held on the stack was a sentence-wide measure of processing complexity and hence difficulty. By annotating these example sentences with dependency links as in (20), it can be seen that the length of the longest dependency extending from a word to the right corresponds directly to the amount of time a word must be kept in memory, and the number of arcs that run concurrently at any point is related to (though not exactly the same as) the number on items on the stack.

- (20) a. This is the malt that the rat that the cat killed ate
- b. This is the malt that was eaten by the rat that was killed by the cat
-

Yngve postulated that the maximum depth that could be processed would be determined by general cognitive constraints, inspired by Miller’s (1956) famous “seven plus or minus two” limit on memory capacity. He also understood the relevance of processing complexity to typology, hypothesizing that “the grammars of all languages will include methods ... so that most sentences will not exceed this depth”, and “when languages change, [stack] depth phenomena will frequently be involved, and will often play an important role.”

Another class of sentences which have provided evidence for memory-based theories of processing difficulty are those containing relative clauses (RCs). It has been observed that “subject-extracted” RCs (SRCs, in which the noun being modified is understood as the subject of the relative clause, as in (21a)) are easier to process than “object-extracted” RCs (ORCs, as in (21b)). The dependency annotations in (21) show that the only difference in dependency lengths in these sentences occurs within the RC (shown in square brackets) and results in longer dependencies in the ORC case.³

- (21) a. The reporter [who attacked the senator] admitted the error
- b. The reporter [who the senator attacked] admitted the error
-

³Here and elsewhere, the particular theory of dependency relations being used may affect the dependencies in minor ways. For instance, Gibson (2000) would assume that *attacked* has no direct dependency on *reporter*, but on the word *who* instead, which he takes as a pronoun that is coreferential with the immediately previous noun. These subtleties make no difference to the qualitative predictions of the theory.

In order to quantify the observation that longer dependency lengths lead to comprehension difficulty, Gibson proposed a simple word-counting metric as part of his *Dependency Locality Theory* (DLT: Gibson, 2000), itself a revision of the earlier *Syntactic Prediction Locality Theory* (SPLT: Gibson, 1998). DLT proposes a simple word counting measure of memory-based-complexity where the difficulty of “integration” at each head word is a function of the words that intervene between it and a preceding dependent.⁴ Because DLT predicts that processing complexity will arise at the ends of dependencies specifically, it can be tested with online experiments on reading times at the word level. Accordingly, repeated studies have shown that reading times are slower in the ORC than in the SRC at the embedded verb (e.g. *attacked*) specifically (e.g. King and Just, 1991; Holmes and O’Regan, 1981; Ford, 1983; Grodner and Gibson, 2005, i.a.).


In fact, DLT does not count all words that intervene between the ends of a dependency as counting towards overall difficulty, but only those which introduce *new discourse referents*: typically, lexical nouns or verbs. Thus, in (21a), the words *reporter* and *attacked* are effectively adjacent, making that dependency’s “cost” 0, while in (21b), the word *senator* intervenes, making the corresponding cost 1. Obviously, if all words were counted, the difference would be qualitatively similar: *who* in the first case for a cost of 1, or *who the senator* in the second for a cost of 3. The choice to only count particular words was intended to capture the observation that sentences like (22b) are rated as harder to understand than (22a), presumably because the pronoun *I* incurs less processing load than a full lexical noun phrase (Warren and Gibson, 2002).

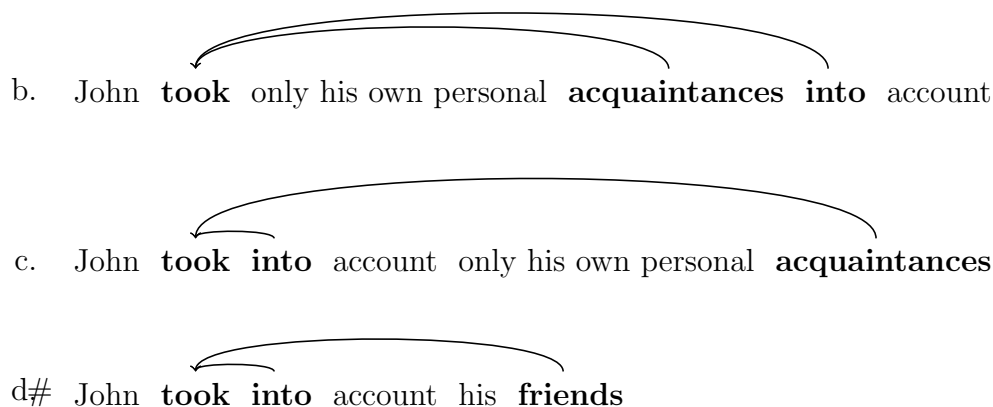
- (22) a. The reporter who the senator who I met attacked disliked the editor.
 b. The reporter who the senator who the professor met attacked disliked the editor.

⁴Strictly speaking, the cognitive load experienced at the end of a dependency is only one of the memory demands proposed by DLT, the “integration cost”. Like its predecessor SPLT, DLT also includes a “storage cost” for incomplete dependencies. As Gibson (1998, p19; 2000, p102) notes, the integration cost alone accounts for most of the relevant findings discussed in the literature, and it is this part of the theory which has received the most subsequent empirical support and been the most influential.

Warren & Gibson concluded that the pronoun *I* led to less processing difficulty because the pronoun refers to a given and highly accessible referent. Subsequent research has refined this picture somewhat, and called into question whether givenness is the most relevant property, rather than some measure of confusability between the referents (e.g. Gordon et al., 2001; Gordon, 2004) or the higher frequency of pronouns in embedded subjects in spontaneous language (e.g. Real and Christiansen, 2007; Genari and MacDonald, 2008). In any case, the empirical literature to date on length as an influence on ordering in *production* has shown that phrase length in words is a useful measure for predicting production choices, regardless of whether the accessibility of the discourse referents within the phrase drives that effect (e.g. Wasow, 2002; Szmrecsányi, 2004; Temperley, 2007). Most of the relevant studies discussed here are on production rather than comprehension, and as corpus studies like the one in Chapter 2 look at aggregates over large samples of language, I will simply count the number of words each dependency crosses, which yields a metric that is well correlated with DLT (and with Hawkins’ MiD, discussed below). In fact, measures like DLT are likely highly simplified proxies that happen to reflect a more articulated cognitive process underlying comprehension and production. For instance, Lewis and colleagues (e.g. Lewis and Vasishth, 2005; Lewis et al., 2006) have implemented a detailed activation-based model of the memory system underlying online sentence comprehension which models the activation and decay of memory traces associated with words, and makes predictions for comprehension difficulty which are highly correlated with those of the simple DLT metric. This model has now been used successfully to explain patterns of comprehension difficulty across diverse structures in several languages (e.g. Japanese: Nakayama et al., 2006; Hindi: Vasishth and Lewis, 2006).

Strictly speaking, DLT is a theory about comprehension: it predicts the amount of difficulty that a comprehender will experience at a certain point in the sentence. However, it is very easy to apply the same principles of dependency length minimization to choices made in production. Consider the sentences in (9), repeated here with dependency annotations.

- (23) a. John  took his friends into account



Choosing the noncanonical order results in a much shorter dependency length when the object is long (23c vs b), but with a short object, there is no difference in the total length (23d vs a). A noncanonical order is therefore chosen in production when it substantially reduces dependency lengths, but not otherwise.

As discussed in the introduction, the most influential work on the relationship between processing and grammar in recent years is due to Hawkins (e.g. Hawkins, 1994, 2001, 2004, 2007), so it is worth taking a moment to relate Hawkins' representational framework to the DLT framework discussed in more depth here. Instead of a dependency formalism, Hawkins assumes a hierarchical phrase structure grammar: that is, one in which a sentence can be broken down into nested phrases, each of which has a representational status independent of the words within them. In Hawkins' framework, however, certain words have the privileged status of being "constructing nodes", meaning they unambiguously signal the presence of the constituent containing them. For instance, a noun or determiner could signal the presence of a noun phrase, and a preposition the presence of a prepositional phrase. In Hawkins' theory, once a constructing node has been encountered in a string of words, its associated constituent is "constructed" immediately, and must be maintained (presumably in some memory buffer, though the theory is deliberately agnostic on implementation details) until all its sister constituents have also been constructed, allowing the structure of the mother constituent to be finalized. The gap between the point at which the first and last child of a given mother constituent are constructed is called the *phrasal combination domain* (PCD). Hawkins discusses a set of theorized "parsing principles",

which summarize preferences seen in language production and comprehension. Most relevant here is his *Minimize Domains* principle (MiD), which states that this PCD should be made as short as possible. Intuitively, the length of the PCD could be equated with the taxation on cognitive resources such as memory in either the producer or comprehender: once the first daughter of a constituent is “constructed”, the constituent must be kept in some kind of working memory until its structure is finalized; any words in the PCD must therefore be produced or comprehended in parallel, without the use of whatever resources are taken up by maintaining that constituent. Broadly speaking, this theory makes the same predictions as the dependency length theory, since longer PCDs equate to longer dependencies between words. One place in which the predictions of the two theories diverge is in cases where there are more than two dependents of a single head word on the same side. Both theories predict that the longest dependent should be placed furthest from the head. However, while the PCD size is unchanged by the order of any other dependents, the dependency length theory predicts that all phrases should be ordered by length, since it proposes that word orders are disfavored in proportion to the sum of *all* dependency lengths. Temperley (2007) uses this fact to discriminate between the two theories, showing that in English clauses with three postverbal adjuncts, the first tends to be shorter than the second. Temperley also discusses the fact that for head-final languages like Japanese, the PCD is shorter for a long-before-short ordering only by virtue of the fact that Japanese case particles come at the end of phrases. By assumption, it is these particles which are the privileged “constructing nodes” that signal the beginning of the PCD, despite the fact that the presence of an argument is probably evident much earlier. It is unclear why the parser should wait until the end of a long noun phrase to begin the construction of the verb phrase that will contain it, given that Hawkins’ other principles consider processing that happens as early as possible to be the most efficient. In fact, Fine (2006) shows that the same long-before-short preference holds in the verb-final subordinate clauses of German, despite the fact that German arguments typically begin with prepositions or case-marked nouns, and hence the PCD must extend over the entire verb phrase regardless of the order of its constituents.

For these reasons, and because the dependency length theory is slightly easier

to illustrate, I will stick to a dependency framework throughout here. Nevertheless, Hawkins' principles seem to make the same predictions in the majority of cases, and nothing in this dissertation should be taken as support for one of these theories over the other. One way in which Hawkins' and Gibson's expositions of their theories differ is that while Gibson (a psycholinguist) states DLT as a theory of working memory in comprehension, Hawkins (a linguist) does not make any specific claims about the cognitive mechanisms which underlie the observed processing preferences. Hawkins avoids basing his principles in a theory of working memory (see e.g. Hawkins, 2004, p27), preferring to state MiD as a theory of general processing "efficiency" without being more specific about how or why domain minimization or early recognition of constituents is efficient, and makes use of evidence from both production choices and comprehension difficulty (though perhaps more from production). There is much more work that could be done on the relationship between production choice and comprehension difficulty. However, any production effects that can be explained by MiD are equally well accounted for by a dependency length theory, as explained above, so this difference in exposition is not a reason to prefer one theory over the other.

3.1.3 Prosodic theories

Before turning to an empirical investigation of the dependency length theory, I will briefly mention another class of production theories which, while not conventionally considered psycholinguistic, attempt to account for similar facts and may make similar predictions. These theories suggest that a major component of word order choice is pressure to preserve prosodic well-formedness, by distributing stresses in accordance with universal or language-specific principles.

Anttila et al. (2010) propose that in English, the choice of words in the ditransitive alternation and heavy-NP shift (both discussed above and in chapter 2) is partially predictable from the lexical stresses in the phrases. For present purposes, the most relevant aspect of their model is the suggestion that lexical stresses prefer to appear in the position which receives sentence-level stress (in Present Day English, the rightmost constituent), and hence phrases which contain more stresses should be placed finally.

Naturally, this captures the End Weight generalization. Given that sentence-level stress in Old English is understood to be final as well (e.g. Speyer, 2008), this extends to the results described in chapter 2.

Although the details of Anttila et al.’s analysis are out of the scope of the present discussion, it is worth pointing out that their model is attractive in that it offers a single framework within which other prosodic influences on word order, and the interaction between them, can be stated. For instance, the Optimality Theory implementation they present additionally captures differences in the ditransitive alternation between rhythmically one-foot verbs (e.g. *give*), which appear with both theme-recipient and recipient-theme orders, and two-foot verbs (e.g. *donate*, *convey*), which take the theme-recipient order frequently or all of the time. Like the dependency length theory, it also offers the ability to account for leftwards movement of heavy material, assuming that in languages where that pattern is observed, sentence-level stress is placed initially.

The stress-counting theory makes certain concrete predictions which may differentiate it from other theories, and which deserve further investigation. One prediction, which derives from the Optimality Theory implementation of their model, is that there is a single, discrete pressure to place the heaviest phrase (in terms of stress count) in the sentence stress position. This contrasts with the dependency minimization theory, which predicts that ordering preferences should reflect the gradient difference in weight between multiple phrases, and that in cases with more than two dependents to be ordered, the preferred outcome should order them by increasing length. Jäger and Rosenbach (2006) have shown that in the case of the English possessive alternation, order choices do reflect a continuous measure of difference in weight, rather than just a binary notion of “most heavy”. Similarly, as discussed above, Temperley (2007) has shown that English clauses with three postverbal dependents do seem to order all three according to their weight, rather than simply placing the heaviest last. However, Anttila et al. (2010) point out that previous work has assumed a word or syllable counting metric rather than using lexical stresses, so it is quite possible that these results will be falsified when re-tested using the prosodic definition. Even if that does not turn out to be the case, an alternative, gradient implementation of the

prosodic model could be explored.

Independently of whether the prosodic explanation is in general correct, Anttila et al. suggest that counting lexical stresses yields a superior metric to simply counting words. If this was the case, it could undermine the psychological basis for the dependency minimization theory as an artifact of human memory resources: there is no obvious reason why stressed syllables specifically would be more prone to interference in memory or otherwise affect recall or integration. At present, there is little evidence that counting stresses, words, or discourse referents makes much difference to the predictive power of any length or weight-based theory. A fundamental problem with testing whether unstressed words do or do not count for the purposes of weight is that in natural language, the number of stresses in a phrase will tend to be highly correlated with its length in words. This makes it hard to tease the two apart. Moreover, most unstressed words are pronouns and function words, which would also be excluded under the DLT metric that ignores words which do not introduce new discourse referents. Therefore, it seems possible that this part of the theory cannot be adequately tested, at least in English.⁵

3.2 Language-wide dependency lengths

Figure 3.1 makes it clear that the total length of all dependencies in a singly-rooted sentence is minimized when the root (the word on which all other words are dependent) is in the middle. It is easy to think through what would happen when additional dependents on this word are added: the additional length each one incurs increases

⁵Anttila et al. do present a statistical model in which they simultaneously enter covariates of phrase length in words and stresses into a logistic regression predicting the ditransitive outcome in a written corpus. They find the stress counts to have higher absolute coefficient sizes than word counts, which might suggest that stresses may carry the effect. However, that analysis lacks many other controls known to be relevant, and does not present goodness-of-fit statistics or model comparisons which would indicate how well each predictor actually accounts for the data: it remains possible that a model containing only length in words would predict the outcomes in their dataset as well as one containing only length in syllables. This is an especially worrying possibility given the potential for collinearity problems in that model, which are apparent from the VIF statistic reported and counter-intuitive direction of the effects of the correlated predictors.

with the number of other dependents on the same side of the verb only, so the minimum total length of a singly-rooted structure is always achieved by putting half of the root's dependents on its left, and half on its right. If one of the dependents is not a single word but a longer unit (ignoring for the moment any internal dependencies that unit may have) then all other dependents which are on the same side of the root but further away from it must extend over that additional length. Therefore, total dependency lengths are minimized when the roots children are ordered shortest-to-longest, beginning closest to the root and moving away.

These principles of dependency layout were used by Gildea and Temperley (2007) to construct an algorithm that takes a dependency graph connecting a set of words, and finds a serial ordering of the words which results in the shortest possible total dependency length. Gildea & Temperley's is a dynamic programming algorithm, relying on the fact that the words within any subtree of the dependency graph can be given an order which is optimal independent of the order of words outside that subtree, except for the single link which connects its root to the rest of the sentence. For details of the procedure, see that article. Gildea & Temperley applied their algorithm to the Wall Street Journal (WSJ) corpus of English newswire text, finding the minimum possible total dependency length for the dependency structures in each sentence (that is, ignoring the grammar of English and rearranging words so that the dependencies between them are as short as possible). They found that the average total minimum dependency length for a WSJ sentence was 33.7 words, while the average *actual* dependency length of the sentences as written was 47.9. This was compared to a baseline where the order of words was simply randomized, yielding an average total length of 76.1 (note that this random baseline is still much less than would be expected of a "worst possible" strategy of trying to *maximize* the total dependency length). These figures are presented in Figure 3.2. See also Temperley (2008) for a more detailed exploration of the average dependency lengths produced by varying types of formal grammar.

The ordering of words in actual English therefore yields dependency lengths that are much closer to the minimum possible than would be expected by chance. This suggests that some kind of pressure for dependency minimization may have played a

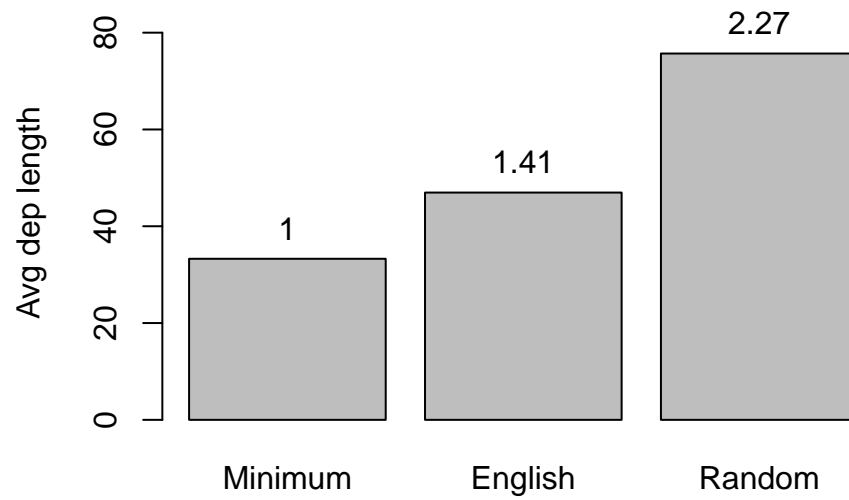


Figure 3.2: Mean per-sentence dependency lengths from Gildea and Temperley (2007). Numbers show length ratios relative to the theoretically possible minimum.


role in the evolution of languages like English. Gildea & Temperley also note that languages like English have grammars in which the order of words is partly fixed, and in fact encodes meaning (“*man bites dog*” and “*dog bites man*” have different meanings). Because of this, it might be unrealistic to expect English to be able to adapt perfectly to yield minimum lengths in every case. They also report the average dependency length given by a new grammar for the WSJ sentences which yields an approximately minimum dependency length under the constraint that a certain head-dependent type (such as verb-subject) must always be serialized in the same direction. This grammar yields an average total length of 42.5, closer still to the actual average length of WSJ sentences.

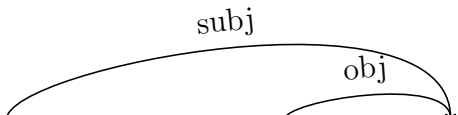
From the point of view of the dependency length theory, then, the grammar of English is a remarkably efficient solution to ordering words for easy processing by human language users.


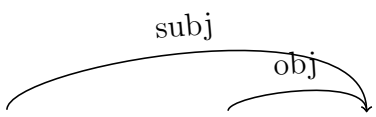
3.2.1 Dependency lengths in Historical English

Given that dependency length minimization apparently influences, or at least predicts the word orders in sentences that modern language users produce, it is not surprising that the same generalization could predict the choices of users of Old and Middle English (henceforth OE and ME). If this phenomenon represents a processing bias, then we would expect the human language users of just a few centuries ago to experience the same bias. In OE and ME, both VO and OV orders were available. As subjects were typically preverbal (and other preverbal dependents common), pressure to minimize dependency lengths might have led speakers to place the object after the verb more often if it was long, therefore avoiding long dependencies between the verb and subject or other preverbal material. The following clauses of OE illustrate dependency lengths for the SVO (24,26a) and SOV (25,26b) orders.⁶

⁶In a co-ordination of two OE clauses like (26), it was common for the first clause to be SVO and the second SOV; however, this is not an absolute rule, and neither SVO nor SOV clauses are limited to co-ordinations like this. Also note that the sentences in (24,25) contain auxiliary verbs, and although I have simplified the dependency structures by assuming the subject is directly dependent on the main verb, the length asymmetry would be identical if the dependency chain ran subject → aux → verb.

- (24)  (SVO)
 Se wolde gelytlian pone lyfigendan hælend
 he would diminish the living lord
‘He would diminish the living lord’
 (*Ælfric’s First Letter to Wulfstan*, c1050, quoted in Pintzuk and Taylor, 2004)

- (25)  (SOV)
 Ac he sceal pa sacfullan gesibbian
 but he must the quarrelsome reconcile
‘But he must reconcile the quarrelsome’
 (*Ælfric’s First Letter to Wulfstan*, c1050, quoted in Pintzuk and Taylor, 2004)

- (26) a.  (SVO)
 Unaberendlic gyhpa ofereode ealne pone lichaman
 unbearable itching overcame all the body
‘Unbearable itching spread over his entire body...’
- b.  (SOV)
 & ungelyfendlic toblawennys his innop geswencte
 and unbelievable bloating his innards afflicted
‘... and unbelievable bloating afflicted his stomach.’
 (Herod’s last days as described in *Ælfric’s Homilies*, c1000)

Clearly, as the object becomes longer, the subject dependency in the SOV clauses will become longer. On the other hand, neither the subject nor the object in the SVO clauses spans the other, meaning that both are always close to the verb. Of course, pressure to place a longer object after the verb would equally well be explained by a simple “End Weight” explanation in which heavy material is placed later in the sentence.

3.2.2 Left dislocation

Nothing in the corpus study reported in Chapter 2 distinguishes between the dependency minimization theory and the simpler principle of putting longer constituents later. However, previous work by Traugott (2007) may bear on the issue. Traugott investigated OE left dislocation structures like the following, where rather than using a full object noun phrase within the clause, the writer has placed the object at the beginning of the sentence, and refers back to it with a pronoun afterwards:

- (27) **ealle pa Romaniscan men** pe Hannibal on Crece
 all those Roman men whom Hannibal into Greece
 geseald hæfde, **him** bebead se consul pæt
 sold had them commanded the consul that
 hie eal hiera heafod bescearen.
 they all their heads shaved
 ‘*The consul commanded all the Roman men whom Hannibal had sold as slaves*
 in Greece to shave their heads.’
 (*Orosius*, c925, quoted in Traugott (2007))

Traugott notes that the majority of these sentences occur with a long and complicated NP: in 49 (94%) of the 52 object left dislocations she finds, the object is followed by an adverbial or relative clause. The tendency for a particularly long object to be followed by a short pronoun bears some similarity to the long-before-short preference reported for modern OV languages like Japanese (Hawkins, 1994; Yamashita and Chang, 2001). This pattern arguably follows from the dependency minimization hypothesis: in (27), the object pronoun is adjacent to the verb, and so there is a short object-verb dependency. If the dislocation construction had not been used, that dependency would extend over the entire relative clause (*pe Hannibal ... hæfde*). Of course, if the only pressure was to place heavier material later in the sentence, then there would be no reason to dislocate long objects more than short ones. I say that the dependency theory “arguably” predicts this result, because it isn’t clear what the dependency structure relationship between the dislocated material and the rest of the sentence is. If *men* in (27) is a dependent of any word in the main clause (say, *bebead*

or even *him*), then including the pronoun in fact increases the overall dependency length of the sentence. In any case, this is a long-before-short preference of a sort.

Unfortunately, object left dislocation structures are relatively rare, so it is difficult to investigate the role of weight in the limited OE data in any more detail. However, in a quantitative analysis of left dislocation in spoken Present Day English, Snider and Zaenen (2006) found a significant effect of the length of the object: longer NPs are more likely to be left dislocated than short ones.

3.2.3 Measuring the change in dependency lengths

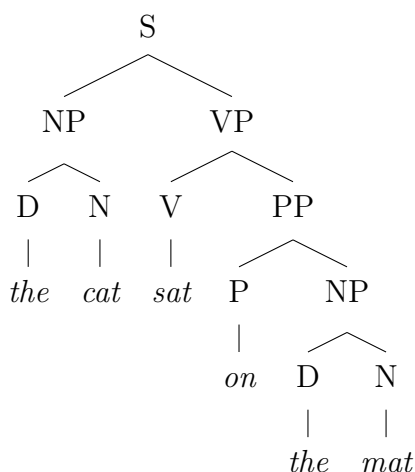
If some kind of processing-driven pressure to make dependency lengths short has influenced the evolution of English, then we might expect to find evidence of this pressure by looking at earlier stages of English grammar. A simple first hypothesis is that Old English was less dependency length optimal than Present Day English, and therefore that the average total dependency length of sentences should decrease over time as the grammar is restructured. This hypothesis seems overly simple, however: language change is likely influenced by a wide variety of historical and cognitive factors, and unless the pressure for short dependency lengths is remarkably strong, its influence on a gross measure like “average total sentence dependency length” may be overwhelmed by other more immediate factors. Potentially, both sudden linguistic changes (such as the borrowing or coining of new words and constructions) and gradual linguistic changes (like a shift in the frequency with which forms are used, or the reanalysis of multiword sequences as morphologically complex words) might be “filtered” through the general preference for shorter dependencies, such that those forms which gain hold in the language are those with the comparative advantage in dependency length. This said, if dependencies do tend to get shorter on average, the historical English dataset described in Chapter 2 covers a long enough period that we might expect to see some kind of trend.

I used all the parsed sentences contained in the two corpora used in Chapter 2, the YCOE and PPCME2. These sentences are parsed with a constituent grammar representation, as illustrated in Chapter 2. Obviously, in order to calculate dependency

lengths, this representation needs to be converted into a dependency structure. This is not a particularly difficult procedure, once the concept of a syntactic “head” is defined: this is the one word per constituent on which all other words in the constituent are directly or indirectly dependent. The head itself is not dependent on any other word within the constituent. We define the head of a terminal word to be itself. The head of a constituent is defined recursively to be the head of its “head child”, one of its children which is given special status. Then, creating the dependency structure is a simple matter of iterating through each word or constituent and drawing a dependency arc from the head of each one to the head of its mother (excluding the case where a given word *is* the head of its mother).

Once the head is known, there is sufficient information in the constituent parse to determine the dependency tree. The main difficulty, then, is deciding for each constituent which of its children will be taken as the head child. To do this, I used a simple “headfinder” algorithm based loosely on the set of rules given in Collins (1999). These rules are hand-written and determine case by case which child to choose for a given phrase type. For instance, the head child of a noun phrase is the rightmost common noun or pronoun if there is one; otherwise the rightmost embedded NP if there is one; otherwise the rightmost (WH) question word, number, determiner or adjective; otherwise the last word. This procedure is heuristic, in that it is not guaranteed to find the correct head, but seems to work very well in practice. As an example, consider the following parse:

(28)



The NP headfinder rules tell us to select the rightmost noun, giving *cat* and *mat* as the head of the two NPs. The PP rule selects the adposition *on* as its head, and the VP rule selects the verb *sat*. Finally, the S rule selects the VP as its head child, so we set the head of S to be equal to the head of VP, *sat*. Drawing arrows from each word and from the head of each constituent to the head of its mother yields the expected dependency graph:



Having converted both corpora into a dependency format, I was then able to calculate the length of each dependency, in words. I excluded any sentences that were longer than 80 words in length (just under 0.2% of the data). Taking the sum of all dependency lengths in each sentence then gives the quantity of interest, which is expected to decrease over time. These values are plotted in Figure 3.3. Each manuscript is plotted as a circle whose area reflects the number of sentences within it (though very short manuscripts are given a minimum size to keep them visible). The y-axis reflects the average total dependency length of all sentences in the manuscripts. Most importantly, the thick line shows the overall trend over time by applying a loess smoother to the data. A loess smoother is essentially a combination of many “local” linear regressions, each focused on one region (here, a certain date range) in the data. By combining the straight lines of best fit for each of these local regressions, we get a smooth curve which illustrates gradual change over time.

Figure 3.3 does seem to indicate that there is some decrease in the total sentence dependency length over time, although it is quite slight relative to the variation even between per-manuscript averages (and presumably extremely small relative to the variation between individual sentences). To try to reduce this unexplained variation and make any change in average length more obvious, I tried a second way of computing the results. This time, I used the algorithm in Gildea and Temperley (2007) to calculate the word order in each sentence which yields the *minimum possible* total dependency length. For each sentence, I then divided the actual total length by the theoretical minimum. This gave a ratio which can be interpreted as an “efficiency ratio”: a measure of how close the actual word order for this sentence gets to the best

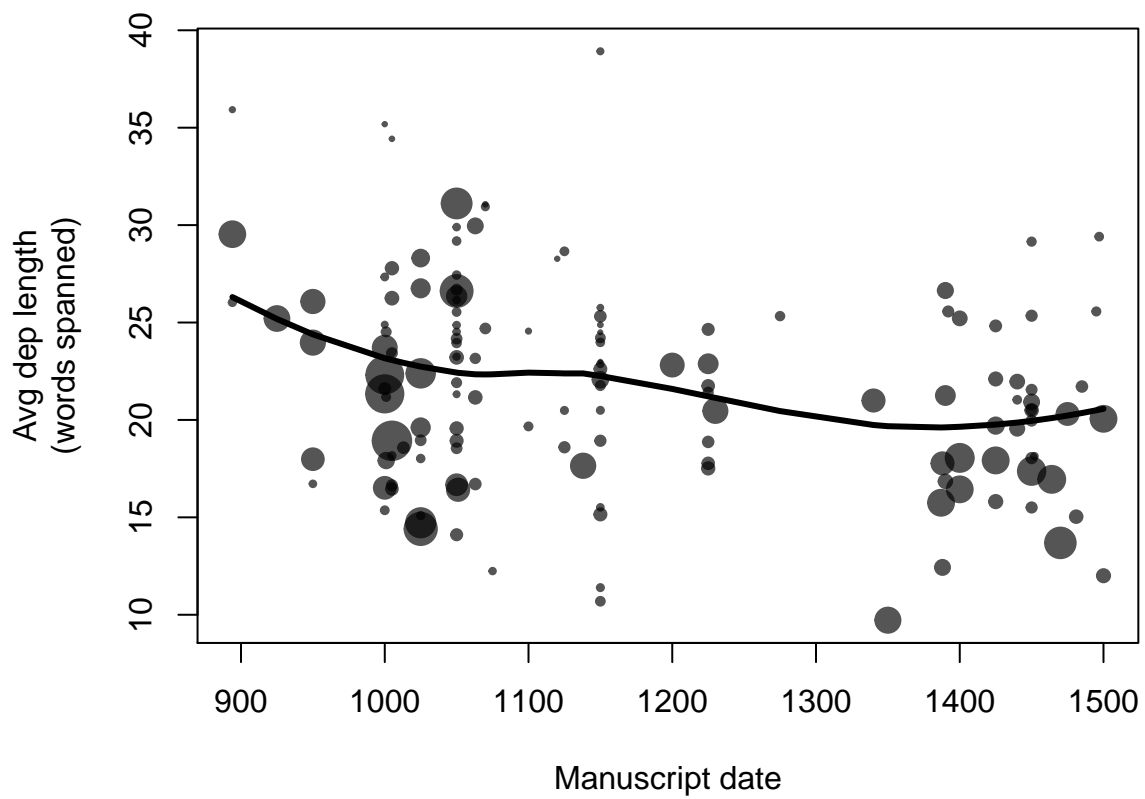


Figure 3.3: Mean total dependency length for each manuscript (raw values)

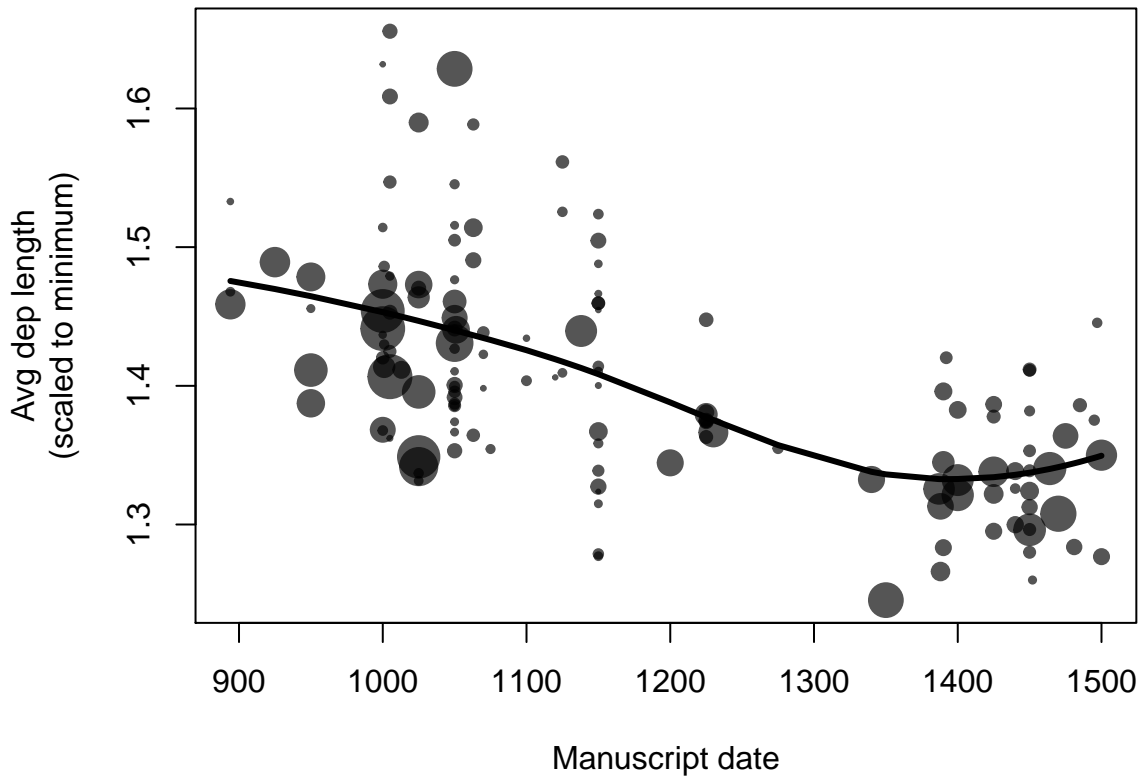


Figure 3.4: Mean total dependency length for each manuscript (relative to theoretical optimum)

possible order. This measure has an advantage over using raw total lengths, because it controls to some extent for the fact that some sentences will unavoidably have higher dependency lengths. Sentences that are simply longer, or have more words with multiple dependents, for instance, will tend to have longer dependencies than short simple sentences. Plotting this measure in the same way as before yields the plot in Figure 3.4.

Having controlled for differences between sentence length and structure in this way, the trend towards shorter dependencies over time is much more apparent. In the earliest manuscripts, dependencies are on average just under one and half times the minimum length they could possibly be. By the end of the Middle English period, this inefficiency has been reduced to about 1.35 times the minimum. This figure is comparable to Gildea & Temperley’s finding that the dependencies in modern WSJ

sentences are on average about 1.4 times the minimal length they could be.

3.2.4 Word order freedom and morphology

The results described above suggest that the configuration of Old English sentences led to words being separated from their dependents by a greater distance, on average, than in Middle or Modern English sentences. If, as the dependency minimization theories suggest, there are processing costs associated with long dependencies, then this change represents movement towards a more processing-optimal state. However, if long dependency lengths are always inefficient, and languages evolve towards the optimal, it is unclear why Old English — or indeed, any language — should ever exist in such a suboptimal state.

One possible answer lies in other changes that were occurring during the Old English period. OE, like Modern German, had a rich inventory of cases, and number and gender morphology, all of which served to clarify the role of words within a sentence and the dependency structure that connects them. If the memory based theories that underly the dependency minimization principle are correct, then these additional clues as to which words need to be recalled and integrated at a given point in the sentence could reduce interference from other words, or otherwise mitigate the difficulty associated with long dependencies. I will return to this point in more detail in section 4.4.1. For the moment, I will summarize some evidence that this may be the case based on work on dependency length optimality in a language with rich morphology .

Old English had relatively more word order freedom than either Middle or Modern English. This is in accord with a general typological pattern that case-poor languages tend to have relatively fixed word order. Intuitively, when word order is all that signals grammatical function, it may become harder to manipulate word order in a given sentence and still express the intended meaning. Modern German is similar in allowing several possible orders of the major constituents, with grammatical function signaled by case. Both Gildea and Temperley (2010) and Park and Levy (2009) have compared modern English and German for their average dependency lengths using

slightly different methods similar to those reported here. Interestingly, both papers show that although English is much closer to the optimal configuration than to a random ordering baseline, as described above, German (while still more optimal than the baseline) is much less so. Park and Levy hypothesize that this may be due to German's greater morphological richness, pointing out that in psycholinguistic experiments such as those reported by Konieczny (2000), short dependencies do not necessarily lead to reduced difficulty in German, and in some cases the opposite may be the case. Similarly, Gildea and Temperley argue that the relatively long dependencies of German may arise from German's ability to place heads and dependents relatively far apart without fear of ambiguity. They also briefly discuss the fact that in German, as in Old English, word order may be harnessed for other functions than simply expressing grammatical function; chiefly, expressing information structure.

A further piece of evidence suggesting that German may not be subject to the same pressure for dependency minimization as English comes from the fact that German — like OE — has crossing dependencies far more often than PDE does (see e.g. Levy, 2005, and references therein). Dependencies between words in many languages tend not to cross each other, and this fact is often taken as a hallmark of linguistic structure, despite the existence of some exceptions (see summary in Pullum and Gazdar, 1982). Ferrer i Cancho (2006) suggests that nested, non-crossing dependencies are a probable outcome of pressure to keep total dependency length small. Ferrer i Cancho shows that after applying an algorithm designed to reorder a sentence to reduce its summed dependency length, both randomly generated and actual linguistic sentence structures display a dramatic reduction in the number of their dependency links which cross. This relationship between non-crossing dependencies and short dependencies indicates that we could use the frequency of crossing dependencies in a language as a heuristic estimate of the extent to which a language's grammar has been shaped by a pressure to minimize dependency lengths. By this logic, both modern German and Old English show less evidence of minimization pressure than Present Day English does.

If this hypothesis is correct, then the relatively long dependency lengths of early Old English may have caused little processing cost as long as morphological cues were

available to its speakers. However, as inflectional paradigms collapsed, longer dependency lengths would have incurred more and more processing complexity, perhaps leading to pressure towards change. I will pick up this point again in 4.4.1.

3.3 Summary

The fact that longer objects in historical English tended to favor VO word order can be explained by multiple psycholinguistic theories. One, conceptual accessibility, predicts that longer phrases should appear later in the sentence. Another, dependency length minimization, predicts that words are ordered to make the dependencies between them as short as possible. This implies that longer dependents should be further from their heads than shorter ones. Although there is substantial cross-linguistic evidence that conceptual accessibility influences word order, findings from head-final languages show that phrase length effects do not always pattern with other components of accessibility. Nothing in the corpus study presented in Chapter 2 differentiates between these two theories, but some limited evidence from length effects in object dislocation suggest that dependency minimization may end up better able to account for the word order preferences observed in Old and Middle English.

If English changed over time to become more processing-optimal, this change should be reflected in the textual record. The dependency length minimization theory therefore predicts that over time, dependency lengths in English sentences should become shorter on average. Calculating those lengths directly for all sentences in the YCOE and PPCME2 corpora shows the expected effect.

Chapter 4

Weight as a causal factor in word order change

In Chapter 2 I presented evidence that the word order of Old and Middle English clauses was influenced by the length of the object phrases within them. In Chapter 3, I described a psycholinguistic theory of processing complexity based on the length of dependencies between words which predicts this influence on word order, and showed that the language-wide word order changes in English resulted in clauses which are less complex to process by this theory. In this chapter I discuss the possibility that length-based processing effects could play a causal role in word order change, and present a simple model exploring this.

4.1 Processing effects in word order change

4.1.1 Variation and change

Variation in synchronic language and gradual replacement of one form with another in diachronic language are often studied as two sides of the same coin, for obvious reasons. Labov (1982, p20) states “change implies variation; change *is* variation”. In Labov’s view, language change is best understood as the propagation of forms throughout a speech community rather than the initial coining of a new form or

loss of an old one specifically. However, when one particular way of expressing some meaning increases in frequency, it is likely to be at the expense of some other existing way the language provides to say the same thing. Therefore, during the time a certain linguistic form is spreading throughout a speech community and becoming increasingly frequent, there is an alternation between the “incoming” form and any previously existing ways of expressing an equivalent meaning. In some cases, the incoming form may end up completely dominating, making any other forms become so infrequent that they eventually become lost from the language altogether. This was the case for the former verb-final sentence templates of Old English: over time, they were used less and less frequently until at some point speakers no longer recognized them as part of the language. However, at any time during the change, individual speakers would have commanded and used both variants (Weinreich et al., 1968). Pintzuk (1999, 2002, 2005) argues that this is also the case for word order change in Old English, and that individuals had command of both basic VO and OV grammars simultaneously.

The historical alternation, then, is amenable to study in the same way as modern alternations are. Although we don’t have access to speakers, and so can’t conduct interviews or experiments, we do have a body of texts which can be analyzed statistically to determine what properties led to a speaker (or writer) using one form or the other. This kind of statistical analysis of variation in corpora has long been used in sociolinguistics, but is becoming increasingly influential in other subfields. In psycholinguistics, it can sometimes shed light on what factors govern a language user’s production choices more effectively than controlled lab experiments thanks to the large amount of data that is available and the naturalistic form of the utterances themselves. Recent examples of work of this kind include Jaeger’s (2006) study of English *that*-omission (e.g. “*the family (that) I cook for*”; “*I told you (that) I did it*”), and Bresnan and colleague’s (2007) study of the English ditransitive alternation (e.g. “*give NP1 to NP2*” vs “*give NP2 NP1*”). Since these are binary outcomes between two possibilities, these researchers used heavily annotated datasets to train a logistic regression to discriminate between the two outcomes given properties of each such sentence. Both studies found that speakers’ choice between the outcomes is

influenced simultaneously by a large number of properties, including (sociolinguistic) properties of the speaker, properties of the preceding discourse context, properties of the intended meaning, and purely linguistic properties of the words and phrases involved. The resulting regression models could then be inspected to determine the exact nature of the influence of each such property.

The diachronic generalization of these kinds of variation is straightforward: at any moment in time, speakers exhibit *conditioned* variation just like in modern languages: therefore, a single speaker of Old English may produce a VO clause more or less frequently depending on whether its object is a pronoun, as shown in chapter 2. However, the “base rate” of the outgoing form will decrease over time, with case-by-case variation being additive on top of this global shift in frequency. In fact, Kroch (1989) has argued that one of the hallmarks of linguistic change is the so-called *constant rate* effect. This holds that a certain form (say, VO) may be more frequent in certain environments (say, clauses with non-pronoun objects) but always by exactly the same amount as a ratio of odds. For instance, suppose that in an early period VO is more likely than OV with a nonpronoun object by a ratio of 3:1 but with a pronoun object by a ratio of only 2:1. Thus the nonpronominal object favors VO by an odds ratio of $3/2 = 1.5$. If in a later period, VO outnumbers OV in clauses with a pronoun object by 4:1, then at the same point VO must outnumber OV with nonpronominal objects by 6:1, since $6/4 = 1.5$.

Kroch argues that the logistic function, as used by the logistic regressions described previously, is a good model for the relative frequency of two variants during change. The logistic function is a simple ratio $\frac{e^k}{1+e^k}$ which can be interpreted as a probability p . One reason to use this function is that its shape is the characteristic “S”-shaped curve associated with the uptake of a new form. Additionally, it provides for easy checking of the constant rate effect. This is because the constant rate hypothesis predicts that the difference in k for two environments should always be constant regardless of the actual value of p . For instance, in the example just given, in the first period nonpronominal objects favor VO with $p = 3/(1 + 3) = .75$, and pronominal objects with $p = 2/(1 + 2) = .67$. The corresponding difference in k is therefore $\log(3/2) = \log(3) - \log(2) = .41$. In the later period, pronominal objects favor VO

with $p = 4/(1 + 4) = .8$; therefore this value of k must be $\log(4) = 1.39$ and so the nonpronominal value of k must be $e^{1.39+.41} = 6$, and the probability $6/(1 + 6) = .86$. In logistic models, a difference in k values for predictions which differ only in one binary feature (such as pronominality) or one unit of a continuous feature (such as length) is simply the value of the coefficient associated with that feature. This justifies using a constant coefficient for some feature even as the base rate of the outcome of interest changes over time.

In this chapter, I implement a model of language change based on this intuition and on the corpus regression model described in chapter 2.

4.2 Evolutionary simulations

Beginning with work in the mid 90s by researchers such as Steels (1995), Niyogi and Berwick (1995/1998) and Kirby (1996/1999), computational simulations have become increasingly widely used to explore models of language change.¹ These studies span a huge variety of different methodologies and explananda, but all have in common that they provide formal implementations of theories of some type of language change, relating long-term change across generations to the language competence and use of speakers within each generation. For motivations and recent summaries of work in these fields, see Perfors (2002); Steels (2006); Niyogi (2006); Oudeyer and Kaplan (2007); Carstairs-McCarthy (2007).

Some published studies explore models that are simple enough to be described in a small number of algebraic rules, yielding specific law-like predictions for the trajectory of language change in the situation under study. However, most theories of interest describe *complex systems*, where language change is understood as an emergent macro-level characteristic of a system as a whole, which supervenes on the micro-level interactions between individuals. A complex system, in the terminology introduced by Poincaré, is one with long-term properties that cannot be factored into

¹I am referring here to simulations of change in language specifically rather than simulations of genetic evolution of a cognitive system for language processing, which also began to be investigated at around the same time or slightly earlier.

the properties of its components, because those components interact in non-trivial ways. For models like these, *multi-agent simulations* are the only way to evaluate the predictions of the theory.

In a multi-agent simulation, individual agents are explicitly represented *in silico*, as are their linguistic interactions with each other. In some studies (e.g. Steels, 2001), the population of agents is kept constant, and language slowly emerges through the agents’ attempts at cooperation. Steels calls this type of study a “language game”. In other studies (e.g. Kirby, 2001), “child” agents are added to the population periodically, and older agents removed. Kirby calls this type “iterated learning”, with each “generation” or “epoch” of agents corresponding to one iteration of a learn-produce cycle.² In either case, as long as the simulation implements a way for agents to set or update their own linguistic representations based on what they hear from others, the languages that are represented in the population can change over time.

By implementing a particular theory’s assumptions about linguistic representation and transmission, a particular prediction emerges from the simulation. To the extent that these results resemble real language change, we can take the simulation results as support for the theory. There is no consensus on exactly *how* such results should resemble real language: variously, simulations might be taken to be successful if they produce languages with any of the following kinds of properties:

- “universal” properties of language, such as:
 - power-law distributed word frequencies
 - regularisation of low frequency forms
 - compositional syntax
- over many tests, a distribution of languages that resembles the real distribution of human languages, in terms of
 - the relative proportion of different word orders
 - tendencies for certain parameters to pattern together

²See Steels (2002) for a discussion of the differences between language games and iterated learning.

- properties of the change itself (rather than the language) mirror actual change
 - in terms of its direction (the correct variant wins out)
 - in terms of its shape (typically, an “S-shaped curve”)
 - in terms of its timecourse (is the speed of change realistic?)

One common criticism of this newly formed field is that good standardised criteria have yet to be developed, which often makes it hard to evaluate the claims that are being made. However, any simulation at the very least provides (a) an implementation of a theory, giving a chance to check for “bugs”, and (b) an *existence proof*, showing that, were the world this way, the predicted results would be a possible outcome. While this latter claim may seem weak, it is at least enough to refute arguments made on the grounds of impossibility (e.g. such-and-such a construct must be innately specified because it could not arise merely to serve communication).

4.2.1 Processing in simulated variation and change

Some researchers have tried to incorporate processing preferences into simulation models. Kirby (1999) presents a series of “Filtered Learning Models” (FLM), exploring simple alternations between some small number of variant grammars in competition. In the simplest of these, child agents learn one of two grammars by selecting an utterance at random from the utterances produced by other agents. Utterances in this model are simple indicators of the grammar of the speaker, essentially just a binary variable that tells the learner which grammar to acquire: picking one utterance is enough to determine the learner’s final grammar. However, learners do not pick the utterance they learn from entirely at random: Kirby uses processing preferences derived from Hawkins (1994) to bias the choice, suggesting that learning may be harder from sentences that have been shown to be harder to comprehend. This “filtering” of the input is an efficient cause for the correlation observed in the Performance-Grammar Correspondence Hypothesis: easier grammars tend to spread throughout the population, leading to greater representation typologically.

A few simulation models have looked explicitly at word order change in historical English as an object of study. Yang (2000, 2002) explores simulations of the loss of the verb-second sentence type in Old English and French. Yang’s family of *variational learning models* (VLMs) is designed explicitly to account for the fact that change implies within-speaker variation: while Kirby’s agents have access to only one grammar, Yang’s simultaneously have access to all grammars in some hypothesis space G , but weight them differently, preferring to produce from some grammars more than others. They do this by storing a distribution $P(G)$ over each grammar $g \in G$; in learning, grammars that can parse an input sentence they hear have their probability increased relative to the grammars that cannot. Specifically, Yang’s learners use a classic learning theory algorithm due to Bush and Mosteller (1951) to update this distribution on the basis of their input. In the two grammar case where $G = \{g_1, g_2\}$ and an agent receives a sentence s , this works as follows:

1. select a grammar g to parse the sentence with probability $P(G = g)$; h is the grammar not selected
2. if g can parse s , set
 - $P(g) \leftarrow P(g) + \gamma(1 - P(g))$
 - $P(h) \leftarrow (1 - \gamma)P(h)$
3. if g cannot parse s , set
 - $P(g) \leftarrow (1 - \gamma)P(g)$
 - $P(h) \leftarrow \gamma + (1 - \gamma)P(h)$

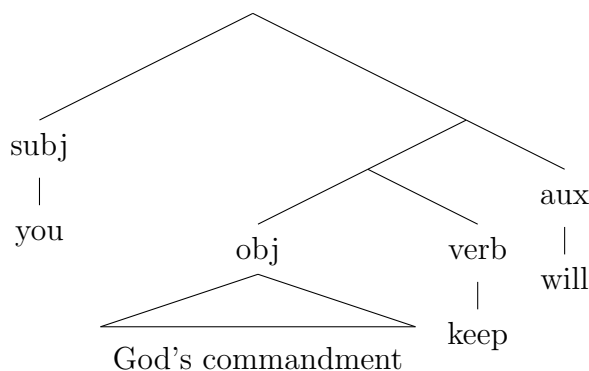
The parameter γ controls the learning rate: setting it to 1.0 results in a “memoryless” learner that always assigns a probability of 0.0 to whichever grammar most recently failed to parse a sentence. Any lower value results in smaller adjustments to the distribution, which over time tends to lead to a converge the probability with which each is able to parse a sentence drawn at random from the input.

In this way, Yang’s agents have access to multiple grammars, and in production generate from them with some probability distribution learned to approximate the

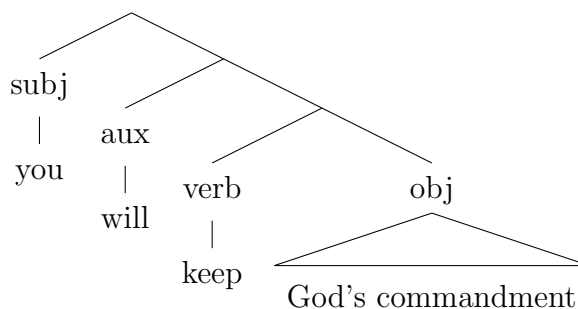
input. This is an implementation of the “grammar competition” proposed by Kroch (1989) and Pintzuk (1999, 2005).

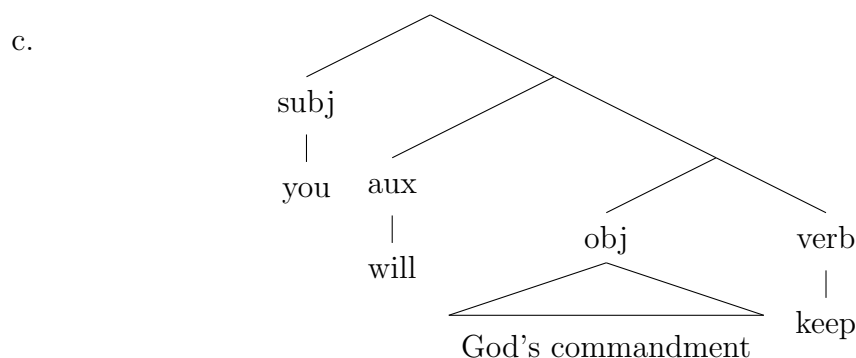
Clark et al. (2008) present a model based on Yang’s VLM, but also incorporating the processing biases of Kirby’s FLM. Clark and colleagues’ agents are based on Yang’s, and similarly have access to a distribution over all grammars. Additionally, they filter the input just as in Kirby’s models. Over several iterations, these agents come to favor a distribution over grammars which has been influenced by processing preferences just as Kirby’s agents do. However, any agent at any point in the simulation has the ability to comprehend utterances from either grammar, and has some probability of producing utterances from both. Clark and colleagues particular object of study is consistent branching in historical English: they show that a posited processing pressure to avoid inconsistent head direction could lead to a change in word order. The examples in (30) (adapted from Clark et al., 2008) show three orders of object, auxiliary verb and main verb that were all possible in Old English.

(30) a.



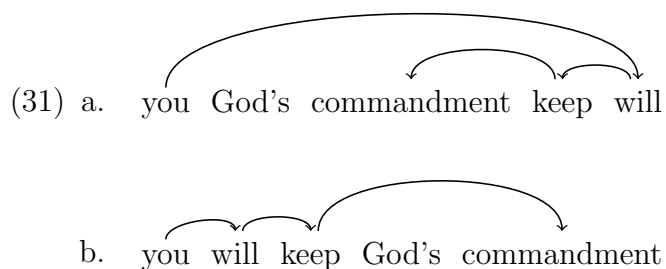
b.

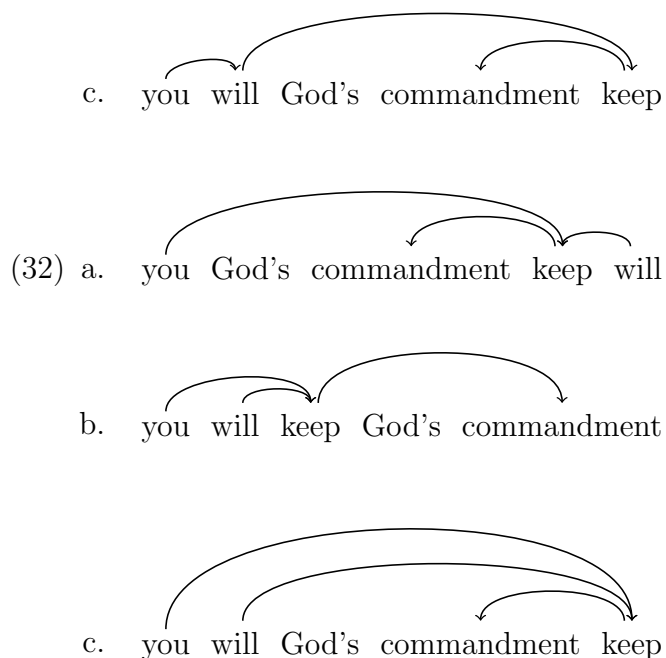




Ignoring the position of the subject, the assumed grammatical representations illustrated by trees show that (30a) is consistently “head-final” (since the verb comes after its object and the auxiliary after the main verb), (30b) is “head-initial”, and (30c), the so-called “brace”, has inconsistent head direction. While (30a) remained in competition with (30b) for some time, (30c) was lost much earlier. Clark and colleagues suggest that the typologically low frequency of inconsistently headed languages (Dryer, 1992) might reflect processing difficulty associated with such structures. By implementing an FLM-like filter, agents in their simulation gradually lose the brace order, even though at any point agents command all the word orders simultaneously.

As an aside, it is interesting to note that the dependency length theory makes different predictions for the relative processing complexity of these orderings depending on whether we assume a “syntactic” dependency chain *subject* → *auxiliary* → *verb* or “semantic” dependencies *subject* → *verb* and *auxiliary* → *verb*. Under both representations, the modern S Aux V O order is preferred. However, the syntactic representation ranks the brace order S Aux O V as less complex than the head-final structure S O V Aux, while the semantic representation yields the opposite order. This can be seen from the following illustrations, though exploring this any further is out of the scope of this chapter.





The variational learners of Yang and of Clark and colleagues represent the relative frequency of two grammars in variation, but they do not model the usage situations in which one variant may be favored over the other. As described above, corpus work on modern alternations has shown that it is the norm for variation to be conditioned on a multitude of factors, which means that the probability of a certain form being used in any given context may be substantially different from the language-wide average frequency of use. In the rest of this chapter, I present a model which corrects this shortcoming.

4.3 A model of OV loss by processing bias

4.3.1 A multi-cue variational learner

In this section, I will present a variational learner similar to Yang's which models variation not just at the level of a distribution over two grammatical options, but as a *conditional* distribution over those options given a variety of factors which could favor one outcome over the other.

This simulation is based on the historical data analyzed in chapters 2 and 3. In this model, agents' knowledge of language is represented by a logistic regression model similar to the one presented in chapter 2, and those used to model synchronic variation, as discussed above. When agents attempt to learn their language, they are fitting the parameters of that model from the data they receive as input sentences. When agents produce an utterance, they are sampling from the conditional distribution their regression defines over the possible word orders, conditioned on certain properties of the utterance they intend to express.

In this model, utterances aren't simply a binary outcome of the two grammars, but a tuple consisting of the features that were coded for in the corpus (argument pronominality and length, etc.) For the simulation reported here, I just consider those object properties that were found to be significant in the chapter 2 corpus model: object length, pronominality, negation, and accusative/direct vs dative/indirect object. When producing an utterance, agents sample from the empirical distribution over these utterance classes in the historical corpus: so, if clauses with an nonpronominal, nonnegative accusative object of length 4 make up 5% of sentences, agents will produce those 5% of the time. This makes the distribution of properties of utterances as close to the actual historical language as possible. I use only texts written before 950, so that the initial state of the simulation is similar to the oldest well-attested period of Old English.

Given an intended utterance, agents must decide whether to realise the sentence using the SOV or SVO outcome. They do this by feeding the properties of their utterance into a logistic regression model estimated from their input data. Therefore, if pronominal objects favour SOV and longer objects favour SVO in their representation, as they did in the historical data, agents producing utterances with those properties will tend to produce SOV or SVO respectively. Here again is the logistic regression formula:

$$p(VO) = \frac{e^k}{1 + e^k} = \frac{1}{1 + e^{-k}}$$

$$\text{where } k = \alpha + \beta_1 \cdot \text{obj pron} + \beta_2 \cdot \text{obj len} + \beta_3 \cdot \text{neg obj} + \beta_4 \cdot \text{case}$$

The parameter α , the intercept, reflects the overall frequency of VO in the ambient

language: it is the closest thing to an equivalent of Yang’s $p(g)$. The coefficients $\beta_{1..4}$ indicate how strongly each variable biases the agent towards or against VO in any particular case. For instance, a negative β_1 will lead to fewer VO outcomes when the object is a pronoun. The task of the learner is to select values for these parameters which make the order outcome in the input sentences they hear as likely as possible: that is, they make their best guess as to the parameter settings of the speakers of the sentences they hear. Given these parameters for an agent a , I will use $P(VO|s, a)$ to mean the probability a assigns to sentence s being realized as VO order given the linguistic properties that are available during planning.

The simulation proceeds in “epochs”, each of which involves a different population of agents. Learners estimate the parameters of their language model from a subset of sentences taken from an input pool, which is initially set to be a set of sentences from the historical data. The agents then produce new sentences with word orders determined by their language model; these become the input pool for the agents in the next epoch. This procedure is stated slightly more carefully below:

- select a random sample of S sentences from the historical data, including their word order outcome
- for each epoch (iteration),
 - create a population of agents P
 - set O to the empty set
 - for each agent $a \in P$:
 - * take a sample I from the set S
 - * fit a ’s parameters $\alpha, \beta_{1..4}$ to the input in I
 - * repeat the following $|S|/|P|$ times:
 - sample a sentence s from the historical data
 - assign s a new order outcome $\sim P(VO|s, a)$
 - add s to O
 - replace the old input S with the new utterances O

The only free parameters determine the number of agents and sentences available to be learned from. For the runs reported here, I set these to $|P| = 50$ agents, $|S| = 3000$ utterances saved per iteration, and $|I| = 1000$ utterances learned from by each agent. The agents’ regression models are fitted by maximum likelihood.³

In this model, it is not very natural to think of there being two separate grammars, with a distribution over them. Rather, there is a single grammar, or a single representation of language (use), but this representation includes probabilistic outcomes *within* it, such as the outcome governing the choice between two word orders. Probabilistic Context Free Grammars (PCFGs: see e.g. Manning and Schütze, 1999) are another example of a language model which *contains* probabilistic outcomes rather than a probability distribution over non-probabilistic language models. This way of understanding the model makes it much easier to “scale up” to language variation more generally. After all, real languages exhibit multiple variations simultaneously: a modern English variation such as the ditransitive (“*give NP1 to NP2*” vs “*give NP2 NP1*”) contains two NPs which can each undergo the possessive alternation (“*NP1 of NP2*” vs “*NP2’s NP1*”), meaning either a distribution over $2 \times 2 \times 2 = 8$ discrete grammars, or a single representation containing two probabilistic outcomes.

In figures 4.1 and 4.2 I plot the results from two independent runs of the simulation. The top left subplot shows the proportion of sentences that are produced with VO order in each epoch. The remaining subplots show the change in each coefficient in the agents’ language models: as in all regression models these can take any numeric value, where 0 indicates that the associated variable has no effect on the outcome, and larger magnitudes indicate progressively larger influences on the outcome. Here, positive values mean that the factor biases towards VO, and negative towards OV. The results from these simulations confirm that there is no particular tendency for the

³Strictly, the models are fitted by MAP probability, since the fitting procedure used here also includes a gaussian prior over coefficients with a standard deviation of 1 (sometimes also called an L2-norm regularizer). This has the effect of discouraging coefficient values that are excessively far from 0, but in fact has very little effect on the simulation. The only noticeable difference is that in the later iterations of the biased simulation, once the population has very nearly converged on all-VO thanks to the high intercept value, and the coefficients no longer influence the outcome, the coefficients swing around wildly rather than reverting to 0. The continuous length variable is standardized by dividing by two standard deviations to be on the same scale as the dichotomous variables and therefore equally regularized in fitting.

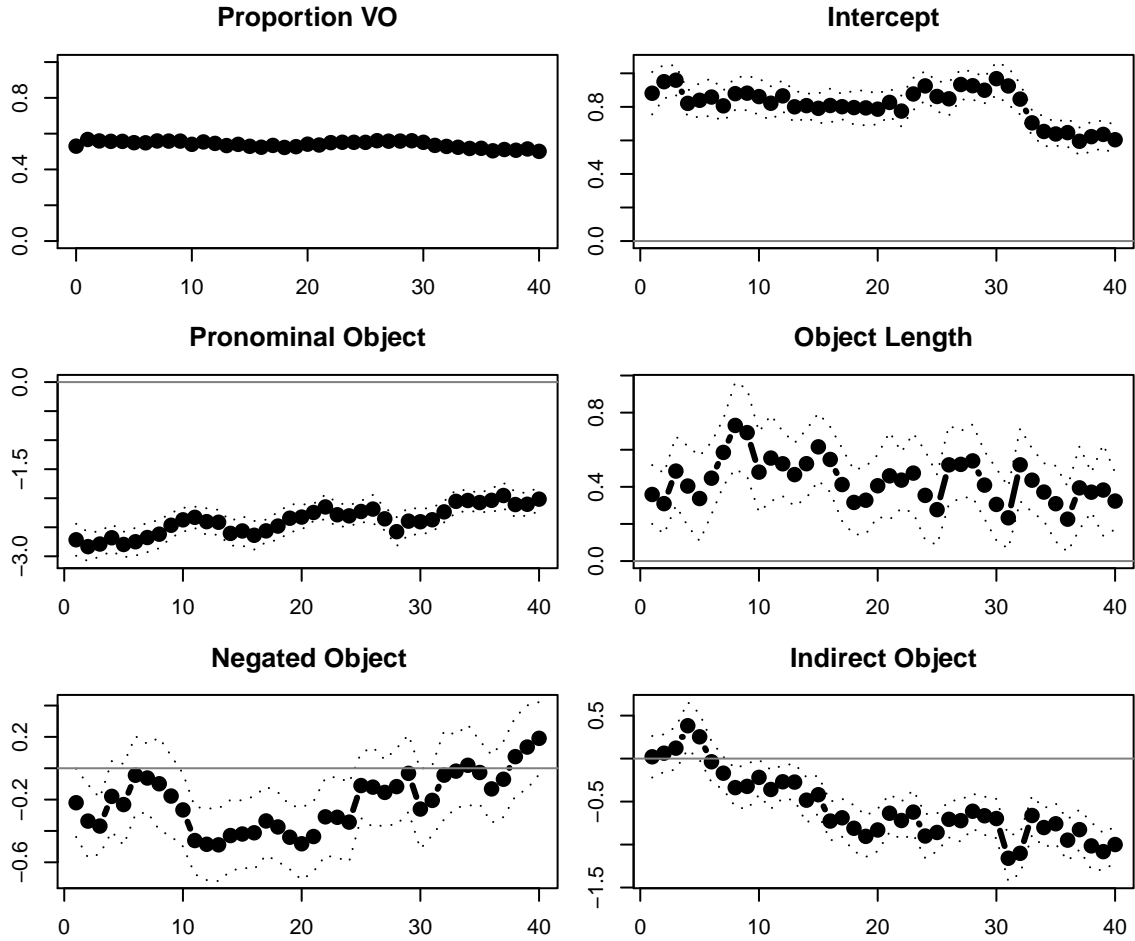


Figure 4.1: Unbiased simulation, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations.

proportion of VO to change over time: there is a certain amount of drift, but with the parameter settings tested here, agents can learn a grammar which reflects the proportion of the two orders in the input quite accurately. The initial historical data is 56% VO. Therefore, agents in these simulations are commanding both word orders throughout, just as Yang’s variational learners do. However, the extent to which the individual predictors of order keep a stable weight throughout the simulation varies depending on how strongly each predicts the order outcome. Object pronominality, for instance, is very strongly associated with OV order in the input data, and this

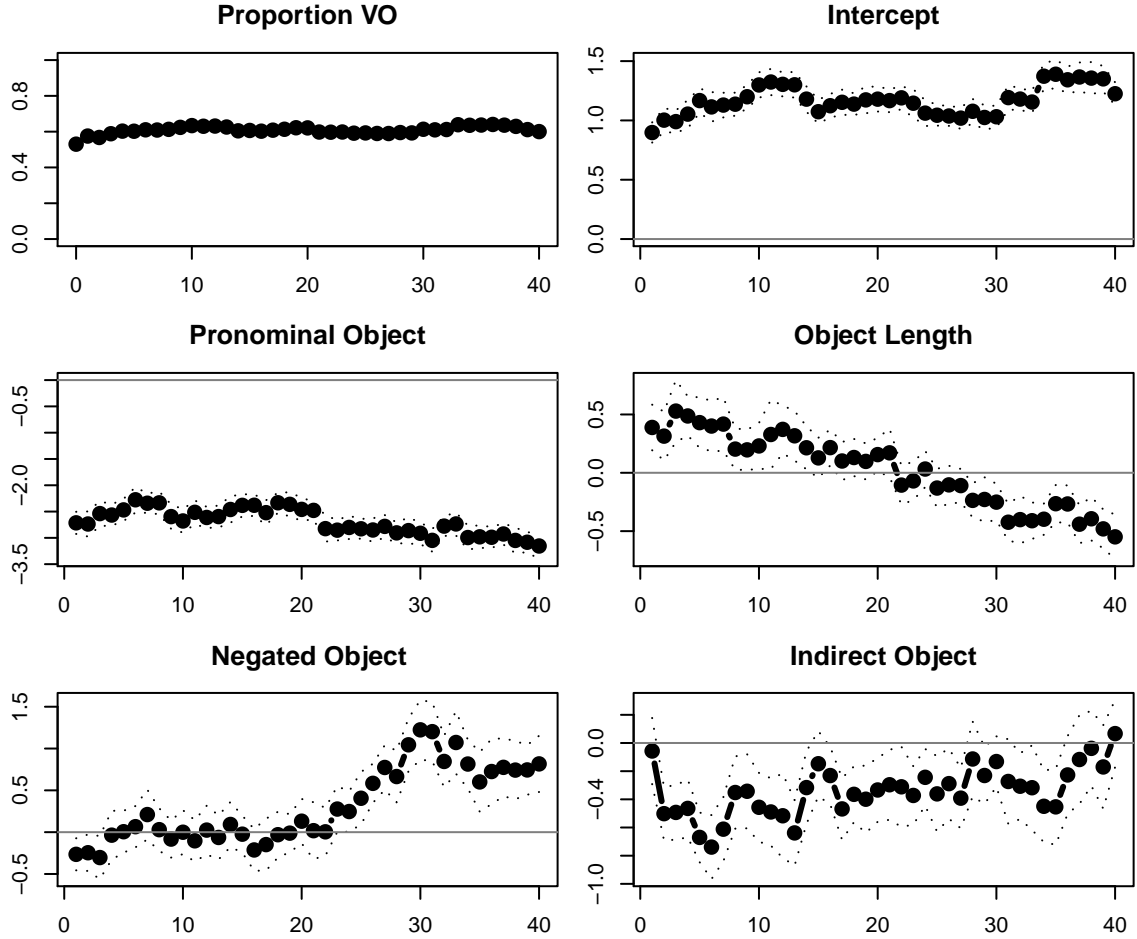


Figure 4.2: Unbiased simulation, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations.

fact is preserved throughout both simulation runs reported here: after 40 epochs, its coefficient is still quite close to the original estimate in both runs. Object negation and type, however, are only weakly predictive of the outcome, and in the simulation they are subject to much more drift. In the first run, negated objects are weak predictors of OV order for at least 25 epochs, while indirect objects begin with no association and gradually come to favor OV. In the second run, negation has a weak bias towards OV order initially, but over time gradually flips its effect to become associated with VO. Indirect objects, however, remain weak predictors of OV for most of the timecourse.

Object length is less strongly predictive of the order outcome than pronominality, but more strongly than either of the other two variables. In the first simulation, it remains fairly constant as a predictor of VO, while in the second it gradually loses this effect and comes to favor OV. If this model were accurate (i.e., if length could be arbitrarily associated with word order rather than having some kind of universal bias towards a certain grammatical configuration) we would expect human languages to favor orders which maximize dependency lengths as often as the favor orders which minimize them. As discussed in chapter 3, this is not the case.

4.3.2 A biased multi-cue variational learner

Next, I adapt this simulation by adding a processing bias, similar to the simulations reported by Kirby or Clark and colleagues.

As discussed in Chapters 2 and 3, there is strong evidence that dependency minimization effects influence both comprehension and production. Therefore, I could conceivably implement a bias in both production (agents produce more VO clauses with longer objects) and comprehension (agents do not learn as much/as often from OV clauses with longer objects, as in Kirby’s FLM). Both of these would have led to the effective proportion of VO/long object sentences from which the learner acquires a grammar being exaggerated relative to OV/long object sentences. Since phrase length is well known to influence frequency in production (chapter 2) while the equivalent bias in comprehension involves an extra step of reasoning (assuming that learners do not learn well from structures that incur more processing difficulty) in this model I implement a processing bias in production. However, either locus of the effect would be expected to give the same results (see Briscoe, 1998, discussed in Clark et al., 2008, for a study confirming this).

The weight effect observed in chapter 2’s corpus study was argued to reflect a universal processing preference, and therefore the weight effect should always act in the same way, rather than being freely learned just like any other cue in the model. To implement this, I simply fix the coefficient to some suitable value, yielding a “biased” simulation. The only differences from the above simulation are that (a) in

learning, agents do not estimate an object length coefficient, and (b) the object length coefficient in production is set to .92, the value found in the empirical corpus study. These results are plotted in figure 4.3.

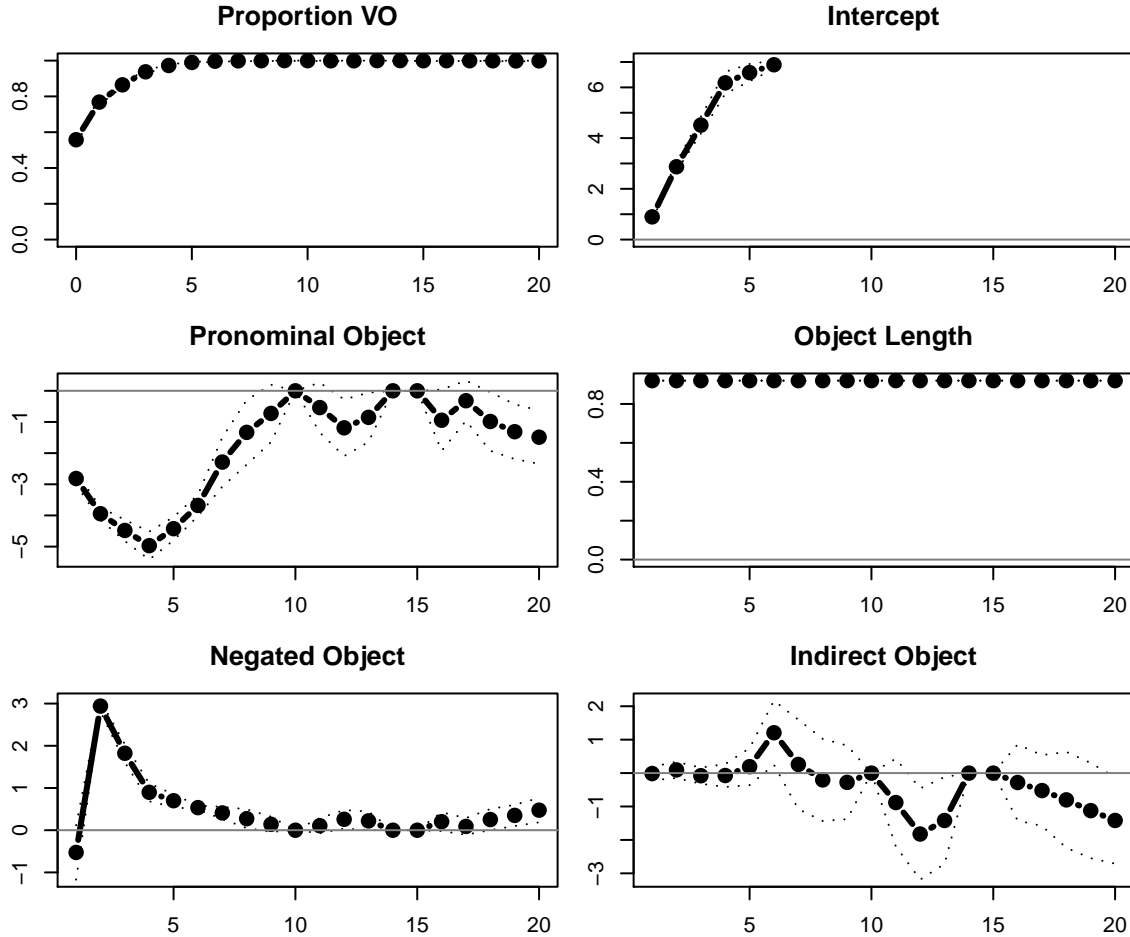


Figure 4.3: Biased simulation: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations.

Because including an object length coefficient in production causes agents to produce, on average, more VO clauses than they were exposed to in learning, the intercept term (indicating the aggregate level of VO in the language) increases rapidly. Accordingly, the proportion of VO order produced epoch on epoch also increases rapidly, and by the 8th epoch there are agents who have only been exposed to VO sentences,

and so acquire a deterministically VO grammar. This is represented in this model by an intercept of infinity, which means that it is no longer possible to plot the population mean intercept once there are such speakers in the population. Early in the simulation, the coefficient for object pronominality becomes more strongly negative. This is because in the input data there is a strong tendency for pronouns to be preverbal, and since pronouns are short, agents “correct” for the increasing VO tendency by correspondingly increasing the OV tendency in pronouns. This effectively offsets the length effect, meaning that pronominal objects remain preverbal even after most other objects are postverbal, just as was observed in the empirical data. However, once the VO order becomes frequent enough, there are too few OV exemplars left for the learner to accurately acquire this tendency, and the population mean of the pronominality coefficient starts to revert to 0.

In this model, the change to VO is not dependent on the proportion of VO in the input to the first generation of agents. To verify this, I ran the simulation once more after manipulating the data set to reduce the number of VO outcomes in the initial input to 10% of the total data.⁴

The resulting simulation run is plotted in figure 4.4. The qualitative patterns of change are identical to those given the historical data, even though the start point is very different.

This simulation captures the basic tendencies observed in historical English: what began as a tendency for OV orders gave way to increasing — and eventually deterministic — VO. At the same time, factors that are strongly associated with an outgoing order (e.g. object pronominality) can remain associated with it even as the overall frequency decreases.

⁴The precise details for the way this was done are as follows. A logistic regression model was fitted to the entire dataset, predicting the order outcome. I then found the 90th percentile of the empirical distribution over the model’s predictions: the value $p(VO)$ such that only 10% of the data had a probability of being VO at least that high. That 10% were set to be VO, and the remainder OV. To ensure that the agents’ models could not fit this new order outcome perfectly, before calculating the 90th percentile I added noise to the probability estimates. This noise was normally distributed in log odds space, with mean 0 and standard deviation equal to the pooled standard deviation of estimates for true VO outcomes and estimates for true OV outcomes.

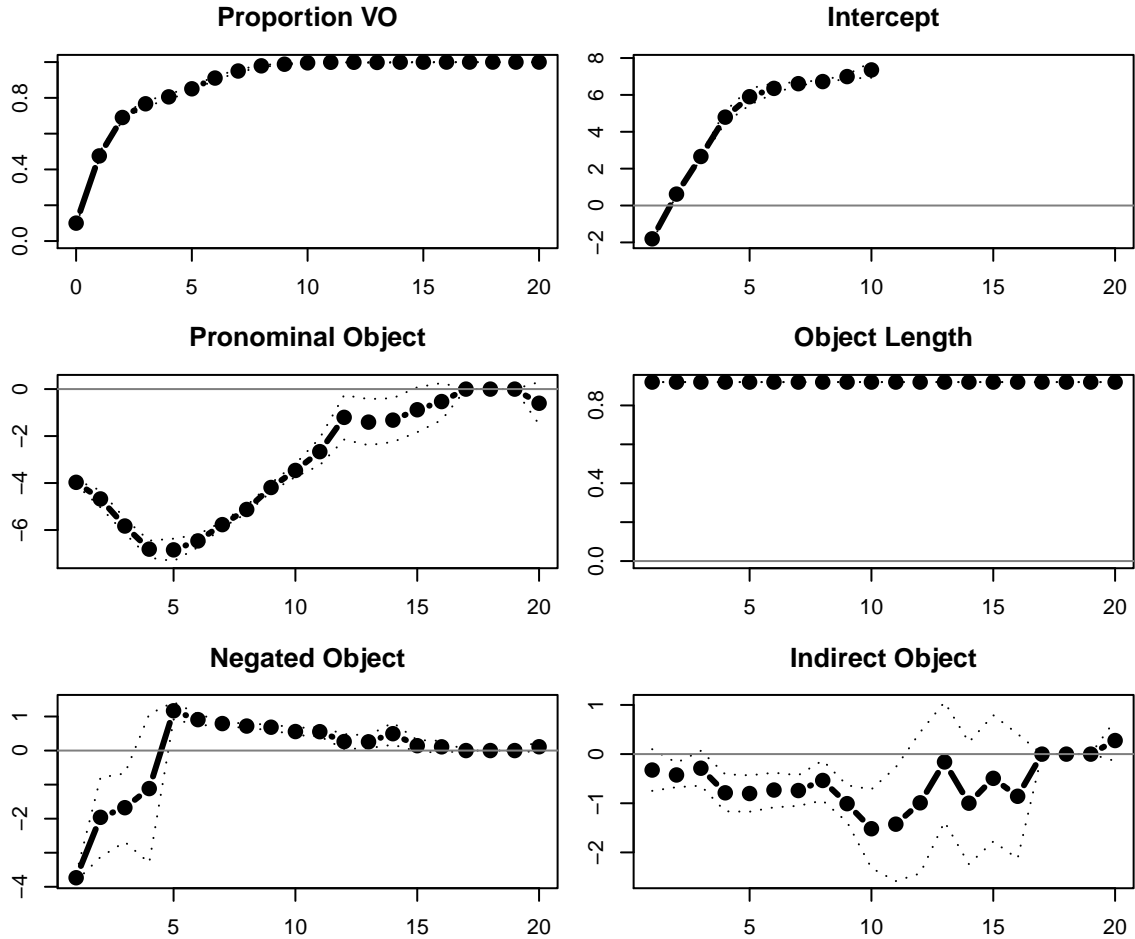


Figure 4.4: Biased simulation (OV start state): Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show two standard deviations.

4.4 Problems with the theory

4.4.1 Is there really a universal preference for SVO?

Perhaps the most glaring inadequacy of this model is that it shows a universal preference for SVO over SOV order. This is at odds with the actual distribution of word orders in the world's languages: as mentioned in chapter 2, languages with basic SVO and SOV order are roughly equally frequent, and together seem to make up something like 90% of all languages (Tomlin, 1986). As the final biased simulation run

shows, the model predicts that even majority SOV languages should quickly switch to a SVO order, implying that SOV languages really should not exist.

Of course, this model only incorporates a single bias — one which favors SVO over SOV — and in reality a large number of factors may interplay to determine the orders eventually favored by human language users. In fact, several potential cognitive biases or strategies have been proposed that favor SOV or eliminate processing cost associated with SVO. I will survey these briefly here.

Ueno and Polinsky (2009) suggest two possible “strategies” that speakers of SOV languages might employ to avoid difficult sentences with multiple preverbal arguments. First, they could omit one of the arguments by “pro-drop”: that is, in situations where a speaker of a language like English would produce a pronoun as the subject due to the referent’s high accessibility, SOV languages might more commonly allow their speakers to omit the subject altogether. However, in a series of corpus studies on children’s spoken English (where subject omission is common) and on scene descriptions in Japanese, Spanish and Turkish (all of which allow subject omission), they found no support for this idea. In fact, speakers of all languages were more likely to omit arguments when the clause was transitive than intransitive, supporting their hypothesis that pro-drop might be employed more in more complex sentences. However, the SOV languages (Japanese and Turkish) showed no more incidence of pro-drop than the SVO languages (Spanish and child English). If pronoun omission did reflect a processing strategy that is particularly useful in SOV languages, we would also expect that SOV languages are more likely to allow pro-drop than SVO languages. This does not seem to be the case (or at least, any effect is very small) in accord with Ueno and Polinsky’s corpus findings. The following data is a cross-tabulation of counts of languages which allow subject omission against basic word order, for those languages in the World Atlas of Linguistic Structures (WALS: Haspelmath et al., 2005) for which both are annotated:

| | | expressed subjects | | | |
|-------|---------------|--------------------|----------|---------------|-------|
| | | always | optional | other marking | mixed |
| Order | SOV | 26 | 23 | 143 | 7 |
| | SVO | 37 | 19 | 182 | 10 |
| | other | 6 | 5 | 49 | 7 |
| | none dominant | 5 | 7 | 83 | 2 |

Looking at the top left four cells, the ratio of SVO languages that have optional subjects to those that have obligatory subjects (1:1.1) is lower than the same ratio for SOV languages (1:2.1), but not by very much and in fact this difference is not significant by a chi-square test ($\chi^2(1) = 1.34, p = .247$). The “other marking” category lumps together several other means of expressing the subject (including verbal affixes, clitic pronouns or agreement markers and pronominal subjects in noncanonical subject position) for which the prediction is not clear.

Ueno and Polinsky’s second proposed strategy similarly avoids clauses with multiple arguments, but this time by favoring intransitive clauses over transitives. By using descriptions made by native speakers in response to the same stimulus scene, they could control for differences in content and test whether SOV speakers tended to favor descriptions containing more intransitive clauses where SVO speakers would favor transitives. This time, there was a sharp contrast between SVO and SOV in the expected direction. Taken together, Ueno and Polinsky interpreted their results as evidence that argument omission is a general strategy for simplifying otherwise complex sentences across languages, while favoring intransitive clauses is a strategy used by SOV languages in particular. If this is the case, then it could be part of the reason for SOV languages being processing efficient despite the apparent long dependencies they incur: speakers may simply choose to express themselves in a way that avoids such dependencies wherever possible.

A further way in which languages could relieve processing complexity due to long dependencies is via a case marking system. Memory-based theories of processing complexity (e.g. Gibson, 2000; Gordon et al., 2001; Lewis and Vasishth, 2005) suggest that part of the difficulty associated with long dependencies is interference from

intervening arguments. For instance, the phrase “*the reporter that the senator attacked*” is predicted to be more complex to process than “*the senator that attacked the reporter*” because in the former, the noun phrase “*the senator*” intervenes between *reporter* and *attacked*, and therefore may cause interference when trying to retrieve or integrate that dependency. Case marking may serve to differentiate the NPs, avoiding interference: at the verb, the processor can search for a unique cue (nominative case) to retrieve the appropriate noun from memory. Additionally, case may allow early prediction of upcoming structure: as soon as an accusative-marked noun is encountered, the processor knows that this noun is an object of some kind and may predict likely upcoming structures on that basis. For instance, in the Japanese sentence (33), the object *neko* bears an object case marker which allows the processor to predict its role in the sentence with high certainty even before the verb is reached.

- (33) inu ga neko wo mikaketa
 dog-SBJ cat-OBJ saw
 “*The dog saw the cat*”

Hawkins (2004) suggests this kind of early assignment of properties makes for more efficient processing. There are at least two possible ways, then, in which we might expect languages with rich case marking to avoid or make up for any processing cost associated with SOV.

Naturally, this makes the prediction that SOV languages will tend to have rich case systems more often than SVO languages do. This, in fact, is Greenberg’s (1963) Universal 41, and is well supported. Below I have cross-tabulated basic word order against the presence of a morphological case marking system, for languages where both are annotated in WALS.

| | | morphological case | | |
|-------|---------------|--------------------|------|------------|
| | | none | case | borderline |
| Order | SOV | 18 | 69 | 13 |
| | SVO | 50 | 17 | 2 |
| | other | 16 | 11 | 1 |
| | none dominant | 10 | 23 | 5 |

Just looking at the top left four cells of the table, the expected pattern is evident: SOV languages with case outnumber those without by almost 4:1. Conversely, SVO languages without case are more frequent by almost 3:1 ($\chi^2(1) = 42.5, p < .001$). SOV languages are also more likely to fall into the “borderline” case category, which refers to languages that arguably have case marking but for only for certain oblique (adverbial) functions and not for functions that would typically be included in subject and object categories. This pattern supports the idea that case may increase a language user’s ability to do early processing online, and reduce interference between arguments or otherwise aid memory at the verb.⁵

If a rich case system for disambiguating or differentiating arguments is part of the reason that SOV does not incur higher difficulty than SVO, then it seems likely that during the period when the English case system was collapsing, clauses where case was already ambiguous would be realized in SVO order more often than those where the subject and object were transparently case marked. However, Pintzuk (2002) finds no effect of this kind. In OE clauses with auxiliary verbs, she finds objects with unambiguous case to appear after the main verb in 35.8% of her sample, while objects whose case is ambiguous appear after the verb 37.1% of the time. In a very preliminary study using the OE corpus data reported in chapter 2, I did not find any sizeable difference in VO rate depending on case ambiguity in clauses without auxiliaries either.

Of course, “strategies” for ameliorating difficulty in SOV can only be part of the picture, since without any additional reason to think that SOV languages might be adaptive in some other way, we would still expect an overwhelming preference for SVO. SOV languages could survive as long as they made good use of one or more of these strategies, but as soon as those strategies fail to be used 100% of the time, we would again expect a one-way bias in favor of SVO. And these strategies are not always used even in stable SOV languages: Ueno and Polinsky record 64 and 56% intransitive clause use in their SOV languages (as opposed to 50 and 45% in SVO);

⁵The discussion here assumes a very simple mapping between case markers and grammatical function. In reality, languages differ in how transparent and universal this mapping is. However, the general principle follows in any situation in which case narrows down the range of possible interpretations in *any* way, and even if only partially or probabilistically.

likewise, syncretism and ambiguity in case marking systems is the norm rather than the exception.

In actuality, there is also a proposal for a general cognitive bias towards SOV. In a series of papers, Susan Goldin-Meadow and colleagues have argued that the SOV structure reflects an ordering which is somehow conceptually more basic or default. They show that homesign — manual communication systems that are invented by deaf children without access to a signing community — typically uses the order agent-patient-action regardless of the basic word order of the background spoken language (Goldin-Meadow and Mylander, 1998; Goldin-Meadow, 2005). Goldin-Meadow et al. (2008) obtain similar results when speaking participants are made to “act out” scenes nonverbally in a lab experiment — essentially, to invent a simple signing system. Additionally, in a lab task in which (speaking) participants were required to assemble pictures of scenes by overlaying pictures of agents, patients, and actions, participants tend to select the component pictures in that same order. Goldin-Meadow and colleagues suggest that agent-patient-action may represent a “cognitively basic” order for representing events, linguistically or non-linguistically. They argue that SOV is therefore the likely candidate for the basic order of some earlier form of all languages, but that as the needs of language users become increasingly complex, other pressures (such as the dependency minimization pressures described in chapter 3) begin to favor a shift to SVO.

Looking for evidence that there is indeed some interplay between pressures for SVO and for SOV, Langus and Nespors (2010) and Gibson, Brink and Saxe (in prep) have independently replicated Goldin-Meadow’s experimental findings with gestural “acting out” of simple scenes. However, both groups find higher incidence of SVO with more complicated sentences. Langus and Nespors report that both Italian (relatively free SVO) and Turkish (relatively free SOV) speakers produced the main clause followed by the subordinate clause in biclausal sentences like “[*the man tells the child*] [*that the girl catches a fish*]”. Thus a clausal object is placed postverbally by both SVO and SOV speakers. Gibson and colleagues find agent-patient-action order most common with English speakers even with complex objects like “*the hat with the star*”

but increased levels of agent-action-patient when both the agent and patient are animate. Both groups attribute these results to an interplay between two pressures. Langus and Nespors suggest that the SOV bias is a cognitive primitive, but not one relevant to language proper, and only arises due to participants not processing overly simple sentences as language. They suggest that the SVO bias, in contrast, reflects an innate linguistic pressure for postverbal objects. Gibson and colleagues suggest that while the SOV bias is some kind of cognitive universal, a pressure to differentiate confusable referents (as predicted by memory-based theories of processing complexity) leads to SVO in cases where both are similar.

In summary, then, it is likely that multiple factors convene to determine the processing complexity of a given linguistic form such as a word order template. Work identifying these and describing their interactions is only just beginning. However, even if there were good reasons to think that the overwhelming SVO bias of this model would be tempered in real language use, there are other significant problems with this as a model of change which I will describe below.

4.4.2 Are language learners really this bad?

Although this model captures the fact that preverbal pronouns remain common late in the change towards VO, it does so in a way which is incompatible with the empirical findings presented in chapter 2. In that model, there was no apparent interaction between date and pronominality, meaning the coefficient remained constant for the duration of the change. In the simulation, however, the pronominality coefficient increased sharply in magnitude, as a way of counteracting the increasing intercept term. The empirical model leads us to expect that object pronouns should always favor OV order, but only by some constant odds ratio, as discussed by Kroch (1989).

In fact, this problem highlights a more fundamental theoretical issue with this model. Human language learners in general do a good job of inferring the patterns in the language data they are exposed to, and are able to learn even relatively weak probabilistic associations between cues. If this is the case, then this model's stipulation that learners cannot infer any relationship between length and order seems

arbitrary. If language users can in fact infer the way in which constituent length affects placement in a grammar, then the model is inadequate as it stands. In the next chapter, I use empirical experiments with human participants to test whether language users do have this ability.

Chapter 5

Inference driven comprehension in a free word order language

In this chapter I report several experiments using speakers of a modern verb-final language with significant word-order variation: Japanese. The variant word orders of Japanese main clauses do not obviously differ in their average dependency lengths. However, I will focus on word order not as an outcome, but as a source of information about interpretation. I show in a number of experimental studies that comprehenders do use word order as a source of information, even when it only gives partial, probabilistic clues to interpretation, as was the case in historical English and likely in all languages with word order variation. In one recall/constrained production experiment, I back up previously reported findings that Japanese speakers prefer to order long material *before* short, supporting the dependency minimization theory outlined in Chapter 3. Finally, I show that at some level, comprehenders themselves know about this long-before-short bias in production, and in fact even use that knowledge to aid their own processing: when they encounter a initial phrase, they are more likely to entertain the possibility that it is an object if it is long than if it is short. This novel finding is predicted by a rational model of inference, but not by classical cue combination models. These results call into question the simplistic model of change discussed in Chapter 4, in which weight affects change via comprehenders' failure to compensate for production bias.

5.1 Word order and case cues in processing

One of the central tasks performed in language processing is establishing the relationships between referents mentioned in a sentence. For instance, in the English sentence *dog bites man*, the *dog* is understood as the agent of a biting action, and the *man* as its patient. English conveys these relationships through word order — the sentence *man bites dog* has a different meaning — so part of the comprehender’s task is to reconstruct these parts of the intended meaning given the order of the words. However, languages differ in the devices they offer for expressing similar relationships: English is in fact at one end of a continuum, being extremely dependent on word order, while other languages use morphological agreement or additional words or clitics to mark the same relationships (e.g. MacWhinney et al., 1984; Croft, 2002; Siewierska and Bakker, 2008).

The fact that different languages employ different solutions to the “mapping problem” between form and meaning means that human learners must be flexible enough to acquire whichever strategies are appropriate for dealing with the particular ambient language.

5.1.1 Processing differences by language type

There is a general consensus within modern psycholinguistics that the probability of particular linguistic forms — as measured by their frequency, or probability in context — influences comprehension difficulty. Less frequent words or structures are processed less quickly and comprehended less accurately. For instance, frequency or probability predict ease of comprehension in domains such as verb subcategorization preferences (e.g. Trueswell et al., 1993; MacDonald et al., 1994) and lexical ambiguity (e.g. Jurafsky, 1992; Gibson, 2006). The relevant probabilities may be purely structural, as in the probability of a certain upcoming word given the dependent phrases or parts of speech seen so far (e.g. Hale, 2001; Konieczny, 2000); or, they may also condition on lexical and “real world” knowledge such as whether a referent is human or not (e.g. Trueswell et al., 1994; Gennari and MacDonald, 2008). Levy (2008) provides a formal model of *expectation-based processing* within which many or all of

these findings could be explained.

Expectation-based processing and related theories can help explain the human ability to adapt to the very different kinds of strategies for encoding meaning seen across the world's languages. Under this account, language users attend to different cues to structure or meaning in proportion to the extent they are used. Therefore, they quickly become attuned to the particular patterns of usage of their language, whatever those patterns happen to be. In this view, there is no need to encode innate knowledge that particular cues such as word order or case should be attended to: over time, language learners simply notice which features of their language reliably correlate with which components of meaning, and learn to attend only to those features. This hypothesis allows for fairly subtle cross-linguistic differences in comprehension strategies. To give just one example, speakers of different languages appear to prefer different readings for the structurally ambiguous sentences like the following, given in both English and Spanish (Cuetos and Mitchell, 1988; Brysbaert and Mitchell, 1996; Zagar et al., 1997; Hemforth et al., 1998):

- (34) a. Somebody shot the maid of the actress who was on the balcony
- b. Alguien disparó contra la criada de la actriz que estaba en el balcón

Some languages, like English prefer a “low” reading where the actress is on the balcony; others, like Spanish prefer a “high” reading with the *criada* on the *balcón*. This finding cannot be accounted for by universal rule-based theories such as Frazier and Fodor's (1978) Late Closure principle, which predicts that *all* languages should prefer to attach “high”. These results motivated Mitchell, Cuetos, Corley and Brysbaert's (1995) *Tuning Hypothesis*, another probability-driven theory. The Tuning Theory suggests that processing preferences like these are a function of language-specific exposure: languages may differ in the frequency with which they use a certain structure, and comprehenders prefer that structure in direct proportion.

Given the multitude of different features of language or meaning a comprehender could potentially condition on when estimating the probability of some word, structure, or meaning, we need theories of how comprehenders select which cues to pay attention to, and how they combine information of different types if they rely on

more than one simultaneously. A large amount of work on this topic has been carried out under the auspices of the *competition model* (e.g. Bates and MacWhinney, 1982, 1989). The competition model holds that speakers of different languages rely on particular cues in comprehension, in proportion to their degree of availability (how often the cue is present to be used in a sentence) and reliability (how often using the cue yields the correct result). For instance, MacWhinney et al. (1984) show that when given two nouns and a verb and asked to guess which noun which was the “do-er” of the action described by the three word “sentence”, English speakers overwhelmingly relied on word order. They interpreted VNN as VOS, NVN as SVO, and NNV as OSV, and paid much less attention to other cues such as number agreement. Italian speakers, on the other hand, paid rather less attention to word order, typically interpreting the noun that agreed with the verb as the subject. German speakers, finally, almost ignored word order, paid some attention to agreement, but largely relied on the animacy of the nouns in choosing an agent. These differences can be understood as differences in the utility of those cues in each language: English has rigid word order, while Italian has frequent SVO order but allows all other possible orders in some contexts. German has relatively free word order, but the canonical nominative subject-accusative object sentence template does not occur with animate objects and inanimate subjects. For example, English oddities like “this tent sleeps four” are prohibited. This makes animacy asymmetries a reliable cue for subjecthood.

Similar results obtain in other experimental paradigms using normal grammatical sentences. For instance, Love and Swinney (1998) use a cross-modal lexical priming task in which subjects listen to a sentence but at some point during it are required to answer a question related to a separate visual stimulus displayed onscreen. English subjects show evidence that the object of a verb is reactivated in their memory immediately after they hear the verb, even in sentences when the object of the verb is absent from that position though understood there (35).

(35) He went to the *bank* which his family always used ____

This reactivation is apparent due to faster reactions when the visual stimulus presented at that point is related to the object (e.g. “money”). Conversely, Bulgarian

subjects in related studies show no such reactivation. Love and Swinney attribute this to the fact that English word order is fixed, and hence its comprehenders rely on it, while Bulgarian has freer order, and hence its comprehenders do not.

The competition model proposes that language users rely on a variety of different *cues* in language interpretation, and that these cues are weighted in importance in proportion to their usefulness in a particular language. Thus English speakers would weight word order cues much higher than other cues, while German speakers would weight animacy and case highly. These weightings are formalized through the notion of *cue validity*, a simple measure of usefulness which has two primary components, *availability* and *reliability*. Availability is calculated as the proportion of time that a certain cue is available to a comprehender. For instance, in English a word order cue like “the NP before the verb” is highly available, since there is typically one NP before the verb, while in German or Japanese there are many sentences with no NPs before the verb, or several. Conversely, “the NP with nominative case marking” is not very available in English since only a handful of pronouns actually mark nominative case (*I, he, she, we, they*), while in German or Japanese many or most sentences will have an unambiguously nominative NP, making that cue more available. Cue reliability measures the proportion of time that relying on a given cue will yield some particular interpretation. For example, in English both “the NP before the verb” and “the NP with nominative case marking” are highly reliable cues for assigning some noun to the subject category, since both will pick out the correct subject a high proportion of the time that they are available. The availability of a cue c can be thought of as its prior probability $P(c)$ of appearing in any given sentence, and its reliability for picking out a category of meaning m is the conditional probability of picking out that category given that it appears, $P(c_m|c)$. The aggregate measure cue validity is given by multiplying availability and reliability. Given the probability theory definitions above, then, by the chain rule, cue validity is simply the joint probability of a cue occurring and picking out a particular category: $P(c, c_m) = P(c) \cdot P(c_m|c)$. The competition model suggests that language users interpret sentences based on cues weighted in proportion to this measure.¹

¹In fact, the competition model also considers a further definition of cue validity, *conflict validity*,

5.1.2 The utility of word order variation

Because English relies so strongly on word order to express meaning, it has relatively little flexibility to rearrange the order of words without changing meaning. Given the extensive evidence that word order influences processing complexity in both production and comprehension (see Chapter 3), this inflexibility might be expected to be a burden on speakers of languages like English: for instance, sometimes the grammar might constrain speakers to producing a sentence-initial phrase that is conceptually inaccessible or otherwise incurs difficulty when placed early in the sentence. Fortunately, English and other languages often provide multiple ways of expressing the same meaning that differ in their word order. From a processing point of view, having these options carries a distinct functional advantage. V. Ferreira (1996) shows this particularly clearly for the English dative alternation: participants were asked to construct a sentence using supplied fragments such as *I/gave/toys/children*. The critical items contained either an “alternating” verb, one that could appear in both the NP NP and NP PP constructions (36a,b) or a “nonalternating” one that could only appear in the NP PP construction (36c,d).

- (36) a. I gave some toys to the children
b. I gave the children some toys
c. I donated some toys to the children
d.? I donated the children some toys

Participants made fewer grammatical errors and were faster to construct sentences when given an alternating verb and nouns which were compatible with either construction. Comparable results obtained in another experiment using the active/passive alternation. Ferreira takes these findings as support for a flexible, incremental model of grammatical production: if the processor is able to plan one of the noun phrases more quickly than the other, then the speaker can choose the construction which places

which determines what language users do when multiple cues are in conflict with each other. This is out of the scope of discussion for the moment.

that NP earlier. When the only choice is theme-goal (NP PP: 36c), then sometimes the speaker may complete planning of the goal first but not be able to use it immediately, making production less efficient. Thus the availability of word order variations makes communication an easier task. Conceptual accessibility factors are obviously predicted to influence which arguments are faster to plan (Chapter 3), but similar arguments could be made for processing benefits gained by choosing the construction that yields shorter dependency lengths, for instance. Only speakers of a language with appropriate word order freedom or word order alternating constructions would be able to benefit from these potential processing “rebates”. The syntactic flexibility offered by meaning-equivalent constructions or optional words, therefore, may be a functional solution to allowing the speaker to produce language efficiently online (Ferreira and Dell, 2000).

In fact, it has been suggested that the emergence and spread of certain constructions in the recent history of English are directly due to the loss of the significant word order freedom of Old English (e.g. Los, 2002). There is evidence that the passive construction increased dramatically in frequency in Middle and Early Modern English to replace other methods of moving given information to the beginning of the clause. Corpora from that period examined by Seoane (2006) suggest that while around 65% of passives result in a given-new order, less than 5% result in the opposite new-given order (comparable statistics for active sentences are unfortunately not given, so it isn’t possible to conclude that information structure actually influences the outcome, although these proportions alone are striking). Seoane also finds a trend for the (*by*-phrase) agents of passives to be longer than their subjects in around 75% of passives in that time period, which she takes in support of the End Weight principle (Wasow, 2002). Again, no statistics are given for active sentences.²

An obvious solution for a way to maintain a mapping between form and function while allowing more freedom of word order is to introduce some device like case into

²Seoane finds that only a minority (28%) of the passivized patients are more animate than the agent, and concludes that animacy has no influence on active/passive choice. She reports a similar null result for the person hierarchy. However, since there are no direct comparisons made with the proportions of animate or first/second person objects in active sentences, this conclusion is not warranted: for instance, the proportion of *active* sentences where the patients are more animate than the agent may quite plausibly be much less than 28%.

a language, to explicitly mark the function each phrase plays in the structure of the sentence. Although mainstream syntactic theories tend to assume that case has some central and universally standardized role, there is increasing interest in developing an understanding of case as a more ad-hoc functional solution to the problem of introducing additional cues where they are needed to specify the role a phrase plays in the sentence (see Nordlinger, 1998; Spencer and Otaguro, 2005). In fact, there is a striking cross-linguistic tendency for case marking to be more widespread on phrases whose roles are less obvious, prototypical, or predictable (Croft, 1988; Comrie, 1989; Aissen, 1999, 2003). This is known as *differential case marking*. Perhaps the most widely discussed example is the Spanish object marker *a*, which is commonly used to mark an animate direct object. Tippetts and Schwenter (2007) suggest that the measure which best predicts whether speakers will include this optional marker is in fact the *relative animacy* of the subject and object: the more animate the object and inanimate the subject, the more likely the *a* is to be included. Similarly, Japanese subject markers are more commonly dropped for animate than inanimate subjects (Fry, 2003). Where the animacy configuration of the arguments is canonical, fewer other sources of information tend to be given, perhaps because even without them the meaning is easily recoverable.

This phenomenon extends beyond animacy to other properties that tend to pattern with a certain grammatical function. In general, subjects are more agent-like, having “proto-subject” properties such as animacy and volition, while objects tend to show “proto-object” properties such as indefiniteness and affectedness (Tsunoda, 1985; Dowty, 1991; Van Valin, 2004). In Japanese, direct object markers are less commonly dropped on strongly definite NPs like proper names and pronouns (Fry, 2003). In Punjabi too, first and second-person objects (which are necessarily animate) require a case particle, while third-person objects take a case particle if they are definite but not indefinite (Shackle, 1972, p69-70). The same is true in Hindi, where animate objects in general are case marked but inanimates only when definite (Malchukov, 2008). Intuitively, first and second person referents are always human and definite, and hence animacy and definiteness cues alone indicate they are unlikely

to be objects.³

The tendency to add additional marking to less prototypical participants can be seen even in English. In the following examples from Anderson (1970) (discussed in Croft, 1988), the marker *at* indicates a less prototypical participant role: specifically, a less affected patient.

- (37) a. John shot Harry
 b. John shot at Harry

Prepositions play a similar role to case in providing additional sources of information about the role a referent plays in the meaning of the sentence. In cases like (37a), where John and Harry play prototypically agent-like and patient-like roles (shooting and being shot respectively), no additional markers are included. In (37b), where the action is less prototypically transitive (no-one gets shot), an additional object marker is included. There is cross-linguistic evidence, then, that markers such as case and adpositions often provide just one probabilistic source of information about the appropriate interpretation of a sentence: word order and simple world knowledge can provide similar information, and languages typically make use of all three in combination.

A final piece of evidence for this interplay between explicit morphological marking and other information sources comes from so-called “word order freezing” phenomena. Jakobson (1936) (discussed in Bouma, 2011) shows that while Russian word order is quite free, and case is typically relied on for the assignment of referents to agent and patient roles, where there is syncretism between cases such that case becomes ambiguous (in competition model terms, unavailable), word order steps in to determine interpretation.

³The discussion here is necessarily more superficial than this well-studied phenomenon deserves, and omits some ongoing debate about the extent and nature of case alternations. Notably, de Hoop and Malchukov (2008) have argued that differential subject and object marking pattern very differently typologically, arguing for an additional function of case marking, that of *differentiation* of arguments from each other, rather than straightforward *identification* of which has each role. Within the context of the current debate, the important point is that case is somewhat informative in many if not all languages that have it about the role of phrases in a clause, and a less probable assignment of roles is overall more likely to receive marking.

- (38) Mat' ljubit doč'
 mother-NOM/ACC loves daughter-NOM/ACC
 ‘*Mother loves daughter*’ (not: ‘*Daughter loves mother*’)

Similar “freezing” phenomena are quite common cross-linguistically, showing that word order apparently plays a role in determining interpretation even in strongly case-centric languages, although primarily so in situations where the case cue is unavailable or weak (see Lee, 2001 for Korean; Flack Potts, 2007 for Japanese; Bouma, 2011 for Dutch).

In summary, although different languages require their speakers to adopt different processing strategies, the typical situation is that multiple sources of information are considered together, even if some languages rely much more heavily on some cues than on others. The degree to which languages provide such sources of information also seems to depend on how far the utterance deviates from a canonical clause. A prototypical transitive clause specifies an agent-like referent acting on a patient-like referent (Hopper and Thompson, 1980; DeLancey, 1981; Comrie, 1989). When those roles can be easily guessed from meaning alone, neither case nor fixed word order are necessary, and many languages allow case omission or freer word order in exactly those situations.

5.2 Japanese word order and case

Japanese is fairly strictly verb-final, but other phrases (for instance, the subject, and direct and indirect objects) can appear in any order before to the verb. Rather than using word order to identify the subject and object as English does, Japanese arguments are marked with a final particle which gives information about its grammatical function. For instance, subjects are typically marked with the particle *ga*, and direct objects with *wo*.

Even though word order is relative free, in simple transitive clauses when a subject and object are both expressed, there is a strong tendency towards using the order SOV. Yamashita (2002) finds a mere 11 OSV clauses out of 2635 sentences examined in a mixed-genre corpus, although she does not report how many sentences in total

were transitive clauses with both arguments expressed, so it is not possible to state this number as a frequency for the OSV order. Even so, it is evident that OSV clauses are infrequent in the language as a whole, occurring in an estimated 0.41% (less than half of one percent) of sentences. Yamashita also reports counts for noncanonical ordering of the two objects in a ditransitive clause (DO before IO, rather than the more frequent IO DO order) and for so-called “long distance scrambling” in which an object from a subordinate clause is placed at the front of the sentence even though other material from the main clause then appears after the object and before the rest of the embedded clause. All of these noncanonical orders are infrequent: in total, she finds only 19 (0.72%) sentences that fit any of the three patterns.

Given the fact that SOV is much more frequent than OSV, word order in Japanese does carry a significant amount of information about the function roles played by phrases in a sentence. “First NP” is a good cue for subject and “second/last NP” a good cue for object. If comprehenders make optimal use of available information in processing, they should use this fact. Indeed, Mazuka et al. (2002) found that participants rated OSV sentences as more difficult and misleading than SOV sentences in an offline task. Yet there is only mixed evidence from *online* studies that comprehenders use any order information in determining grammatical function assignment. In self-paced reading experiments, OSV and SOV sentences are read equally quickly (Nakayama, 1995; Yamashita, 1997). Mazuka et al. (2002) find in an eyetracking self-paced reading experiment that the subject itself is read more slowly in OSV sentences, though reading times at other regions match those in SOV sentences. Miyamoto and Takahashi (2002) find slower reading times for ditransitive sentences in the less frequent S DO IO V order as compared to the standard S IO DO V order; again, they only find a reading time difference at the argument immediately before the verb, comparing across different types of phrase. Likewise, ERP studies have found that an initial object can elicit the LAN and P600 patterns of brain activity, which are often associated with a violation of syntactic rules (Ueno and Kluender, 2003; Hagiwara et al., 2007; Wolff et al., 2008). Taken together, these results suggest that Japanese comprehenders determine the grammatical function of argument NPs immediately and incrementally, without waiting for the verb (Sakamoto, 2001; see Schlesewsky

and Bornkessel, 2003 for an applicable model based primarily on related findings in German). Any processing complexity associated with the infrequent word order is experienced at the argument NPs themselves, and may well reflect an unexpected structure (say, OSV rather than an expected OV with no expressed subject) and not necessarily difficulty determining the role the NP plays in the sentence. Accordingly, Wolff et al. (2008) find that auditorily presented sentences with initial objects elicit brain activity associated with difficulty when there is a prosodic boundary after an initial object, but not otherwise. This boundary is commonly inserted in spoken OSV sentences, but not in OV sentences.

In summary, previous research has found that Japanese speakers report OSV sentences to be difficult, and ERP results and (a subset of) reading time studies have found evidence that this difficulty is due to the detection of an infrequent word order early on in the sentence. There is little evidence to suggest that comprehenders actually use word order as a cue to determining the grammatical function played by the phrases. The intuitive reason for this is that Japanese also has a rich case system, in which the subject and object are typically explicitly marked. In a series of offline experiments, Ito et al. (1993) (reported in Sasaki and MacWhinney, 2006) investigated whether word order was used to determine function in an offline interpretation task where case was withheld. They used the task of MacWhinney et al. (1984), in which participants are given bare words such as “*saw gorilla elephant*” and asked to interpret them as a sentence, answering a question like “who was the do-er?” Ito and colleagues presented Japanese strings of this form, manipulating the cues available to participants. They varied (i) case marking on the two nouns; (ii) animacy of the two nouns; and (iii) the order of the three words. Accordingly, they found that Japanese speakers relied most heavily on case marking to determine the agent, followed by animacy, and were the least reliant on word order. Ito and colleagues take this ordering of preference to reflect the validity of the three cues. Therefore, comprehenders do seem to be aware of the information carried by word order, even if they appear to use only case markers to guide their online comprehension in the normal case.

However, it would be somewhat surprising if comprehenders rely entirely on case particles in determining the role of phrases, since Japanese case particles can be

relatively ambiguous. For instance, the particle *ga* is most often a subject marker, but can also mark the object of a stative verb (39a) and can be used to merely mark the focus of a predicate rather than the subject proper, as in the so-called “double subject” construction (39b). The topic marker *wa* can mark the topic whether it is a subject or an object; therefore (39c) has two possible interpretations. The same is true for the particle *mo*, meaning “also” (39d). Finally, particle omission is relatively frequent (39e).

- (39) a. dare *ga* susi *ga* tukureru no
 who-GA sushi-GA can.make Q
 “Who can make sushi?”
- b. zou *wa* hana *ga* nagai
 elephant-TOP nose-GA long
 “Elephants have long noses”
- c. o isyasan *wa* yonda no
 HON-doctor-TOP called Q
 “Did (you) call a doctor?” / ? “Did the doctor call?”
- d. akio san *mo* mikaketa
 Akio-TOO saw
 “(I) saw Akio as well” / “Akio also saw (someone)”
- e. kanozyou asoko *ni* tureteiku
 she-Ø there-TO bring.go
 “(Someone) will take their girlfriend/her there” / ? “She will take (someone) there”

An even clearer example is the fact that Japanese exhibits word order freezing effects like those described for Russian above. Flack Potts (2007) discusses the *ga*-marked stative verb objects of (39a) above, and shows that the OSV interpretation is not possible when case marking is ambiguous (unavailable in competition model terms) as in (40a-b). Moreover, even when the SOV interpretation would be completely implausible, an OSV order is apparently disallowed (40c-d).

- (40) a. tarou ga hanako ga kowai
 Taro-GA Hanako-GA afraid
“Taro is afraid of Hanako” (Not: *“Hanako is afraid of Taro”*)
- b. tarou ga hanako ga osoreru
 Taro-GA Hanako-GA fear
“Taro fears Hanako” (Not: *“Hanako fears Taro”*)
- c. tarou ga zisin ga kowai
 Taro-GA earthquake-GA afraid
“Taro is afraid of earthquakes”
- d.* zisin ga tarou ga kowai
 earthquake-GA Taro-GA afraid
 Intended: *“Taro is afraid of earthquakes”*

In the terminology of the competition model, although Japanese case particles may have high validity, they do not reach a validity of 1.0. Particle-based cues are not always available, either due to there being multiple NPs marked with a given particle (39a,40) or none at all (39e). They are also not perfectly reliable, because a single particle may denote several different functions (39b-d,40). However, speaker judgments on the acceptability of *ga-ga* sentences like (40) and offline interpretations of sentences with case removed (Ito et al., 1993) suggest that at least in these marginal cases, other cues also influence interpretation, and these may step in to cover any insufficiencies of the case system as a cue to role identification.

Evidently, word order is not a primary cue in Japanese. This makes perfect sense: most of the time case marking fully determines the roles of the participants, making order redundant. However, where case is absent or ambiguous, comprehenders should fall back on probabilistic word order cues, just as they do in the competition model experiments for interpretation, and in word order freezing phenomena. The experiments reported later in this chapter indicate that this is the case in online processing as well as offline interpretation.

5.3 Comprehension as Bayesian inference about production choice

The competition model, like most probabilistic models of sentence interpretation, directly models the *comprehension* process. That is, it directly estimates the probability of a certain interpretation, conditioned on some set of features available in the input to the comprehender. An alternative would be to estimate the likelihood that the speaker would have *produced* the observed utterance given each of the possible meanings.

This distinction is closely related to the dichotomy of *generative* versus *discriminative* models in computational modeling. A discriminative model directly estimates the conditional probability $P(h|f)$ of a hypothesis h (in this case, an interpretation) given certain observable features f (the cues). Conversely, a generative model estimates the joint probability of the hypothesis and observable features $P(h, f)$. From this, a marginal distribution over hypotheses can be obtained with Bayes' law by normalizing over the probability assigned to that set of features under any hypotheses in the possible space H : $P(h|f) = \frac{P(h,f)}{\sum_{i \in H} P(i,f)}$. Similarly, an alternative to directly estimating the probability of a certain interpretation would be to estimate the probability of a speaker producing a certain sentence and its interpretation together, and then using that estimate to compare interpretations. This approach can be thought of as “speaker modeling”, or “comprehension by simulation” as it relies on the fact that the comprehender is also a speaker of the language and therefore has access to the approximately the same knowledge of language as the actual speaker of any given sentence. From a theoretical standpoint, this is an attractive property, since it means that the language user only needs to have one representation of the language, rather than one system for production and a separate system for comprehension.

The speaker-hearer's knowledge of language can be modeled as a directed acyclical graphical model, or Bayesian network (Narayanan and Jurafsky, 1998, 2002). A graphical model specifies the probabilistic relationships between all the variables which determine the final form of a sentence (including both meaning features like

animacy and form features like word order, case marking, etc). Based on these relationships, the model can be used to calculate the joint probability of any specific state of the network. To calculate a conditional probability, say, that of a meaning variable having some value m given that a form feature has some value f , we total up the joint probability of all states where the form features have value f and simply take the proportion of that probability mass given to states in which the meaning features have value m . Logistic regression models like those presented in chapters 2 and 3 are actually a form of graphical model of this kind, in which the conditional probability table (CPT) of the variable corresponding to the outcome (e.g. word order) assigns probability to its two outcomes according to the logistic function applied to a weighted sum of the values of its parent nodes.

Narayanan and Jurafsky (1998, 2002) suggested the usage of networks like these for modeling human language comprehension. In their model, they include both syntactic variables, such as those indicating the presence of a certain phrase or word type, and lexical/thematic variables such as the semantic fit between a verb and its argument and the argument structure that a verb expects. By connecting these variables with appropriate relationships between the variables, they were able to observe how a few words of input would change the distribution over the probabilities assigned by the network to the outcomes of each variable. For instance, they explored the predictions of the model when given the main verb/reduced relative ambiguous sentences studied by Trueswell et al. (1994). In these sentences, the first few words (e.g. “*the defendant examined...*”) are compatible with either a main verb continuation (“*...the document*”) or a reduced relative (“*... by the lawyer*”). Trueswell and colleagues found that when the sentence was continued as a reduced relative, comprehension difficulty varied depending on the degree to which the noun would have been a good object of the verb. Good objects (like “*evidence*”) biased the comprehender more strongly towards the correct reading than less good objects (like “*defendant*”), resulting in less difficulty. In Narayanan and Jurafsky’s model, hearing a particular verb changes the distribution over probable argument structures, tenses, and the semantic fit between subject and verb. These updated distributions would in turn influence the probability distribution over the main verb and reduced relative outcomes. Their model was able

to integrate several sources of information into a probability value for the outcome which turned out to be a good predictor of actual human comprehension difficulty. Narayanan and Jurafsky only discuss comprehension, but similar models could be developed which would apply to both production choices and comprehension as well.

From this perspective, the competition model terminology of mapping referents to “categories” on the basis of “cues” is not exactly appropriate. Rather, both categories like “agent” and cues like “nominative case” should be considered as *variables* over which inference is performed. The difference between the two amounts to whether a variable is *observed* or *unobserved*. In production, properties of meaning are observed, since they are specified in the meaning to be expressed. The speaker’s job is to determine a distribution over values of the unobserved variables which determine linguistic form, such as case and word order. In comprehension, form variables can be directly measured from the input, so they are observed. The hearer’s job is to determine the most probable configuration of the unobserved variables which determine meaning, and so determine the most likely intended interpretation. I will refer to this proposal as “comprehension as inference about production”. With this in mind, consider a simple model for Japanese in Figure 5.1. This model needs some explanation, especially as the notation has been simplified somewhat. Each node in the graphic corresponds to some variable, which at any given moment may be observed or unobserved. For all nodes that have no incoming edge (arrow), we specify a prior distribution over the possible values that the variable can take. For nodes with incoming arcs, we specify the distribution over values of the variable to be a function of only those variables connected to the node by an incoming edge. Thus the graph structure displays the *independence* between variables: only variables connected by edges can directly influence each other, although indirect information about one variable might be given by another through a longer chain involving other variables.

In this particular model, we consider only two properties of the subject and of the object: their length in words, and whether they are animate. Both of these properties are known to affect word order: animate entities are more likely to be placed first (Branigan et al., 2008) as are longer NPs (Yamashita and Chang, 2001). Additionally, the probability that the subject and object are animate is very different: this will be

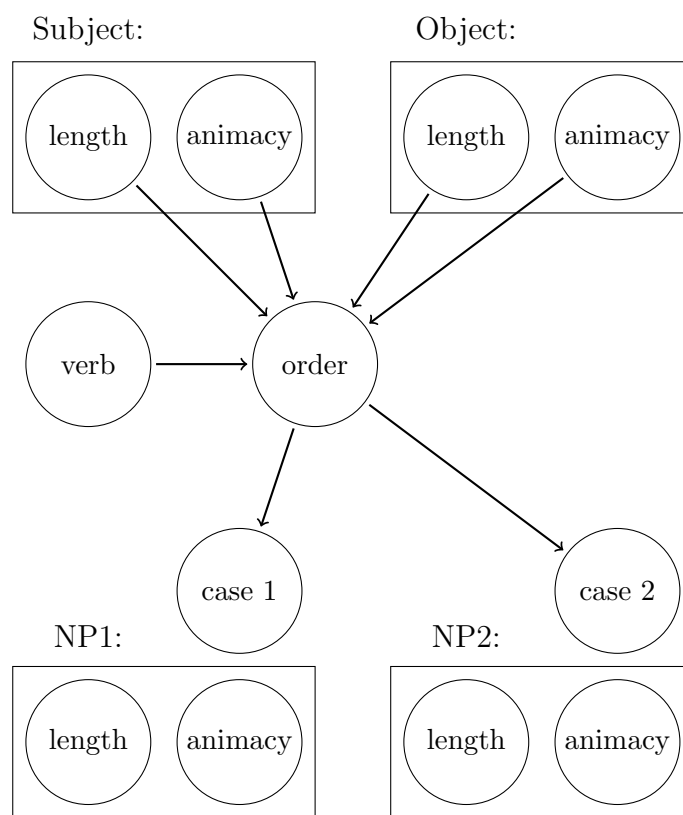


Figure 5.1: Sketch of graphical model for speaker's choice between SOV and OSV orders in Japanese.

represented by setting the prior probability of each node differently, as follows:

| variable | type | prior |
|-------------------|---------|------------------|
| subject length -1 | integer | $Pois(\lambda)$ |
| subject animacy | boolean | $Bern(\omega_s)$ |
| object length -1 | integer | $Pois(\lambda)$ |
| object animacy | boolean | $Bern(\omega_o)$ |

I set the prior distributions over the animacy of the subject and object to be Bernoulli distributed with probabilities ω_s and ω_o respectively, meaning that the expected proportion of animate subjects and objects are ω_s and ω_o respectively. These are parameters of the language, which should be set to the proportion of all subjects and of all objects that are animate. Without any particular reason to think that the subject and object will differ in length, I set the prior on both to be Poisson distributed with some parameter λ to be set according to the distribution of NP lengths in the language (the -1 is a correction to make this distribution proper, since the Poisson distribution assigns mass to an outcome of 0, which is not a possible length). The verb node is a multinomial outcome over all Japanese verbs, and its prior should be set according to the empirical frequency distribution. Given values for these variables, we can specify the probability of each order outcome conditioned on those variables:

$$\begin{aligned}
 P(\text{order} = \text{SOV}) &= \frac{1}{1 + e^{-(\alpha[\text{verb}] + \beta_1 \cdot s \text{ len} + \beta_2 \cdot s \text{ anim} + \beta_3 \cdot o \text{ len} + \beta_4 \cdot o \text{ anim})}} \\
 P(\text{order} = \text{OSV}) &= 1 - P(\text{order} = \text{SOV})
 \end{aligned}$$

This is just the standard logistic regression formula. The free parameters α and $\beta_{1..4}$ correspond to a regression's intercept and coefficients respectively. I allow different values of α for each verb, as in a multilevel regression model (Gelman and Hill, 2006). This allows for “verb bias”, verb-specific differences in the preference for one order or the other. Again, these need to be estimated from empirical data.

However, values are expected to be generally greater than 0, since this intercept term is where the language-wide bias for SOV order is implemented. The coefficients for the four predictors could also be estimated from corpora or lab experiments. β_1 and β_2 are expected to be positive and β_3 and β_4 negative, implementing the preference for animate and long phrases to come early. If it turns out that those preferences are entirely independent of whether the phrase is a subject or object, we could equate $\beta_3 = -\beta_1$ and $\beta_4 = -\beta_2$.

The final part of the model reflects “surface” properties of the sentence, the properties of the first and second NP. Of course, when the order is SOV, the first NP’s properties are equated with the subject’s properties, and for OSV the objects’ properties. The converse holds for the second NP. This means that each “surface” node has three dependencies: one on the order outcome variable, and one on the corresponding property (length or animacy) for both the subject and the object. I have omitted these arrows in the diagram for clarity.

| order | NP1 length = | order | NP2 length = |
|-------|-----------------|-------|-----------------|
| SOV | subject length | SOV | object length |
| OSV | object length | OSV | subject length |
| order | NP1 animacy = | order | NP2 animacy = |
| SOV | object animacy | SOV | object animacy |
| OSV | subject animacy | OSV | subject animacy |

The conditional probability table for the possible values of the case marking variables are just as simple. Case 1 refers to the particle on the first NP, and case 2 to the particle of the second NP.

| order | $P(case1 = \cdot)$ | | order | $P(case2 = \cdot)$ | |
|-------|--------------------|-----|-------|--------------------|-----|
| | ga | wo | | ga | wo |
| SOV | 1.0 | 0.0 | SOV | 0.0 | 1.0 |
| OSV | 0.0 | 1.0 | OSV | 1.0 | 0.0 |

Of course, a fuller model would also condition on the verb, assigning probabilities to outcomes like *ga* and *ni* marked objects depending on the verb type, and allowing NPs with no case marking at all.

In production, the subject and object property variables are observed, as is the verb. The speaker then performs inference over the model (essentially, plugs the appropriate values into the regression which determines the probability at the word order node) to determine the probability of choosing one order or another. The speaker then chooses SOV or OSV with the resulting probability, fixing the value of the order node. In turn, the order then fully determines the properties of NP1 and NP2, including case marking.

In comprehension, the only observed nodes are the NP1 and NP2 properties and the verb. In the typical case, this includes the case variables, and because I have made case deterministically dependent on word order in this example, the value of the order node can be inferred exactly: for instance,

$$\begin{aligned}
 P(\text{order} = \text{SOV} | \text{case1} = \text{ga}) &= \frac{P(\text{ga} | \text{SOV})P(\text{SOV})}{P(\text{ga} | \text{SOV})P(\text{SOV}) + P(\text{ga} | \text{OSV})P(\text{OSV})} \\
 &= \frac{1.0 \cdot P(\text{SOV})}{1.0 \cdot P(\text{SOV}) + 0.0 \cdot P(\text{OSV})} \\
 &= 1.0
 \end{aligned}$$

Thus when case is observed, this model assigns a probability of 1.0 to one interpretation or the other regardless of the animacy or length of the arguments. If we map these probabilities to processing complexity, the model predicts that varying other cues like word order or animacy will not lead to difficulty.

However, even if case is not observed, the model can still make inferences about the most probable interpretation. For instance, using subscripts a and i to indicate animate and inanimate respectively on NP1, NP2, subject (S) and object (O):

$$\begin{aligned}
 &P(\text{order} = \text{SOV} | \text{NP1}_a, \text{NP2}_i) \\
 &= \frac{P(\text{SOV}, \text{NP1}_a, \text{NP2}_i)}{P(\text{SOV}, \text{NP1}_a, \text{NP2}_i) + P(\text{OSV}, \text{NP1}_a, \text{NP2}_i)}
 \end{aligned}$$

$$\begin{aligned}
&= \frac{P(\text{SOV}, S_a, O_i)}{P(\text{SOV}, S_a, O_i) + P(\text{OSV}, O_a, S_i)} \\
&= \frac{P(\text{SOV}|S_a, O_i)P(S_a)P(O_i)}{P(\text{SOV}|S_a, O_i)P(S_a)P(O_i) + P(\text{OSV}|O_a, S_i)P(O_a)P(S_i)}
\end{aligned}$$

At this point, we can use the regression equation to determine the conditional probabilities of the two word orders. In this example, we would have to first calculate expected values for the length and $\alpha[\text{verb}]$ variables since they are unobserved, though in a more realistic example those values would have been observed together with animacy, so we would be conditioning on them as well. Together with the prior distributions over animacy defined above, this is enough to determine the probability of the SVO order even without observing case. Since $P(S_a)$ and $P(O_i)$ are likely to be much higher than $P(S_i)$ and $P(O_a)$ thanks to ω_s, ω_o , and since $P(\text{SVO}|\dots)$ will be higher than $P(\text{OSV}|\dots)$ thanks to α , we expect the probability of SOV to be very high, if not quite as high as 1.0. Thus the model uses both animacy and word order in determining the interpretation with high certainty. However, if the NP1 and NP2 animacy assignments were opposite, then they would act against the SOV bias, making the probabilities assigned to the two orders closer together. In that case, we might expect processing difficulty for an $O_i S_a V$ sentence when the particles are not available.

I now report some experiments designed to test this prediction of the model.

5.3.1 Study 1: Case and word order in comprehension

I ran a self-paced reading experiment to investigate whether Japanese speaking comprehenders do in fact rely on word order when case cues are unavailable, as predicted by the model above. This standard paradigm involves presentation of written sentences on a computer screen piece by piece. To continue to the next part of the sentence, comprehenders must press a key, which causes the currently displayed part to be hidden and the next word(s) to appear. By timing the delay between key

presses, we can observe the amount of time taken to read any particular part of the sentence, a behavioral measure of comprehension difficulty (see Just et al., 1982). In this way, I can test whether sentences with canonical SOV order lead to faster reading times than those with OSV order.

Design and materials

I produced a set of 16 Japanese sentences, each with a subject, object and transitive agent-patient verb. In all cases, the subject was animate and the object inanimate, and the intended meaning was not reversible (that is, the object was not a possible agent for the action) making the meaning fully unambiguous even without case marking or order information. All subjects were common given names, and the objects were high frequency concrete objects. These sentences were embedded as complements of verbs of communication, such as “said”, “heard”, “reported” and so on, with given names as the subject. This main clause had the form “... to *name verb*-ta” (“*name verb*-ed that ...”). Note that in Japanese, the main clause comes *after* the embedded clause. An example item is given in (41) with bars indicating the region boundaries. Actual presentation was done in Japanese script (which has no spaces between words).

- (41) tarou ga | mizu wo | nonda to | nomura san ga | itta
 Taro-SUBJ | water-OBJ | drank-THAT | Nomura-SUBJ | said
“Nomura said that Taro drank the water”

Given sentences which differ in the word order, the crucial behavioral measure in this experiment is the reading time at or just after the embedded verb. This region is of interest because it is at this point that the interpretation of the sentence becomes clear, even in the absence of case markers: up until this point, the sentence is compatible with interpretations like “the water splashed Taro”, or similar. As discussed above, a well established literature in psycholinguistics has shown that words which are compatible with linguistic structure which is expected by comprehenders tend to be processed more quickly (see Levy, 2008, for a survey). Therefore, if comprehenders assume an initial case-less NP is the subject and the second the object, they should

read the disambiguating verb quickly in the SOV condition relative to the OSV condition. This would confirm that comprehenders are making use of order information, at least when case is unavailable. The case-absent conditions will serve as a control.

The primary purpose of the final embedding verb was to add material after the verb of interest. It is difficult to measure participants' reading times on the final word in a sentence, since they may not hit a key again immediately, knowing there is no more material to come. It is also possible that adding slightly more complexity to the sentences will help to stop participants pacing through every sentence without paying much attention and then relying on memory to answer the questions.

Four versions of each sentence were produced, corresponding to the four experimental conditions. The word order variation was introduced by switching the order of the first two regions. The case particle variation was introduced by replacing the particles *ga* and *wo* in those regions with a “missing” symbol, which was a small circle appearing in the sentence where the particle should have appeared. Participants were instructed to expect this symbol in place of a particle in some sentences, and trained on example sentences with the symbol before the experiment began. The motivation for using this instead of simply removing the particle was that although Japanese allows particle omission in certain situations, long sentences like these with two missing particles would be unusual. Therefore, I wanted to avoid participants thinking that phrases they read with no particle were “natural” particle omissions, which might signal a certain kind of upcoming structure. However, in study 3 I will report a test of stimuli which simply removed the particle, obtaining similar results.

A yes/no comprehension question was devised for each question. Questions were designed so that a roughly equal number questioned the subject, embedded verb, object, and main clause subject/verb. The correct response to half of the questions was “yes” and to half “no”. This was done to ensure participants read the sentences carefully.

Participants

Participants were recruited through word of mouth in Tokyo and around the Stanford University campus in California. All participants were native Japanese speakers and

had not learned English or another foreign language before attending school at the earliest. Ages ranged from 18 through around 40, with almost all being college age. Participants were paid \$7 (or approximately the yen equivalent) for their time. In total, 25 participants took part.

Methods

The experiment was conducted using Linger, a freely downloadable experiment presentation software system written by Doug Rohde. In addition to the 16 experimental items, a further 32 filler sentences of similar complexity were created. The 16 experimental items were assigned to conditions following a latin square design, so that each participant saw one quarter of the items in each condition. Items were presented in a random order with at least one intervening filler sentence between experimental items.

On each trial, participants pressed the space bar to reveal the first region of the sentence. Each subsequent press hid the currently displayed region and revealed the next. When the last region was hidden, the comprehension question was displayed on the screen. Participants responded by pressing a key for “yes” or “no”. Feedback was given for incorrect responses only. Breaks were offered one third and two thirds of the way through the experiment. The entire procedure took just over 20 minutes on average. I remained in the room in the room in a seat where I could see the participant but not the screen throughout.

Results

Reading times were analyzed by linear multilevel regression with random intercepts for item and subject (see Gelman and Hill, 2006; Jaeger, 2008). Datapoints which exceeded the per-region per-condition mean by 3 standard deviations were excluded (2.00%) as were trials with the following comprehension question answered incorrectly (5.25%). The amount of data removed from each condition did not vary greatly. In the full dataset, the proportion of questions answered correctly was distributed as follows:

| | | Particles | |
|-------|-----|-----------|--------|
| | | Present | Absent |
| Order | SOV | 97% | 94% |
| | OSV | 91% | 97% |

At the critical region (the embedded verb) there was no significant main effect of order ($\beta = -22.8, p = .558$) but effects of particle absence ($\beta = 125, p = .0014$) and an interaction ($\beta = 117, p = .0345$). Together these meant that particle absent sentences were read slower at the verb, and the particle absent/OSV sentences slower still. The condition means are plotted in figures 5.2 and 5.3.

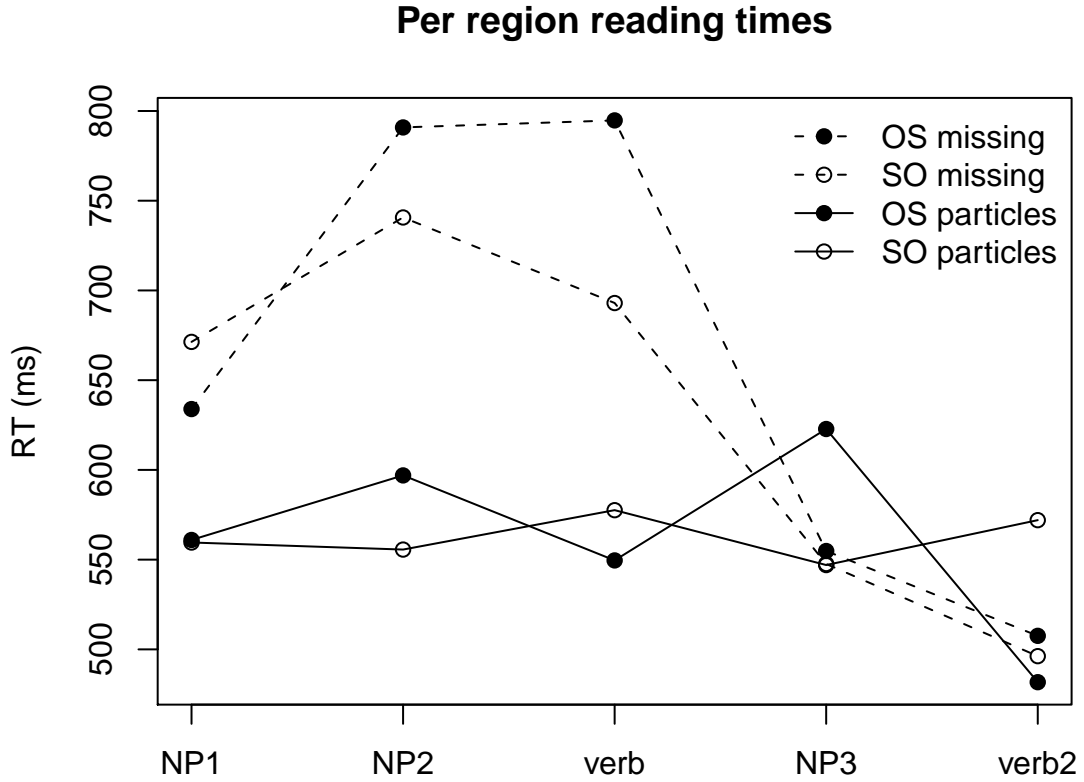


Figure 5.2: Reading times at each region in study 1.

At the initial NP, there was a significant effect of particle absence ($\beta = 95.3, p < .001$). At the second NP, there was an effect of particle absence ($\beta = 186, p < .001$) and a marginal effect of order such that the subject was read more slowly than the object in second position ($\beta = 55.2, p = .0588$). At the post-verb region (the

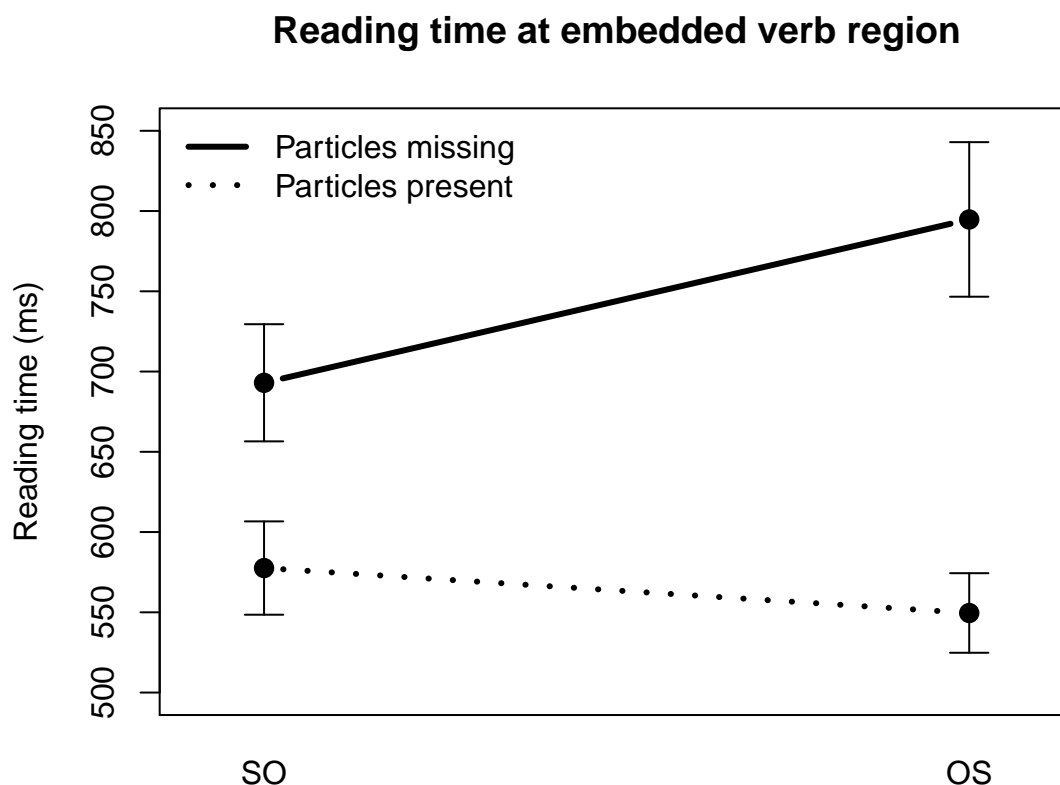


Figure 5.3: Reading times at embedded verb in study 1. Bars show standard error of the mean.

main clause subject) there was a marginal effect of order such that SOV sentences were read faster ($\beta = 46.9, p = .0622$). At the final verb region there were effects of particle absence ($\beta = -80.7, p = .0334$), order ($\beta = -94.5, p = .0144$) and an interaction ($\beta = 110.09, p = .0413$) such that the SOV/particles-present condition was read somewhat slower than the other three. All other effects were non-significant ($ps > .1$).

Discussion

In line with previous findings, there was no difference in reading times (RTs) between the OSV and SOV word orders when case particles were present. Removing the case particles increased reading times substantially, perhaps because participants had to think more about the probable function of each NP in the sentence, though perhaps

also simply because the sentences with “missing” symbols were less familiar. The fact that this slowdown was observed at the verb, which was identical across conditions, suggests the former explanation may be more likely, although it is possible that increased RTs there were due to “spillover” from slower processing of the immediately preceding NP. The most interesting result, however, is the fact that in the particle-absent conditions there was a significant slowdown at the verb when the object was presented before the subject. This is captured statistically by the interaction between the two manipulations. This effect has not been previously reported, and suggests that Japanese comprehenders may in fact make use of probabilistic order cues, but only to an extent which is observable when more reliable sources of information (case) are not available.

5.3.2 Study 2: Interaction of NP type with other cues

While most Japanese transitive verbs take a direct object which is marked with *wo* (when the particle is included, and when the object is not instead marked as a topic), some verbs take an object marked with *ni*. This particle is also used to mark the recipient or of dative verbs, the causee of causatives, the goal of movement verbs, and most types of adverbial phrase. Typically these objects are less affected by the action denoted by the verb. For instance, *sawaru* (“touch”), *suwaru* (“sit (on)”), *sitagau* (“follow (instructions)”) and *butukaru* (“bump into”) all take *ni*-marked objects. Thus, these less affected (and hence less prototypical) objects receive differential case marking.

The difference between *wo*- and *ni*-marked objects could influence comprehension difficulty in two ways. First, because *ni* has more functions than *wo*, there may be some (probabilistic) uncertainty about the function of a *ni* marked NP even when the particle is present. While *wo* always indicates a direct object, *ni* also marks locations, instruments, times and other adverbials as well as goal and recipient arguments. Second, the *ni*-marked objects themselves tend to be less prototypically object-like, and the function the NP plays in the sentence may be less predictable, or at least less predictable before seeing the verb. If prototypicality acts as a cue for interpretation,

as psycholinguistic work in the probabilistic constraints framework has shown (see e.g. Trueswell et al., 1994; MacDonald and Seidenberg, 1999), then the comprehender may find it slightly harder to correctly interpret these NPs as objects.

These properties are not directly modeled in the simple Bayesian model described previously: however, it would be very easy to add new variables for subject and object prototypicality that would work in much the same way as the current animacy variables.⁴

Design and materials

I produced a set of 20 items with the same basic structure as those used in study 1. Ten of these were chosen to have verbs taking an object marked by *wo*, and ten had an object marked by *ni*. I also made two small changes to the structure of the sentences which resulted in them being slightly more complex than those in the previous study. First, a (time or location) adverb phrase was placed at the beginning of each sentences. This was done simply to make the sentences comparable in length to those used in another experiment being run back-to-back with this one, which is not reported here. Second, instead of using simple proper names and single nouns for the subjects and objects, I used slightly longer phrases (each one of the form “noun *of* noun”) for both. Two example items are shown below. (42a) has an object marked with *ni*, while (42b) has an object marked with *wo*.

- (42) a. kouzigenba no naka de | hassai no kodomo ga |
 construction.site-OF inside-AT | eight.years-OF child-SUBJ |
 denki no keeburu **ni** | sawatta to | komatu san ga | siraseta
 electric-OF cable-NI | touched-THAT | Komatsu-SUBJ | informed
 “*Komatsu informed (us) that an eight-year-old child touched the power cable
 at the construction site*”

⁴In the absence of any evidence that subject and object prototypicality influence word order directly (rather than through a correlation with accessibility properties such as animacy and definiteness), I would probably not model an influence of prototypicality on the order outcome, but instead just include these variables under the subject and object properties. Naturally, the prior probabilities for these variables would be very different for the subject and object NP, so information about order would be carried indirectly once the NP1 and NP2 prototypicality properties are observed in comprehension.

- b. itiniti no uti ni | hikikomori no syounen ga | saisin no
 one.day-OF inside-IN | stay.at.home-OF boy-SUBJ | latest-OF
 syousetu **wo** | yomikitta to | watanabe san ga | nobeta
 novel-WO | read.finished-THAT | Watanabe-SUBJ | mentioned
 “*Watanabe mentioned that the reclusive boy read the latest novel(s) within a day*”

Again, the particles-missing conditions were identical except that the final particles *ga*, *wo* and *ni* were replaced with the “missing” symbol. The verbs chosen were approximately balanced for frequency between the *ni* and *wo* conditions using counts from the Japanese Google search engine.

Participants

Participants were recruited, selected, and paid in the same way as study 1. In total, 25 participants took part.

Methods

Presentation of the materials was identical to Study 1. The 20 experimental items were randomized among 40 fillers with the constraint that no two experimental items were adjacent.

Results

Reading times were analyzed as in study 1. Outliers at three standard deviations from the per-condition per-region mean accounted for 1.8% of the data and incorrectly answered questions 9.8%. In the full dataset, the proportion of questions answered correctly was distributed as follows:

| <i>wo</i> -marked object | | | | <i>ni</i> -marked object | | | |
|--------------------------|-----|-----------|--------|--------------------------|-----|-----------|--------|
| | | Particles | | | | Particles | |
| | | Present | Absent | | | Present | Absent |
| Order | SOV | 91% | 93% | Order | SOV | 96% | 91% |
| | OSV | 87% | 85% | | OSV | 91% | 87% |

At the critical verb region, there was no main effect of particle absence ($\beta = 33.3, p = .524$) or order ($\beta = -3.68, p = .944$) but a marginally significant interaction ($\beta = 143, p = .0535$) such that the particles-absent/OSV conditions were read more slowly. Additionally, there was a main effect of object type such that *wo*-marked objects were read more quickly ($\beta = -104, p < .01$). Condition means are plotted in figures 5.4-5.6.

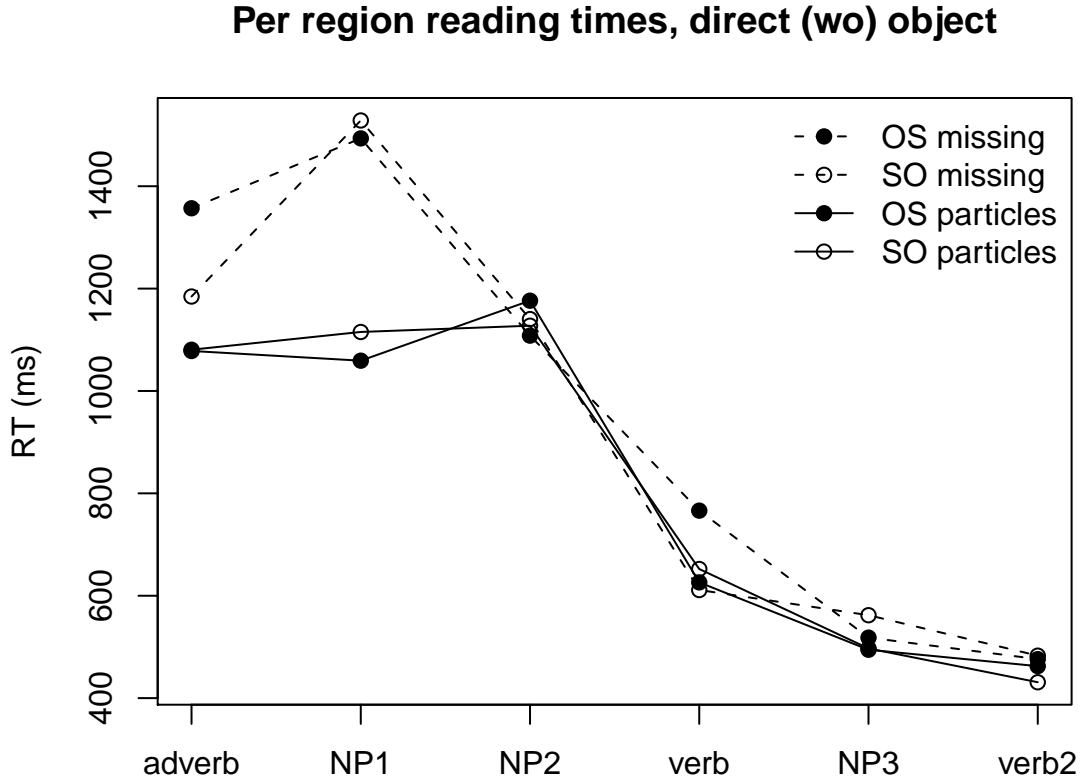


Figure 5.4: Reading times at each region in study 2 for direct (*wo*-marked) object conditions.

At the initial adverb, there was a main effect of particle absence ($\beta = 183, p = .0322$) and order ($\beta = -252, p = .037$). There was no significant effect of object type ($\beta = 49.0, p = .751$), but this interacted with order ($\beta = 337, p = .0477$) such that *wo*-marked/OSV conditions were read slower. At the first NP, there was an effect of particle absence ($\beta = 307, p = .0015$). At the second NP there was a marginal effect of object type such that *wo*-marked objects led to faster RTs ($\beta = -239, p = .0534$).

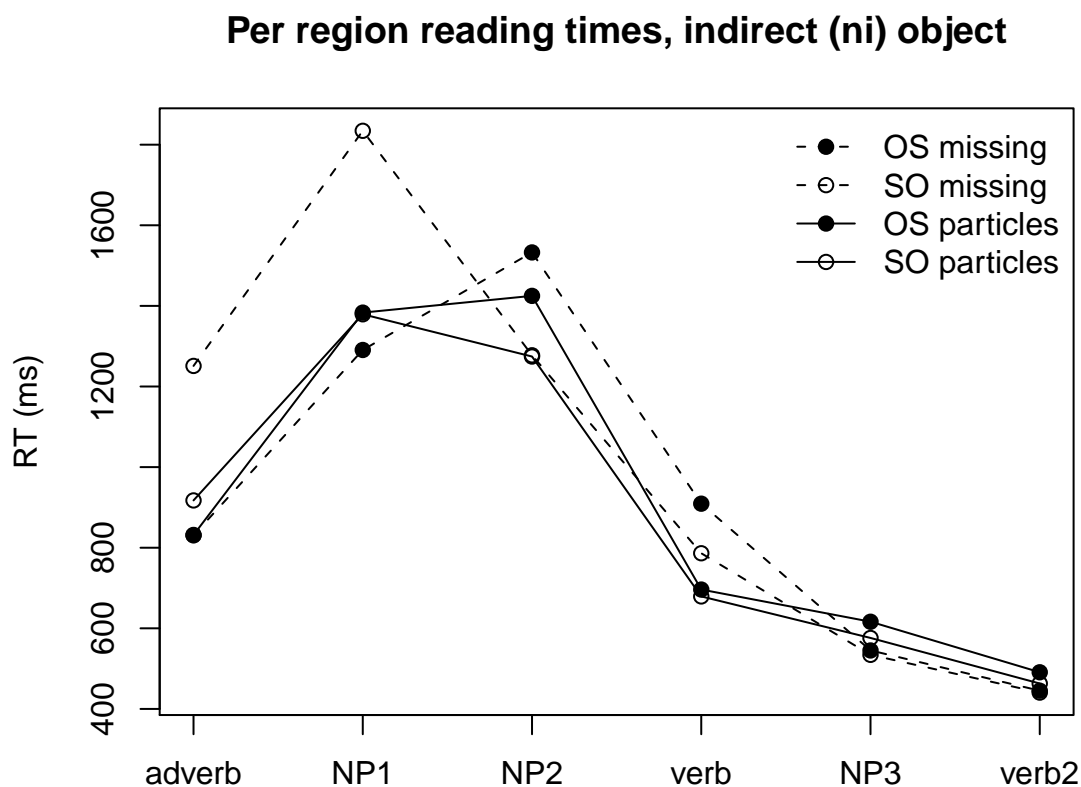


Figure 5.5: Reading times at each region in study 2 for indirect (*ni*-marked) object conditions.

At the post-verb main subject there was a marginal effect of object type such that *wo*-marked NPs led to faster RTs ($\beta = -50.0, p = .0688$). There were no other significant effects (all $ps > .1$).

Discussion

The *wo*-marked object conditions can be compared directly to the results of study 1, since they differ only in minor ways: the NPs are slightly longer and more complex/less accessible, and there is an additional adverb phrase at the start of the sentence. Accordingly, the pattern of results is extremely similar: at the verb, there is an interaction such that the missing particles/OS order condition is slower than the other three. In fact, the results are even more striking than in study 1, since there now appears to be no further effect of particle presence beyond this interaction: the

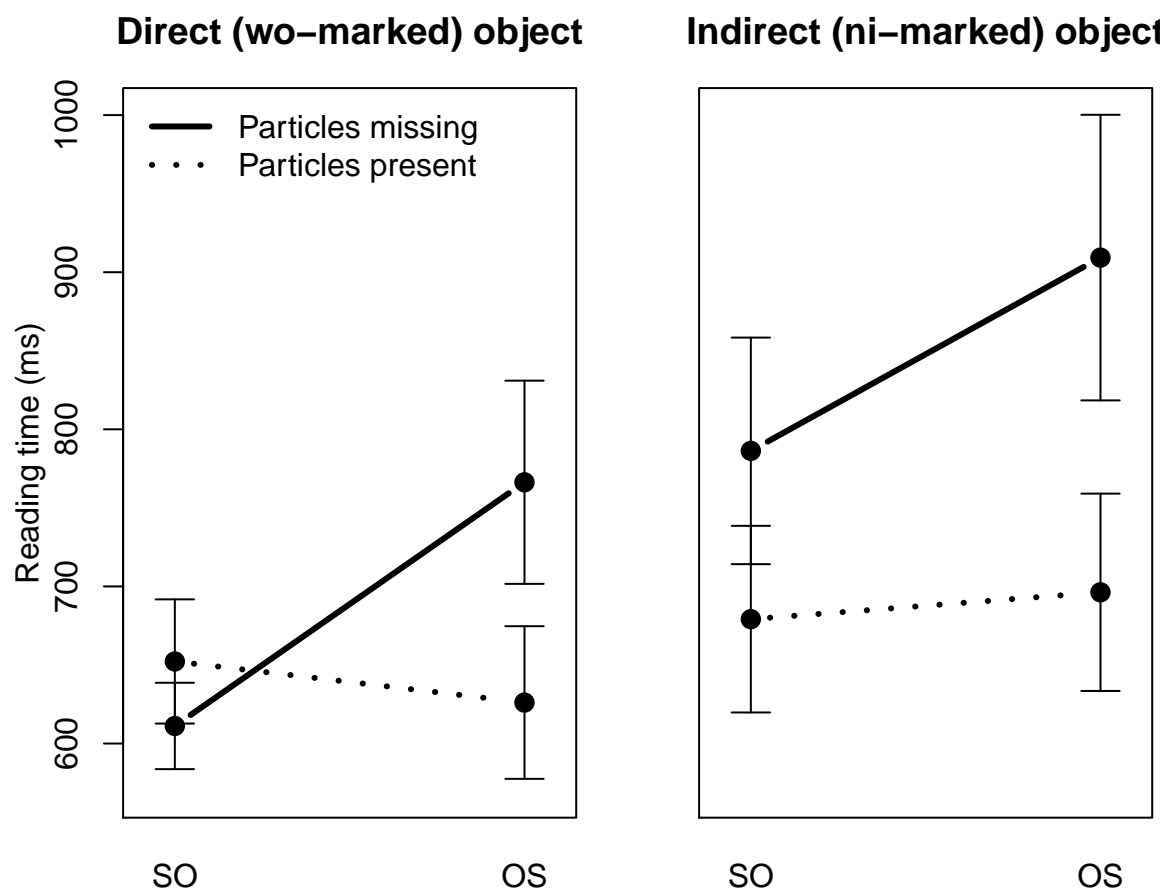


Figure 5.6: Reading times at embedded verb in study 2. Bars show standard error of the mean.

particles missing/SO order condition is read at the same speed as the two particles present conditions. This difference may be due to there being less spillover from the preverbal NPs, since in this study the phrases were longer. One additional difference is that the second NP shows no effects of condition at all here, while in study 1 the particles missing conditions were read more slowly. This may be due to participants having more time to get used to the particle-less phrases, since they had longer NPs and an additional (particle-less) adverb in these conditions. In any case, this makes it less likely that any effect seen at the verb is a spillover from previous phrases.

Turning to the *ni*-marked object conditions, a similar pattern emerges. The chief

difference between the two conditions is that there is a significant main effect of object type: *ni*-marked objects lead to slower reading times at the verb than *wo*-marked objects. This is expected, since *ni*-marked objects are less prototypical, and hence less expected, as objects.⁵

One unexpected effect is that the object is read very quickly in the particles missing/OS/*ni*-marked condition, in contrast to all the other 7 conditions. This may just be statistical fluke, especially as the initial adverb is also read quickly in that condition. Even so, it is possible that it reflects some kind of misanalysis of initial *ni*-marked objects as subjects when the particle is missing. The higher reading times on the initial adverb in the other three particles-missing conditions are expected because the final particle *ni* was missing on the adverb in those conditions too. This was done for consistency, since *ni* particles were removed from the object in the missing conditions.

5.4 “Explaining away” of mutually informative cues

Cue-based comprehension accounts such as the competition model and probabilistic constraint-based models must specify the way in which comprehenders combine the multiple sources of information available to them. The competition model as originally implemented assumes that each cue contributes *independent* information: for instance, the probability of a certain NP being the agent is the product of the probability that it is the agent given its word order, the probability that it is the agent given its case, the probability that it is the agent given its animacy, and so on (see e.g. McDonald and MacWhinney, 1989). In the general case, this is not mathematically correct, since if any cues are mutually informative (say, such that animates tend to appear in a certain word order position more often than inanimates) then the information each one carries will be “overcounted”. The simplifying strategy of assuming independence is a common trick for making complicated probabilistic models tractable

⁵As *ni*-marked direct objects occur with a smaller set of verbs than *wo*-marked objects, they might be expected to increase expectation for the upcoming verb. However *ni* is also the particle used for recipients and causees, and many types of obliques and adverbials. Therefore seeing a *ni*-marked NP probably actually constrains the parse less than an *wo*-marked NP in most cases.

in the computational literature, where it is known as the “naïve Bayes” assumption (see Jurafsky, 2003).⁶

Although it is a priori plausible that human comprehenders could make the same simplification, work in constraint-based parsing frameworks has shown quite clearly that information sources are not treated independently, at least in determining between multiple ambiguous structures. For example, Spivey-Knowlton and Sedivy (1995) show that prepositional phrases which are structurally ambiguous between modifying a verb (“shot the cowboy *with the bow and arrow*”) or a noun (“shot the cowboy *with the leather vest*”) are resolved on the basis of lexical properties of the verb (action verbs like *shoot* preferring NP attachment, perception verbs like *see* preferring VP). However, this is modulated by NP definiteness: when the object NP is indefinite, the effect of the verb type is much smaller. Therefore, at least two properties interact in a complex way in determining the interpretation of an ambiguous sentence. Moreover, comprehension difficulty associated with this interaction of information sources can be observed in online experiments. Garnsey et al. (1997) studied the direct object/sentential complement ambiguity (“*the senator regretted the decision ... immediately/had been made public*”). They found that both verb bias and noun plausibility affected comprehension difficulty: verbs that more commonly take a sentential complement (SC) led to faster reading at the SC continuation, and nouns that were plausible as a direct object led to slower reading. However, these factors interacted such that the smaller the verb’s bias towards one structure or the other, the stronger the effect of object plausibility. In other words, when one cue (verb bias) was not informative, another cue (object plausibility) was weighted more highly by comprehenders.

Relatively less research has demonstrated an interaction between information sources in sentences *without* (temporary) ambiguity. One potential source of evidence that there is such an interaction is from the role of context in comprehension.

⁶The competition model was elaborated with a second cue validity measure, *conflict validity*, which attempted to address these shortcomings (see McDonald, 1987). This is less parsimonious in that it assumes two different constructs, although a certain amount of research was prompted looking for evidence for both simple cue validity and conflict validity in human behavior (see e.g. McDonald, 1986).

While it is well known that context influences the difficulty experienced when comprehending ambiguous structures (Crain and Steedman, 1985; Altmann and Steedman, 1988), recent work by Kaiser and Trueswell (2004) suggests that in unambiguous SVO/OVS sentences in Finnish, discourse context also reduces the processing complexity normally associated with the infrequent OVS order. Specifically, Kaiser and Trueswell found that when the object was given by the context and the subject new, OVS sentences were read faster than when the subject was new and the object given. They understood these results as showing that the OVS order is associated with a particular given-new discourse structure, and using it with new-given order violates expectations. This explanation is broadly compatible with an interpretation as an interaction between information sources. In Finnish, word order is correlated with grammatical function and therefore provides a (probabilistic, partial) source of information for identifying the subject and object. Because of this, cue-based models suggest there may be some processing complexity associated with clauses in which the first phrase is not the subject. This complexity reflects the probability that the comprehender (incorrectly) places on that NP being the subject, which itself reflects the strength of word order as a cue to subjecthood. However, word order is more strongly associated with information status of the NPs: Kaiser and Trueswell discuss the fact that the OVS order is strongly associated with a given object, while the SVO order is relatively neutral. If comprehenders reason directly about what leads speakers to make the choices they do, then we predict a “chain” of inference in comprehension: the NP is given; therefore the speaker placed it first; therefore the act of placing it first does not indicate that it is a subject. If this were the case, then information status would modulate the degree to which the comprehender uses word order as a source of information about grammatical function, yielding behavioral effects like those reported by Kaiser and Trueswell.

Such chains of reasoning automatically fall out of a Bayesian model like the one discussed above. In the inference literature, this kind of effect is called “explaining away”. For instance, first consider a situation in the Japanese model above where the first NP is inanimate and 2 words long, and the second NP animate and also 2 words long. I will use $NP1_i$, S_i and so on as abbreviations for the length of NP1, the

subject, and so on.

$$\begin{aligned}
& P(\text{order} = \text{SOV} | \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2) \\
&= \frac{P(\text{SOV}, \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2)}{P(\text{SOV}, \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2) + P(\text{OSV}, \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2)} \\
&= \frac{P(\text{SOV}, \text{S}_i, \text{O}_a, \text{S}_l = 2, \text{O}_l = 2)}{P(\text{SOV}, \text{S}_i, \text{O}_a, \text{S}_l = 2, \text{O}_l = 2) + P(\text{OSV}, \text{O}_i, \text{S}_a, \text{O}_l = 2, \text{S}_l = 2)} \\
&= \frac{P(\text{SOV} | \text{S}_i, \text{O}_a, \text{S}_l = 2, \text{O}_l = 2) P(\text{S}_a) P(\text{O}_i) P(\text{S}_l = 2) P(\text{O}_l = 2)}{P(\text{SOV} | \text{S}_i, \text{O}_a, \text{S}_l = 2, \text{O}_l = 2) P(\text{S}_i) P(\text{O}_a) P(\text{S}_l = 2) P(\text{O}_l = 2) + P(\text{OSV} | \text{O}_i, \text{S}_a, \text{O}_l = 2, \text{S}_l = 2) P(\text{O}_i) P(\text{S}_a) P(\text{O}_l = 2) P(\text{S}_l = 2)}
\end{aligned}$$

Again, we are left calculating the probability of SOV order being the probability assigned to SOV under the logistic regression with NP1 properties taken as subject properties and NP2 properties taken as object properties, multiplied by our prior belief that the subject and object would have those properties in the first place. This is then normalized by dividing by the total probability placed on both the SOV and OSV orders calculated in the same way. To make a concrete prediction, we will assume the following parameter estimates (recall that these should be fixed from corpus or experimental estimates).⁷

⁷These parameter estimates are set fairly arbitrarily, but at least their directions (positive or negative for the coefficients, greater or less than .5 for the animacy probabilities) are known from previous literature to be as set.

| parameter | value | probability | parameter | value |
|------------|-------|--------------------|-----------|-------|
| ω_s | .7 | $P(S_a) = .7$ | α | 2.0 |
| | | $P(S_i) = .3$ | β_1 | 1.0 |
| ω_o | .3 | $P(O_a) = .7$ | β_2 | 1.0 |
| | | $P(O_i) = .3$ | β_3 | -1.0 |
| λ | 3 | $P(S_l = 2) = .15$ | β_4 | -1.0 |
| | 3 | $P(O_l = 2) = .15$ | | |

This leaves us estimating the probability of the two orders as follows:

$$\begin{aligned}
 P(\text{SOV} | S_i, O_a, S_l = 2, O_l = 2) &= \frac{1}{1 + e^{-(2+2+0-2-1)}} \\
 &= .73
 \end{aligned}$$

$$\begin{aligned}
 P(\text{OSV} | O_i, S_a, O_l = 2, S_l = 2) &= 1 - \left(\frac{1}{1 + e^{-(2+2+1-2+1)}} \right) \\
 &= .05
 \end{aligned}$$

Together with the priors, we can plug these values into the formula for SOV/OSV preference. Because the prior over lengths is identical for subject and object, when both are observed the terms $P(S_l = \cdot)$ and $P(O_l = \cdot)$ will in fact always balance out and can be ignored. This leaves us calculating the following strong preference for an SOV interpretation, despite the animacy configuration.

$$\begin{aligned}
 P(\text{order} = \text{SOV} | \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2) &= \frac{.73P(S_i)P(O_a)}{.73P(S_i)P(O_i) + .05P(O_i)P(S_a)} \\
 &= \frac{.73 \cdot .3 \cdot .3}{.73 \cdot .3 \cdot .3 + .05 \cdot .7 \cdot .7} \\
 &= .73
 \end{aligned}$$

Given equal length phrases and the animacy order inanimate-animate, the model favors an SOV order. This is due to the high frequency of SOV, encoded in the α parameter. In sentences like those in the experiments here, this is the wrong

prediction, and therefore difficulty is predicted at the verb (which is semantically incompatible with the SOV parse). Compare this with the same calculations for an 8 word NP1 and a 2 word NP2.

$$\begin{aligned} P(\text{SOV}|S_i, O_a, S_l = 8, O_l = 2) &= \frac{1}{1 + e^{-(2+8+0-2-1)}} \\ &= 1.0 \end{aligned}$$

$$\begin{aligned} P(\text{OSV}|O_i, S_a, O_l = 8, S_l = 2) &= 1 - \left(\frac{1}{1 + e^{-(2+2+1-8-0)}} \right) \\ &= .95 \end{aligned}$$

$$\begin{aligned} &P(\text{order} = \text{SOV} | \text{NP1}_i, \text{NP2}_a, \text{NP1}_l = 2, \text{NP2}_l = 2) \\ &= \frac{P(\text{SOV}|S_i, O_a, S_l = 8, O_l = 2)P(S_i)P(O_a)}{P(\text{SOV}|S_i, O_a, S_l = 8, O_l = 2)P(S_i)P(O_a) + P(\text{OSV}|O_i, S_a, O_l = 8, S_l = 2)P(O_i)P(S_a)} \\ &= \frac{1.0P(S_i)P(O_a)}{1.0P(S_i)P(O_a) + P(\text{OSV}|O_i, S_a, O_l = 8, S_l = 2)P(O_i)P(S_a)} \\ &= \frac{1.0 \cdot .3 \cdot .3}{1.0 \cdot .3 \cdot .3 + .95 \cdot .7 \cdot .7} \\ &= .16 \end{aligned}$$

When the first NP is 2 words long, the ratio of $P(\text{SOV}|S_l = 2)/P(\text{OSV}|O_l = 2)$ is .5:.02 or 1:.04, reflecting the higher likelihood of SOV with short arguments. The resulting probability of the intended order being SOV is .73. However, when the length increases to 8 words, this probability drops to .16. This is the “explaining away” effect: whether the subject or object is long, the logistic regression will almost certainly predict that a speaker would place the long constituent first. Therefore the ratio of $P(\text{SOV}|S_l = 8)/P(\text{OSV}|O_l = 8)$ drops to 1:.95, almost 1:1; that is, order no longer provides much information about likely grammatical function assignments. Only animacy information, encoded in prior probabilities, remains available to guide the comprehender’s interpretation.

This explaining away property is a direct consequence of rational Bayesian inference. If processing complexity is correlated with the comprehender’s certainty about the parse given rational use of all available information, as suggested by the previous experiments, it follows that we should see explaining away effects in behavioral measures. I will now present an experiment designed to test this novel prediction.

5.4.1 Study 3: Length as an information source in comprehension

To test whether the explaining away prediction of the Bayesian “comprehension as inference about production” model holds, I ran an additional experiment in Japanese. This study tested whether the previously reported long-before-short preference would explain away an OSV word order, and therefore reduce any comprehension difficulty associated with it. By manipulating the length of the initial constituent, I created sentences which either conformed to the long-before-short order, or had no length asymmetry. The prediction for this experiment was that just like Kaiser and Trueswell (2004), comprehension difficulty associated with OSV orders would be reduced when that order was the natural order (here due to length rather than discourse structure).

This hypothesis might be understood intuitively in the following way. When a short initial object is encountered, the comprehender implicitly asks the question “why was this phrase placed first?” and after considering the factors that predict word order in production, will (mistakenly) assign a high probability to it being the subject. However, when the initial object is long, the most likely answer to that question is given instead by the long-before-short preference in Japanese. If length is taken to be a likely sufficient explanation for the word order, then there is no need to posit other causes like grammatical function to explain it, so the comprehender does not place a high probability on that phrase being the subject. Hence length should modulate the effect of word order on comprehension

Design and materials

I designed a set of 24 sentences similar to those from studies 1 and 2. The subject and object NPs were short lexical NPs (not names), and all sentences had *wo*-marked objects. Additionally, the adverb phrase from study 2 was moved to provide a buffer region between the second NP and the verb. To make sure this buffer would “soak up” any spillover from the previous words, a relatively long multiword adverbial phrase was used. To manipulate the order, the first two NPs were switched as before. To manipulate case particle presence, I simply deleted the *ga* and *wo* particles on those two NPs in the particles absent condition (no “missing” symbol was used this time). The length manipulation was achieved by adding a relative clause or other modifier of the first NP. As it was difficult to come up with modifiers which would sound equally natural for either the animate agent or the inanimate patient, I wrote a different version for each. The two modifiers in each item were designed to be close to the same length. The four particles-present conditions of one item are given below.

- (43) a. kaineko ga | hanabin wo | nitiyoubi no mayonaka ni |
 pet.cat-SUBJ | vase-OBJ | sunday-OF late.night-DAT |
 hikkurikaesita to | yamamoto san ga | syabetta
 knocked.over-THAT | Yamamoto-SUBJ | spoke
 *“Yamamoto said that the pet cat knocked over the vase in the middle of the
 night on Sunday”*
- b. hanabin wo | kaineko ga | nitiyoubi no mayonaka ni |
 vase-OBJ | pet.catsUBJ | sunday-OF late.night-DAT |
 hikkurikaesita to | yamamoto san ga | syabetta
 knocked.over-THAT | Yamamoto-SUBJ | spoke
 *“Yamamoto said that the pet cat knocked over the vase in the middle of the
 night on Sunday”*
- c. kinu no you ni tuyayakana ke-NO kaineko ga |
 silk-OF like glossy hair-OF pet.cat-SUBJ |
 hanabin wo | nitiyoubi no mayonaka ni |
 vase-OBJ | sunday-OF late.night-DAT |

hikkurikaesita to | yamamoto san ga | syabetta
 knocked.over-THAT | Yamamoto-SUBJ | spoke
“Yamamoto said that the pet cat with the silky smooth hair knocked over the vase in the middle of the night on Sunday”

- d. bara no ippai haitteiru hanabin wo |
 rose-OF full entered vase-OBJ |
 kaineke ga | nitiyoubi no mayonaka ni |
 pet.cat-SUBJ | sunday-OF late.night-DAT |
 hikkurikaesita to | yamamoto san ga | syabetta
 knocked.over-THAT | Yamamoto-SUBJ | spoke
“Yamamoto said that the pet cat knocked over the vase that was filled with roses in the middle of the night on Sunday”

Methods

While the basic self-paced reading paradigm was identical to studies 1 and 2, for study 3, data was collected over the web instead of in person in the lab. Web-based data collection for psycholinguistics and empirical linguistics is being increasingly widely used, and in many reported cases yields findings that are closely comparable to data from participants or annotators in the lab (see Snow et al., 2008; Keller et al., 2009; Munro et al., 2010). The experiment was run using custom-written software produced in Adobe Flash, which participants could view using the web browser on their own computers. Participants were recruited by word of mouth and were unpaid. Because participants were unpaid and were in their own homes I was worried that they might not be willing to participate in a long experiment, so I reduced the length by having the program select a random subset of 8 experimental items and 16 fillers for each participant. One experimental item was assigned to each of the 8 conditions. The 24 sentences seen by participants were broken into 4 blocks of 6. Each block contained 2 experimental items and 4 fillers. Presentation order within each block was randomized, with the constraint that no two experimental items were ever displayed adjacently.

Additionally, a new secondary recall task was inserted into this experiment. After

each block, subjects completed two recall trials in which the end of the main clause (e.g. “Yamamoto said that ...”) was presented together with three phrases taken from the sentence presented in a random order as “recall cues”. For experimental items, these three phrases were always the subject, object, and verb, meaning only the adverb was not displayed. For fillers, the recall cues were randomly selected. Participants were instructed to type in the cued sentence they had previously seen as accurately as possible. The two recall trials were always one filler and one experimental item selected randomly from the immediately previous block, but excluding the most recently seen sentence.

Participants

At the time of analysis, 69 participants had taken part.

Results

The data were analyzed as in studies 1 and 2. 11.7% of the data were excluded because the following comprehension question was answered incorrectly, and 1.3% because reading times exceeded 3 standard deviations from the per-region per-condition mean. Per condition means at each region are plotted in figures 5.7 and 5.8.

At the immediately preverbal adverb phrase, there were significant effects of particle absence ($\beta = 346, p < .001$) and a marginal effect of word order ($\beta = 230, p = .0706$). There was no significant effect of length ($\beta = -29.4, p = .783$) but this interacted with word order ($\beta = -360, p = .0344$) such that OS conditions were slower than SO conditions with a short initial phrase, but SO conditions were slower than OS conditions with a long initial phrase.

At the verb, there were no significant main effects of order ($\beta = 56.4, p = .548$), particle absence ($\beta = 40.5, p = .669$) or length ($\beta = 26.0, p = .832$) but an interaction between particle absence and word order ($\beta = 263, p = .0270$). There were no significant interactions between particle absence and length ($\beta = 66.4, p = .572$) or length and word order ($\beta = -103, p = .457$). There was a significant three-way interaction ($\beta = -325, p = .0460$) such that the particle/order interaction was

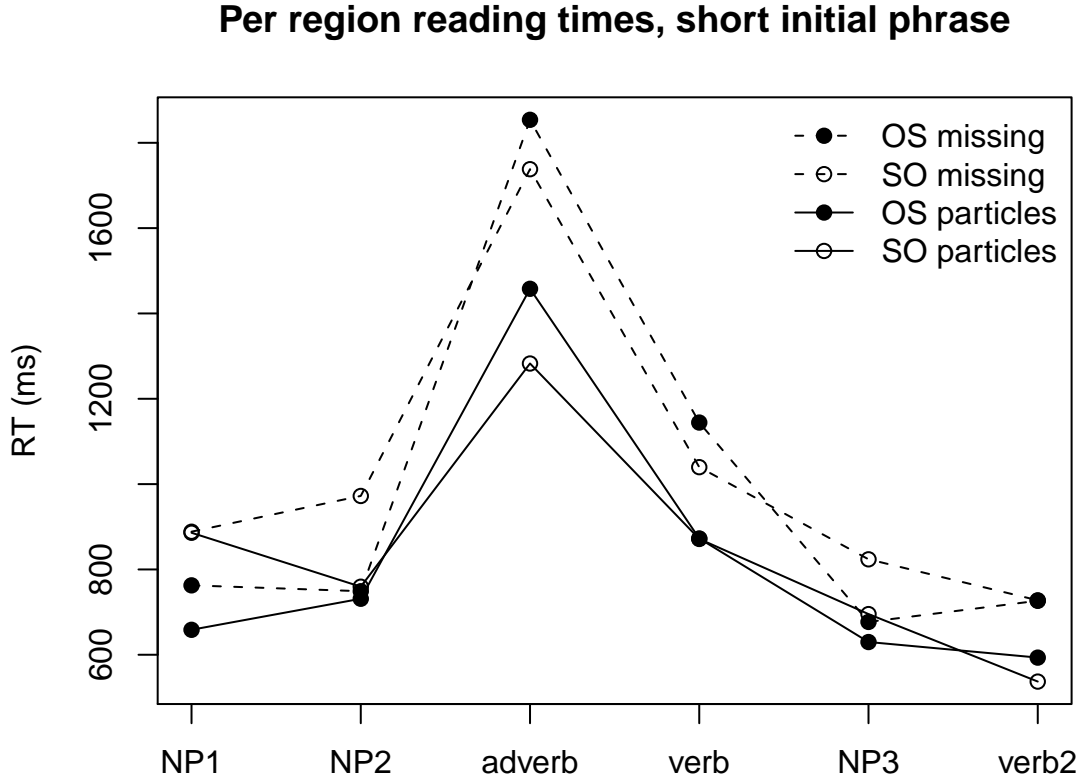


Figure 5.7: Reading times at each region in study 3 for short-short conditions.

observed only in the short conditions. The condition means at the two regions of interest are displayed in figure 5.9.

At the initial NP, there was a significant effect of length ($\beta = 2020, p < .001$). At the second NP, there was no significant main effect of particle absence ($\beta = 43.9, p = .545$) but an effect of length ($\beta = 271, p < .001$) which marginally interacted with particle absence ($\beta = -172, p = .0686$). At the main clause subject, there was a marginal effect of particle absence ($\beta = 47.0, p = .0650$). At the final verb there was a significant effect of particle absence ($\beta = 63.0, p = .0241$). All other effects estimates were not significantly different from 0 ($ps > .1$).

The recall task data was analyzed after removing all responses which were not genuine attempts at recall, or which did not contain both the correct subject and object (15.8%). The remainder was coded only for the relative position of the subject and object, ignoring the position of the adverb. The distribution of “correct” recalls

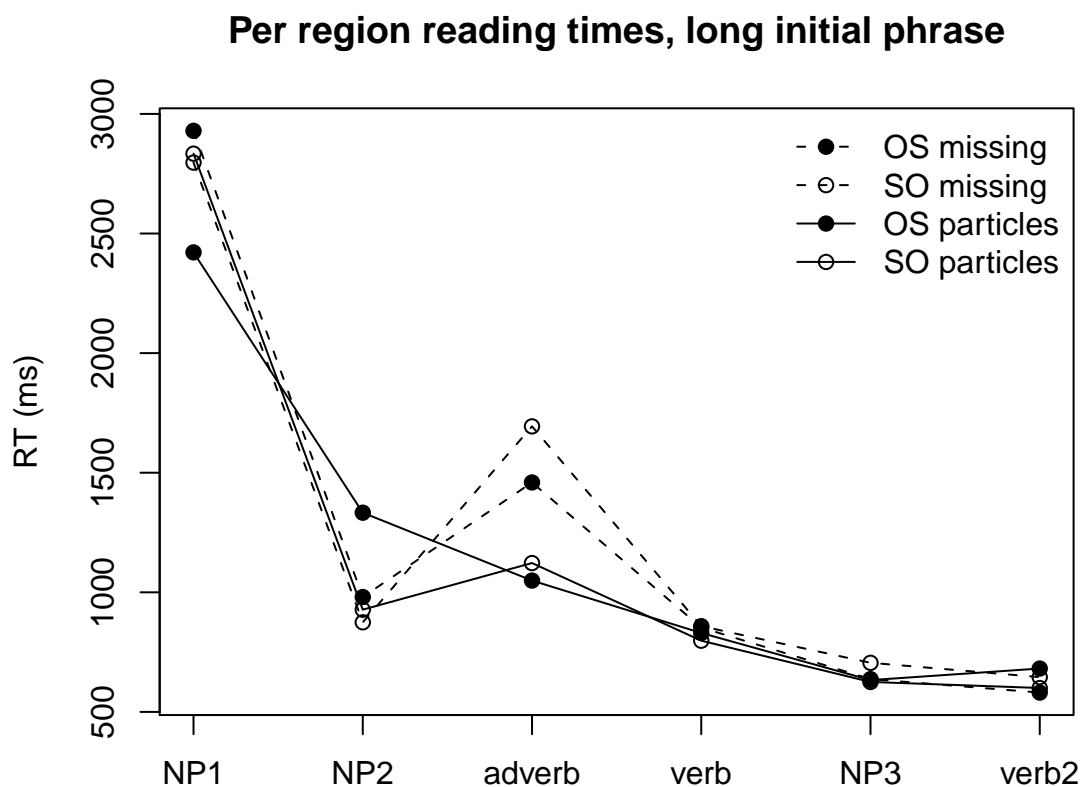


Figure 5.8: Reading times at each region in study 3 for long-short conditions.

(that is, those in which the recalled word order matched the actually presented word order) is as follows:

| Short initial phrase | | | | Long initial phrase | | | |
|----------------------|-----|-----------|--------|---------------------|-----|-----------|--------|
| | | Particles | | | | Particles | |
| | | Present | Absent | | | Present | Absent |
| Order | SOV | 94% | 100% | Order | SOV | 100% | 100% |
| | OSV | 28% | 42% | | OSV | 78% | 86% |

I analyzed these data using a multilevel logistic regression with crossed random intercepts for subject and item. As there were no incorrect responses in three out of the four SOV conditions, only the OSV data were analyzed. The only significant effect was of length, with long initial phrases leading to a higher proportion of correct OSV recalls ($\beta = 1.93, p < .001$).

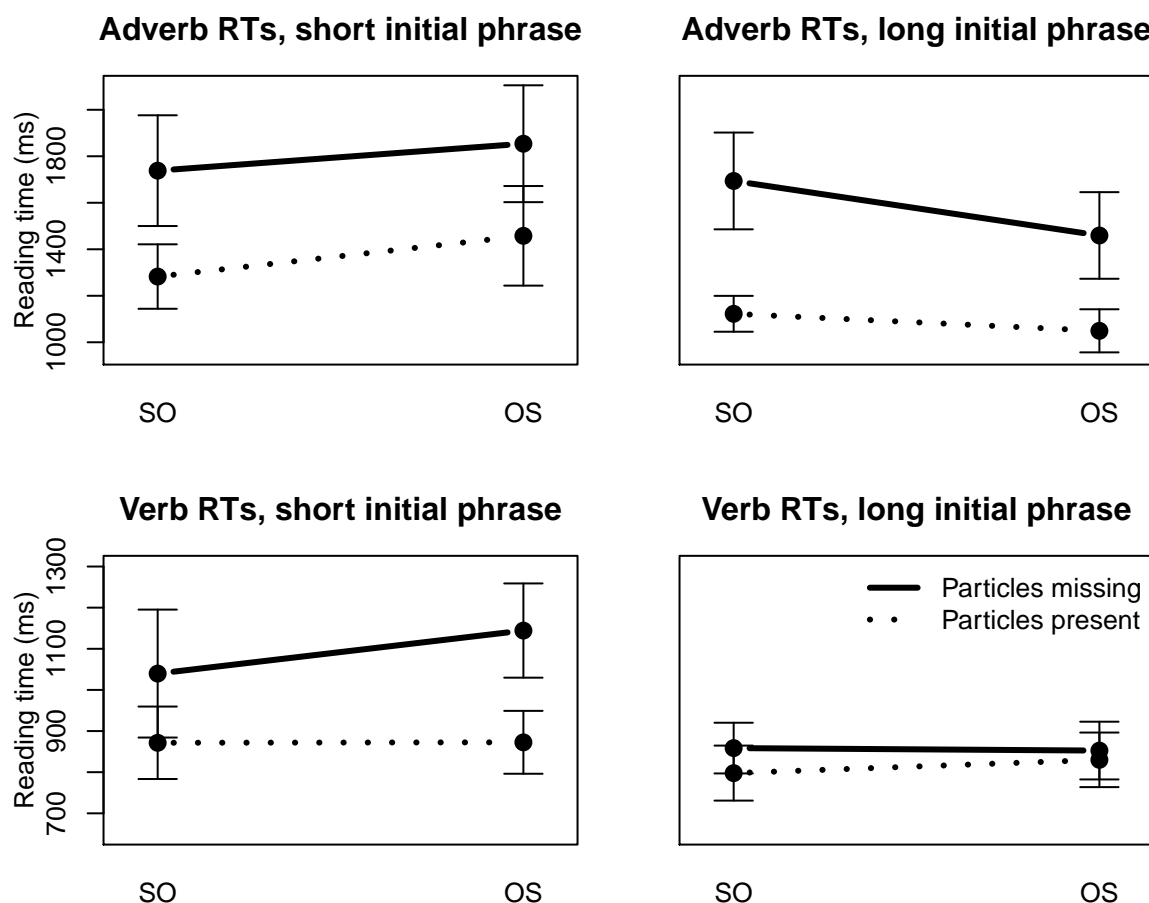


Figure 5.9: Reading times at adverb and embedded verb in study 3. Bars show standard error of the mean.

Discussion

Again, the short initial phrase conditions are directly comparable to study 1 and the *wo*-marked conditions of study 2. As expected, the pattern of reading times at the verb is very similar to those studies. This study differs in that the presence of the adverb after the subject and object allows the possibility of teasing apart effects obtained at the verb specifically from effects due to parsing decisions made after the NPs have been read, from simple spillover from the previous NPs. Immediately following the second NP, there is an effect of particle presence such that NPs without particles lead to slower reading times. However, there is only a marginal effect of word order, and no sign of any interaction between particle presence and word order.

This expected interaction only surfaces at the verb, as would be expected if this interaction effect is due to a parsing decision made at the verb specifically, once the semantic contribution of the verb rules out an incorrect assignment of subject and object properties to the two NPS.

We can compare these results to those obtained when the initial phrase is long (i.e., when the order of the two arguments is long before short). The pattern of results at the adverb is the same as in the short phrase conditions: there is an effect of particle absence, but not of word order and no interaction. Again, this suggests that the locus of the effect of interest is the verb specifically, rather than just the first point after the noun phrases. However, when the first phrase is long, no effects of word order surface at the verb: all conditions are read at roughly the same speed as the particle present conditions. This finding can be accounted for if the comprehender is using their knowledge of the production system to explain away the OSV word order as a result of the length asymmetry, making it lose influence as a source of information about grammatical function. Crucially, the fact that word order is explained away by multiple possible causes is only predicted by models where the nonindependence of information sources is explicitly modeled. This pattern of results is similar to those reported by Kaiser and Trueswell (2004), where Finnish variant word orders incurred no difficulty when predicted by previous context. Both of these sets of findings can be interpreted as evidence that online comprehension involves high-level reasoning about the relationships between probable causes of the speaker's choice of linguistic form.

The one unexpected result in the reading times is the interaction between word order and length at the adverb. This is mostly carried by the fact that with a long initial NP, RTs are significantly faster for OSV than for SOV sentences. Exactly why this should be is not clear: if an adverb following OS was read slower than following SO with short NPs, but *equally* fast with a long initial NP, it would suggest that the OSV order causes some difficulty (uncertainty) early on even with case marking, and the length differential eliminates that uncertainty. However, the OSV order actually appears to be read *faster* here, which is surprising given that SOV order with a long subject is even more expected than OSV with a long object.

Turning to the sentence recall results, subjects recalled SO orders correctly almost 100% of the time. OS orders were mistakenly recalled as SO significantly more often when both subject and object were short than when the object was long. This is evidence for the long-before-short preference. In fact, it is a direct replication of the constrained production task of Yamashita and Chang (2001) using the recall task of Bock and Warren (1985) (also Branigan and Feleki, 1999; Tanaka et al., 2005) described in Chapter 3. As in those previous studies, this recall task can be interpreted as a kind of constrained production task.

In a previous pilot for this experiment conducted in the lab, there was also a significant effect such that the OSV sentences *without* particles were recalled better than those with particles. If this is the case, it would be evidence that comprehenders pay greater attention to word order cues in sentences when a more reliable source of information (case particles) is not available. In this data, there is a trend of a similar size in that direction, but it does not reach significance. It is possible that this is simply because the small number of recall trials each subject completed here (just 4 experimental items) is too little data to get an estimate of effect size which is not confounded by between-subject differences. In the pilot, subjects completed 12 such trials each. I intend to follow up on this question in future work.

5.5 Discussion

Although a great deal of empirical work in psycholinguistics literature has concentrated on English and closely related languages, more recent years have seen a renewal of interest in sentence comprehension and production in other languages (see Cutler, 1997; Costa et al., 2007; Jaeger and Norcliffe, 2009). For the program spelled out in Hawkins' Performance-Grammar Correspondence Hypothesis to be tenable, cross-linguistic support for theories of processing is extremely important. To support the notion that certain psycholinguistic preferences are cognitive universals and therefore shape typological patterns, we need to be able to test that any putative cognitive universal is not in fact specific to whichever language or language type it was documented in (see Cutler, 1997; Hawkins, 2007; also Evans and Levinson, 2009 for descriptions

of the substantial differences between languages more generally). The results presented here bolster previous findings that the “heaviest last” preference observed in head-initial languages does not generalize to head-final languages: in fact, the reverse preference is found.

In this chapter, I discussed the fact that languages differ greatly in the ways in which they encode meaning, and evidence that the strategies by which language users process language differ accordingly. One way to account for this diversity is to assume a comprehender who rationally and simultaneously takes advantage of multiple sources of information, but weights them according to their utility in determining the intended meaning. I presented a simple Bayesian model of comprehension as inference about the speaker’s choices in production. Under this model, we expect to see certain sources of information stepping in to drive interpretation automatically when other sources are no longer available or useful. In a series of experiments on Japanese, I demonstrated that despite being highly predictive of grammatical function, word order is not used much by comprehenders in the normal situation that case markers fully determine the interpretation. However, when case is unavailable, comprehenders fall back on word order. In a final experiment, I then tested a prediction that falls out of the Bayesian model: when there are multiple possible factors might influence a speaker to make a certain production choice, the phenomenon of “explaining away” can lead the comprehender to drastically reweight the information sources available to them.

The fact that comprehenders appear to be able to make high-level inferences about what drives speaker’s choices in production is a novel finding, and is not predicted by simple additive cue combination models. Additionally, it means that comprehenders should be able to infer and adjust for processing-driven effects such as short-before-long or long-before-short preferences in production. This is evidence against the simple language change model of chapter 4, in which a failure to account for processing biases was responsible for driving language-wide word order change.

Chapter 6

Discussion and extensions

Here I survey the major findings of the previous chapters, and discuss their implications. I also discuss some extensions to this work which are currently being planned or underway.

6.1 Dependency minimization in English

In chapters 2 and 3, I presented corpus results based on data from Old and Middle English, stressing two principal findings. First, in main clauses in Old and Middle English in which the subject preceded the object, VO order was chosen over OV more often in proportion to the length of the object. Second, changes in the grammar of Old and Middle English led to words that were dependent on each other for interpretation being placed increasingly close together as time progressed. Both of these results can be understood in the context of a family of psycholinguistic theories which I gave the umbrella term *dependency minimization*: the online production or comprehension of a word incurs effort or difficulty in proportion to the distance that has passed since any previous words on which it is dependent (Hawkins, 1994; Gibson, 1998, 2000; Hawkins, 2004).

The direction of word order change in English (towards VO) leads to a more efficiently processed language under the dependency minimization theory. In fact, changes in English taken as a whole led to a more efficiently processed language,

since average dependency lengths in sentences of a similar length grew shorter over time. From these findings, it is tempting to conclude that a pressure towards orders that minimized dependency lengths had a causal role in the change of English word order. This conclusion would be overstated, of course, since there could be many other factors that are correlated with these dependency lengths, and it could be any of these other factors which is responsible for change rather than dependency minimization itself. For instance, information structure correlates with phrase length (given and topic phrases are more likely to be referred to with a pronoun or short noun phrase) and with word order (crosslinguistically, given phrases tend to be mentioned first). Modern psycholinguistic research described in chapter 3 has shown that in head-final languages like Japanese and Korean, information status and phrase length can be dissociated cleanly: given phrases are placed earlier, as in English, but longer phrases are *also* placed earlier, rather than later (Choi, 1999; Yamashita, 2002; Hawkins, 2004). It may well be possible to dissociate the effects of information status and length in historical English, too. Unfortunately, the data that is currently available is not sufficiently annotated, so as yet it has not been possible to do this systematically.

A further potential issue with the dependency minimization theory as a driving factor in change is that it predicts a universal bias towards SVO and away from SOV, despite the fact that SOV order is roughly equally frequent as SVO across the world's languages. In chapter 3, I considered the hypothesis that other factors, particularly the presence of case, might reduce or eliminate the processing complexity associated with longer dependency lengths. Certainly, German (which has rich case marking) shows less evidence of minimal dependency lengths overall (Park and Levy, 2009; Gildea and Temperley, 2010) and there is psycholinguistic evidence that German speakers do not always suffer from the comprehension difficulty associated with long dependency lengths in English (Konieczny, 2000). In chapter 4, I also discussed some evidence that speakers may use alternative strategies for expression — in particular, favoring intransitive clauses over transitives — of expression to avoid long dependency lengths in SOV languages (Ueno and Polinsky, 2009). In order to evaluate whether differences between languages modulate the effect of dependency minimization on word order change, further work will need to investigate the average

dependency lengths (and particularly, any changes in those averages) in empirical data from languages that differ in their case systems and basic word order.

6.2 Should processing effects be treated as priors or biases?

In chapter 4, I presented a series of simulations of word order change based on the empirically observed change in Old and Middle English. I showed that a population of agents whose production was biased towards placing longer objects after the verb would converge on an all-VO language. This bias was implemented as a mismatch between the knowledge of language the learners acquire from their input, and the sentences they actually produce. Agents used the linguistic input they receive to learn a grammar determining when to use each order, based on other properties of the utterance. However, in production they experienced an additional fixed bias towards using VO order with longer objects, which could overrule the grammar. This bias was intended to capture the processing bias to avoid word orders with long dependency lengths. Since the agents did not infer that some VO sentences might be VO merely because of this processing bias, each generation estimated the background rate of VO to be higher than the last, leading to population-wide change.

However, the results of the psycholinguistic experiments in chapter 5 suggest that in fact, human language users probably would be able to infer when a certain word order is used because of a processing bias rather than some other cause. In that case, agents that are unable to correct for processing biases may be unrealistic, so in future work I intend to explore further simulations where the agents are able to infer whether there is a processing-based cause to an observed word order. One possible way forward is to explore models where the processing effect is not implemented as a hard-and-fast bias, but a “soft” bias, which favors processing-optimal orders, but does not prevent agents from learning any conceivable grammar. This could be done by specifying a *prior* over learning: essentially, making learners regard processing-optimal grammars as more likely and therefore choose them preferentially.

The simulations in chapter 4 were *iterated learning* models, following those of Kirby (2001). Iterated learning refers to a learn-use cycle in which agents first learn from the language of the previous generation, and then generate the language that the next will learn from. Griffiths, Kalish and colleagues (Griffiths and Kalish, 2005, 2007; Griffiths et al., 2008) have explored language evolution in a population of *Bayesian* learners. A learner is Bayesian if it does not simply select the grammar that is most likely to have generated the language it receives as input (a maximum likelihood learner) but combines evidence from their input with some *prior expectations*. With Bayesian learners, Griffiths and Kalish show that iterated learning results in convergence to the learners’ prior. This means that in repeated simulations of language change through iterated learning, the distribution over agents’ grammars will always eventually come to look like the distribution defined by the the agents’ expectations of which grammars are likely. The mathematical details by which this occurs are this is explained in Griffiths and Kalish (2007), but the underlying mechanism is in fact extremely intuitive. Essentially, each agent learns a grammar according to a probability distribution over possible grammars which can be factored into two parts: $P(l|g)$, the probability of the observed language l given the hypothesized grammar g , and $P(g)$, the prior probability of that grammar. This product is then normalized to give a value from a proper distribution, the *maximum a posteriori* (MAP) probability distribution. The prior represents the learner’s expectations about which grammars are possible or likely, and may be very noninformative (say, assigning equal probability to all and only the class of context-free grammars) or may strongly guide the learner between possible grammars (say, by assigning higher probability to grammars with fewer rules or consistent branching direction). In the first generation, learners will be influenced by the initial input language, which is not affected by the prior distribution. However, the language they then produce as input to the second generation will reflect both that initial input and the prior. The second generation, in turn, will combine evidence from that language with the prior again. Each generation in turn will be influenced equally by the prior, but the initial input in its pure form will only influence the first generation. This means that the influence of the initial input will be gradually “watered down” by the repeated influence of the prior at every step.

This effect can in fact be seen in the “unbiased” simulations reported in chapter 4. In those simulations, agents’ grammars were logistic regression models estimated from the input data. The prior in that case was noninformative, being a very weak prior on each coefficient, centered at zero. This means that agents did not consider any potential grammar (i.e., any particular setting of coefficients in the logistic regression) more probable than any other, except to avoid values that are extremely far from zero. Over sufficient time, we would expect to see the distribution over agents’ grammars converging on this noninformative prior: all values of the coefficients within some distance of zero would occur roughly as frequently as each other. Within the number of iterations reported in chapter 4, the influence of the initial language is still evident for the coefficients with stronger effects on the outcome: the intercept and object pronominality. However, for the other coefficients, which are not as strongly associated with the outcome, we see gradual random drift. Over (a potentially extremely long) time, we would expect to see that this drift had led to almost the entire space of possible values of each of the coefficients being explored equally. The distribution over those values would then have converged to the noninformative prior over them.

The biased simulations in chapter 4 cannot be interpreted in the same way, because there the agents do not generate from the same grammar they learn: agents acquire a logistic regression that does not include a coefficient for object length, but they produce sentences from a model with a fixed object length coefficient. This means that learners are not really estimating $P(l|g)$, since they are unable to estimate the length effect, and so are not really Bayesian learners. An alternative to these biased agents would be to use fully Bayesian learners like those of Griffiths and Kalish, and implement the processing preference for longer objects to appear after the verb by introducing a prior bias on the space of possible grammars. That is, learners would consider all possible settings of their coefficients, but their prior $P(g)$ would weight grammars with a positive object length coefficient more highly. This can be implemented quite simply with a gaussian prior centered at some reasonable value for the coefficient. I will briefly report a simulation that does exactly this.

The simulation was set up in the same way as the simulations of chapter 4: each

agent’s grammar is represented as a logistic regression predicting the word order outcome (VO vs OV) as a function of a set of four properties of the object: length, pronominality, whether it is direct or indirect, and whether it is negated. Again, the initial language input is taken from Old English corpus data for texts before 950. Since the fitting procedure used in the previous simulations does not allow arbitrary priors to be placed on the coefficients, I instead used WinBUGS (Spiegelhalter et al., 1999) to fit each learner’s grammar from its input. WinBUGS is a freely available piece of software for fitting probabilistic models using *Monte-Carlo* simulation methods; for details, see Gelman and Hill (2006). Using WinBUGS, I could implement agents which differ from the unbiased learners of the original simulation in one crucial way: they prefer to set the coefficient for object length close to the value .92 (the value obtained in the empirical corpus study in chapter 2). Because these learners combine evidence from the input with priors to yield a distribution over the grammars they are likely to acquire, they can be called MAP learners.

The prior was implemented as a gaussian distribution with mean .92 and standard deviation 4, meaning that values quite far from .92 were still given quite high probability (see the “weak prior” plot in Figure 6.1). Therefore this simulation differs from the previous simulations in that it allows learners to notice and correct for processing biases; learners could even acquire a grammar with “reverse” weight effects, preferring to place longer objects before the verb, if there was enough evidence for that in the language being learned from. However, learners expect that grammars which conform to the observed processing bias are more likely; therefore, where evidence in the input does is not enough to determine the correct relationship between length and word order, they will tend to prefer grammars with the relationship between length and order predicted by the dependency minimization theory.

One further difference from the previous simulation setup is that here, I only test “chains” of agents: each generation contains only one agent, which learns from the previous and provides the input to the next. This was done to save time, as fitting models using WinBUGS is orders of magnitude slower than doing so using the R packages used previously. The results from two independent runs of the simulation are displayed in Figures 6.2 and 6.3.

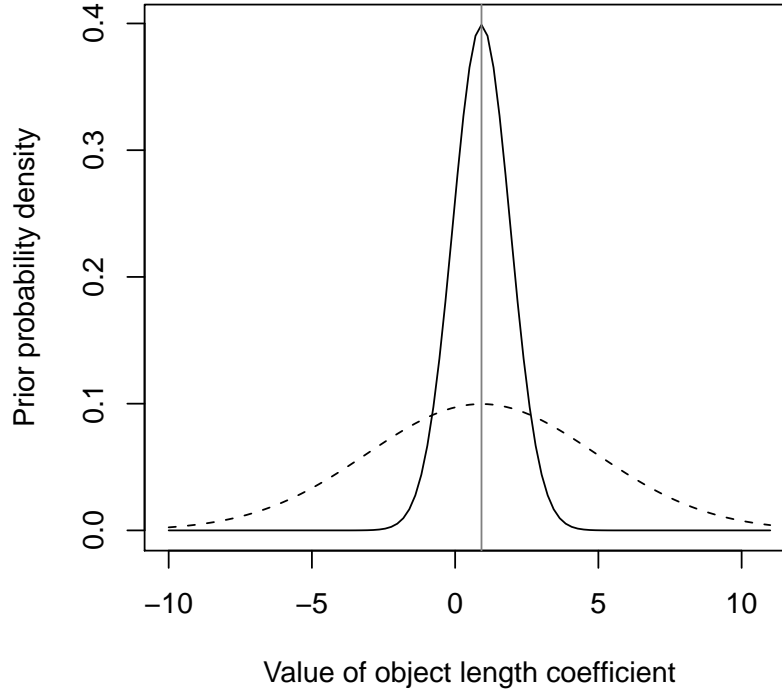


Figure 6.1: Prior probability over the object length coefficient for weak prior (dotted, $sd=4$) and strong prior (solid, $sd=1$) simulations

The results of this simulation are extremely similar to the unbiased simulation in chapter 4. There is more noise (that is, drift occurs faster) in this simulation than previously, probably because the results come from a single chain of agents rather than being averaged over a larger population. The intercept and pronominality coefficient, as before, remain fairly stable within the 60 iterations shown, indicating the strong effect of pronominality on word order in the initial language sample. Object length tends to stay positive in this simulation (these two runs are typical in that respect), reflecting the influence of the prior reinforcing the observed object length effect in the initial data. However, the crucial similarity between this simulation and the previous unbiased simulations is that the overall word order trends do not vary across time.

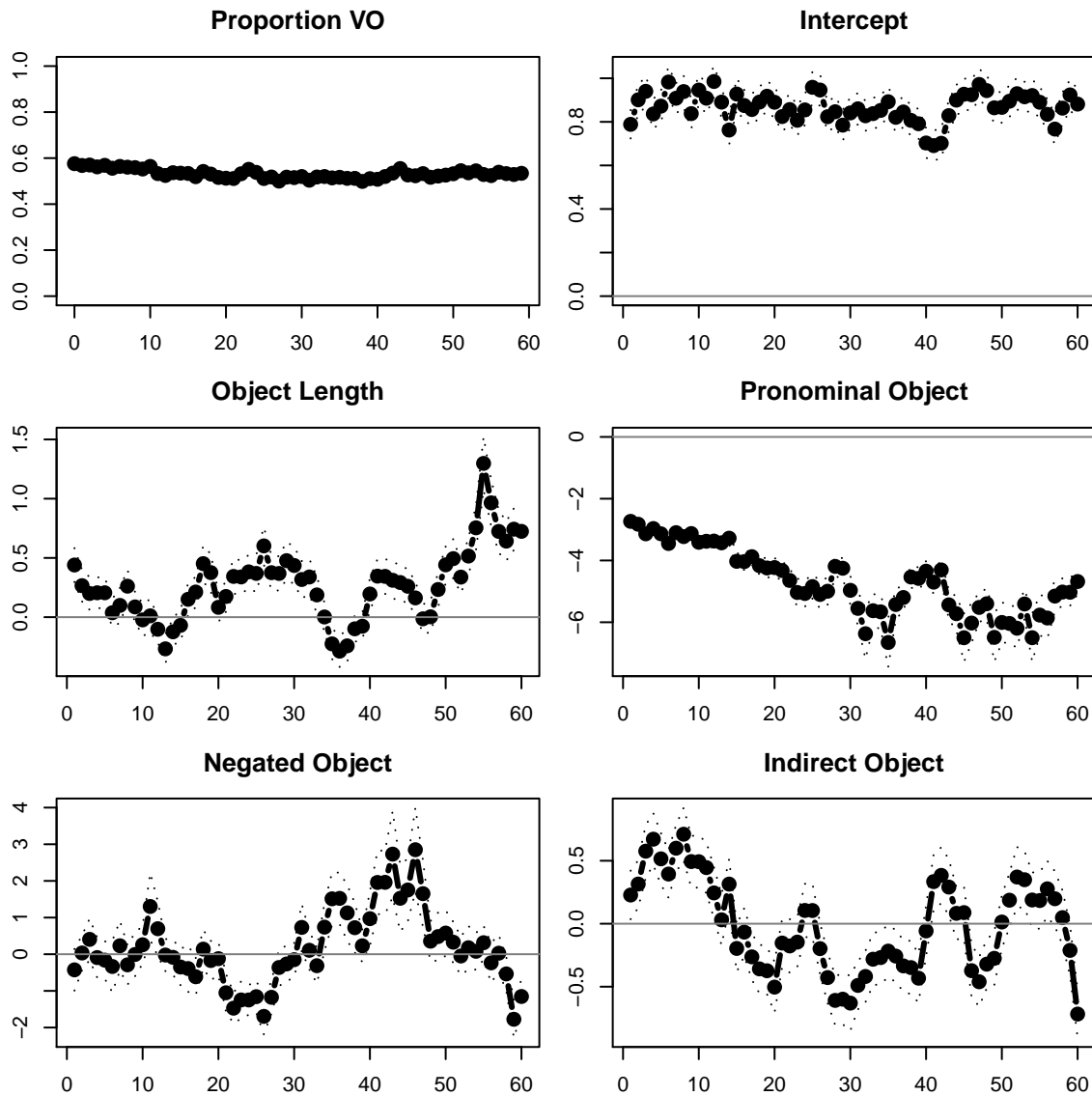


Figure 6.2: MAP simulation with weak object length prior, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates.

Even though the coefficients of the learners' grammars vary greatly as the simulation progresses, the overall proportion of VO orders produced in each iteration hardly differs at all. This is not at all due to the fact that the relatively weak prior over

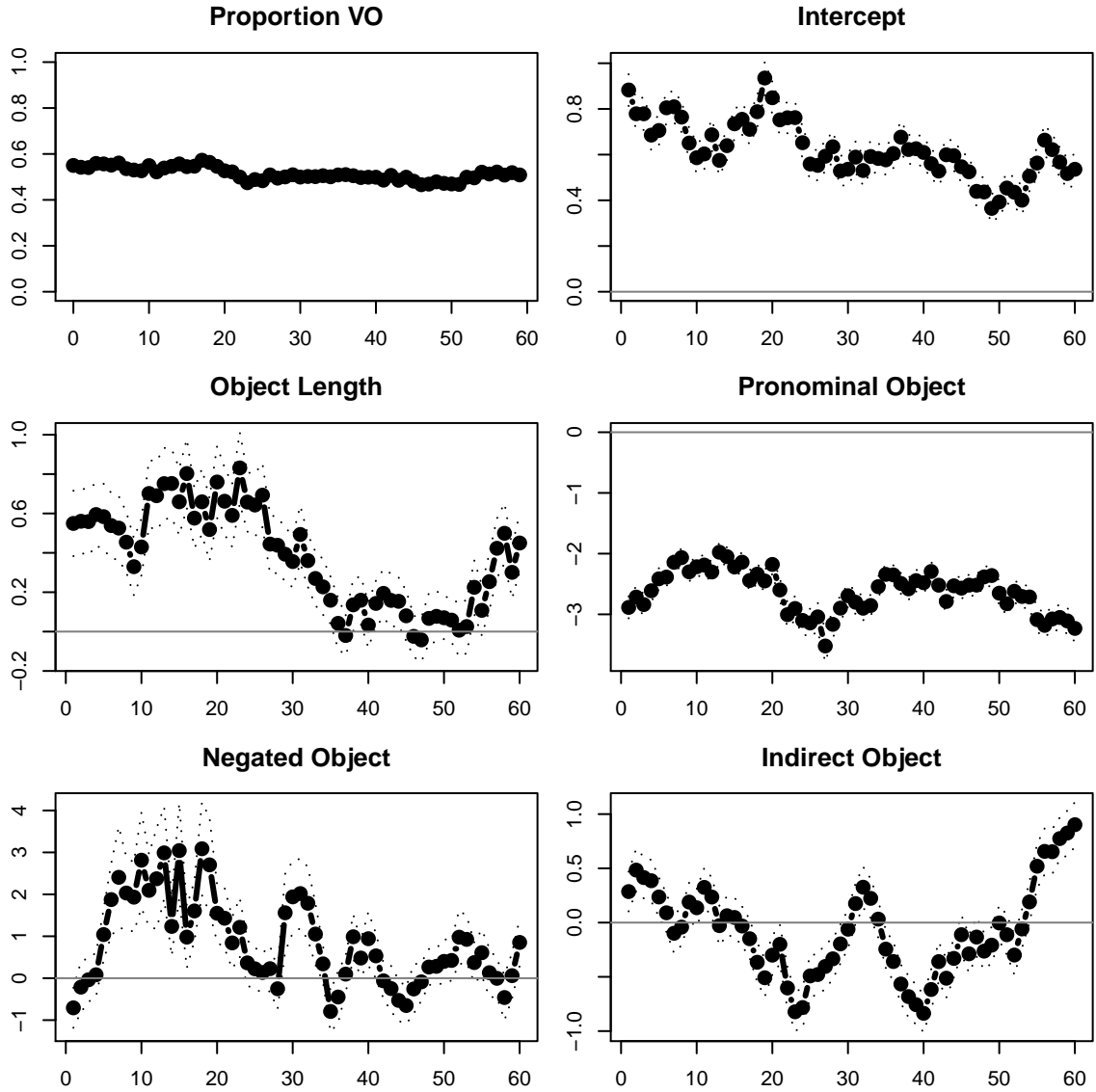


Figure 6.3: MAP simulation with weak object length prior, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates.

object lengths fails to bias learners enough towards acquiring grammars which prefer to place longer objects postverbally. To illustrate this, in Figures 6.4 and 6.5 I display the results from a simulation which is identical except that the prior over the object

length coefficient is now set to have mean .92 and a standard deviation of only 1. This “strong” prior is plotted in Figure 6.1.

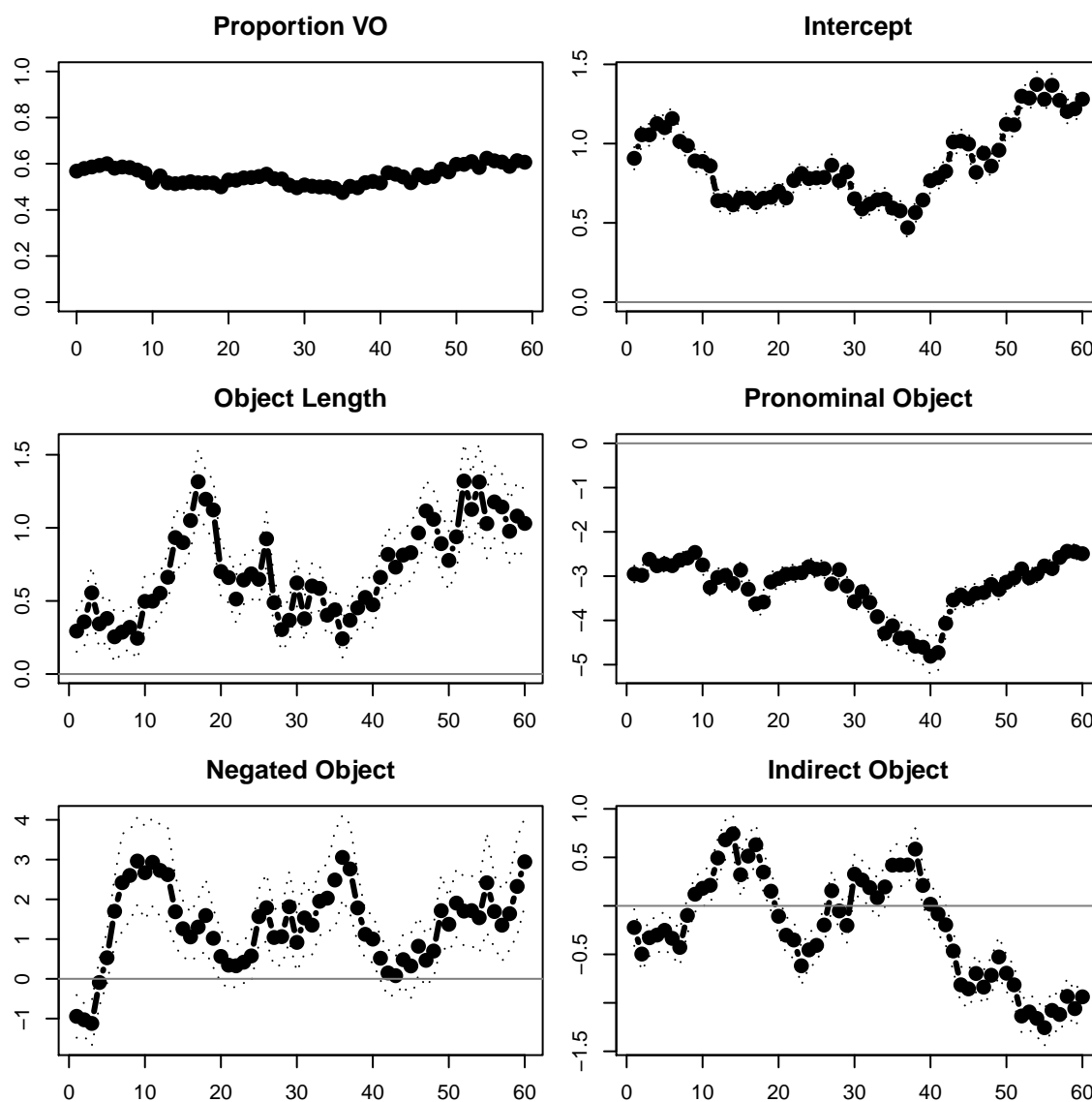


Figure 6.4: MAP simulation with strong object length prior, run 1: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates.

As can be seen in Figures 6.4 and 6.5, with a stronger prior bias on the object

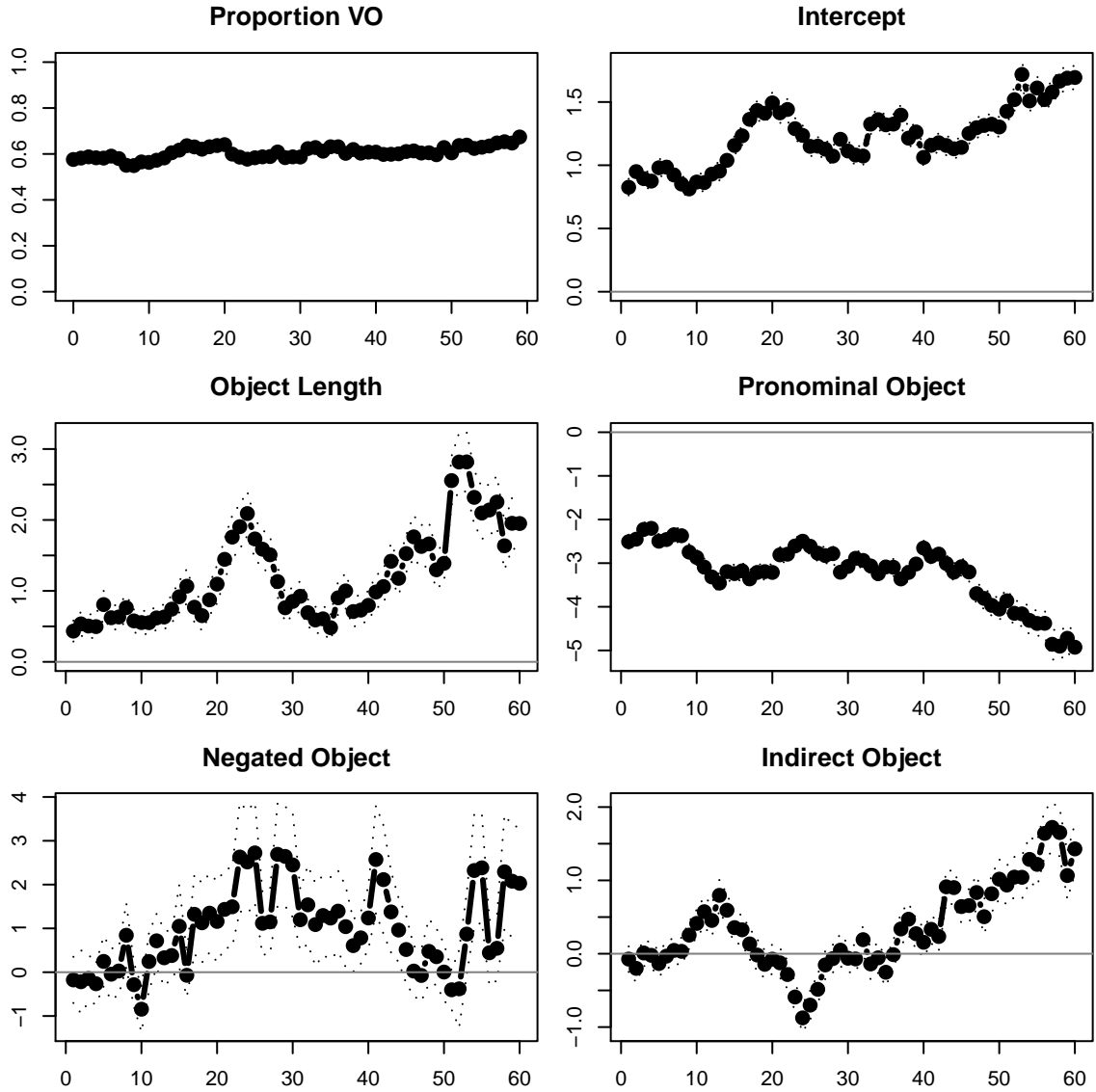


Figure 6.5: MAP simulation with strong object length prior, run 2: Proportion VO produced (top left) and change in coefficients (other plots) over time. Solid lines are point estimates. Dotted lines show 25th and 75th percentiles on the distribution of estimates.

length coefficient, agents are somewhat more likely to end up with a coefficient close to .92, meaning they more consistently prefer to place longer objects after the verb. This is in accordance with Griffiths and Kalish (2005), who predict that the distribution of

values for that coefficient should eventually converge on the prior distribution over it: in this case, the one displayed in Figure 6.1. However, the overall word order results are unchanged: the simulation preserves the proportion of VO in the initial output quite accurately.

These additional simulations show that merely having learners who favor grammars that are in line with the processing preferences I have argued to be universal is not enough to cause change. Rather, what causes the word order change in the biased simulations of chapter 4 is an explicit failure to account for those biases. When learners are able to correctly attribute word order variations to processing biases, they do not misattribute them to other causes — like a high background rate of the favored word order — and hence no change necessarily takes place.

Taken together with the results of chapter 5, which suggested that adult comprehenders are able to infer when hypothesized processing biases might have lead to the word orders they encounter, these simulations call for a more nuanced model of how processing biases might affect language change. One alternative would be to reject the applicability of the findings of the psycholinguistic experiments in chapter 5. It could be that findings from adult online comprehension do not extend to children, or do not extend to language acquisition in general.

An alternative — and perhaps more natural — explanation is that the simulation reported in this chapter has lost the underlying intuition behind the processing theories, which is that certain configurations of words are efficient for the speaker or hearer. In the biased simulations of chapter 4, the processing effect stems from a *bias* built into the system, which is intended to model a universal preference speakers have for processing-efficient orders. In the final simulation reported here, the processing effect is a *prior* that represents the learner’s expectations, reducing the status of the effect to something to be merely observed and accounted for, not a psychological fact that offers any inherent benefit or justification. If processing efficiency is enough to motivate speakers to choose a certain word order even though they are aware that the alternative order is otherwise more common or natural under their grammar, we would expect to see a bias towards efficient word orders even in language users who were at some level aware of that bias. This would lead to the change observed in

the biased simulation of chapter 4, where the order that is more common or natural under the grammar comes to align with the order that is more efficient under the processing bias.

In summary, while the simulations in this dissertation show that learners who do not account for production biases stemming from processing effects could lead to widespread word order change, learners who are aware of and correct for processing effects would not be expected to lead to the same outcome. The psycholinguistic studies reported in chapter 5 suggest that human language users are indeed aware of the influence of weight on order, an effect which I have argued is due to processing efficiency. The question of whether language users are able and motivated to “correct for” this effect on the frequency of word orders in alternation, however, is an open question for future research.

6.3 Language evolution and the relationship between processing and grammar

This dissertation has tried to describe some ways in which processing might influence language change within an evolutionary framework. As discussed in chapter 1, language change can be understood as an evolutionary process thanks to three key features. First, the forms that make up one speaker’s linguistic knowledge (words, constructions, and so on) are “replicators” in that they recreate themselves by being copied by other language users. Second, linguistic knowledge is subject to small changes due to being mislearned or deliberately extended or altered. Third, certain linguistic forms may be more or less “fit” than others, and — by hypothesis — this differential fitness may lead to them being replicated with higher or lower probability. I have argued that one component of the fitness of a linguistic form is the extent to which it minimizes the dependency lengths between words in a sentence. A large body of psycholinguistic work summarized in chapter 3 has argued that comprehension is facilitated when words which are dependent on each other for interpretation appear adjacent, or as close as possible to each other. Likewise, when speakers have a choice

between multiple possible orders of words, there is evidence that they prefer to place them in the order which minimizes the dependency lengths between them. If this processing pressure does turn out to play a role in how likely it is for constructions and word orders to be replicated by incoming generations of speakers, then the evolutionary framework predicts gradual language change such that orders with minimal dependency lengths become more frequent, perhaps even pushing out any alternative orders.

This account is fully compatible with the Performance-Grammar Correspondence Hypothesis of Hawkins (2004), the observation that processing preferences in some languages tend to be mirrored by hard grammatical constraints in others. In fact, this is an attempt to give a causal explanation by which Hawkins' noted correlation might arise. However, the explanation is still stated at a fairly abstract level, which does not fully specify the actual nature of the changing representations. Similarly to Hawkins, I have tried to make only minimal claims about the nature of the mental representations which compose human knowledge about language. In principle, typologically frequent forms could be very directly preferred in online processing, meaning that processing effects directly lead to many of the (statistical or absolute) constraints on possible languages observed by typologists. Alternatively, forms which are theorized to be more efficient could be preferred by only a very minimal amount in online processing, or even have been preferred only in some earlier stage of the language's development, or only during adult use but not at all during acquisition. In that case, the typological universals which are associated with processing pressures would in fact be mere fossilized preferences, which have become part of the grammar but no longer reflect psycholinguistic effects that could be observed to any substantial degree in online processing.

Bates and MacWhinney (1982, 1989) define several possible levels at which there could be a relationship between the *form* of language and functional pressures, such as for efficient processing. Bates and MacWhinney's *level 1* describes purely diachronic correlations, or fossilized pressures. This level might describe the position on Universal Grammar taken by Pinker and Bloom (1990). There, UG is understood as a collection of previously communicatively advantageous constraints on language

which have become fossilized in the human genome. Pinker and Bloom suggest that UG arises by the *Baldwin effect*. Many biological organisms have some flexibility to adapt to their environment during their lifespan. The Baldwin effect allows these species to encode traits that are commonly acquired in response to environmental adaption into their genotype, over generations. As an illustration, ostriches spontaneously develop calluses on their keels and sterna, which are adaptive since they protect those areas which come into contact with the ground when sitting. It has been theorized that these calluses were originally only developed after repeated abrasion during the lifespan of the ostrich, thanks to the ability to develop calluses as a protective measure in response to repeated environmental damage. However, since ostriches always developed the calluses in similar areas, individuals who happened to have a slight propensity to develop them quickly in those areas had a slight fitness advantage, and over generations of continuous selection, what was originally an environmental adaptation become encoded in the DNA (Waddington, 1942; discussed in Chater et al., 2009). Similarly, human ancestors may have had the ability to learn arbitrary communication systems from peers in their environment, but individuals with some predisposition towards learning a type of system that proved more useful would have a fitness advantage, and over generations more and more specific features of language would be pushed into the DNA. Pinker and Bloom’s UG, therefore, imposes constraints that may have conveyed some communicative benefit in prehistory, but do not necessarily need to have any directly observable functional advantage or foundation today.

Bates and MacWhinney’s *level 2* describes a relationship between form and function in which the evidence for the functional of linguistic forms can still be observed in adults today, but does not guide acquisition. It is possible that the apparent conflict between the evolutionary simulations of chapter 4 and the psycholinguistic results of chapter 5 could reflect a relationship at this level. If children learning language were unaware of or unaffected by the relationship between dependency lengths and processing difficulty, they might only look at the raw frequency of word orders in a language when acquiring the alternation between them. In this hypothesized situation, learners would not acquire the ability to condition their use of one order or the other on

phrase lengths until later in their development. Therefore, we would expect to see functional and processing biases like dependency minimization lead to language-wide change, but we might not be surprised to see adults being influenced by or even aware of those biases in online language use, as was demonstrated for the case of Japanese in chapter 5. *Level 3* describes the state in which functional relationships not only influence adult language use, but also acquisition. In this situation, functional relationships can help “crack the code” when learning a native language: for instance, processing biases like accessibility might actually be beneficial in allowing children to limit the set of likely referents for a newly encountered word, or understand the range of allowable participants for a grammatical function like subjecthood. Chater et al. (2009) criticize Pinker and Bloom’s account of UG as a product of the Baldwin effect, showing using simulations that when a fast-adapting system like language (which changes dramatically over the course of centuries, as shown in chapter 2) is placed in tandem to a slow-adapting one like the human genome (which requires millenia to adopt significant adaptations), only the faster system shows any sign of adapting to environmental pressure, such as the functional or processing pressures described here. If this is correct, then any innate language-specific cognitive abilities would not have been encoded into the DNA by the Baldwin effect. Rather, any functional adaptations would have been encoded into the *living languages themselves*, resulting in an explanation for the PGCH that resides at least at Bates and MacWhinney’s level 2.

The highest level Bates and MacWhinney discuss is their *level 4*, which describes the situation in which functional relationships essentially drive all of human language use, and any fixed grammatical patterns are epiphenomena of some active pressure for communicative efficiency. For this extreme position to hold, all apparently arbitrary features of natural languages would have to be given purely functional explanations. As Bates and MacWhinney state, it seems unlikely that all such distinctions can be given a functional motivation. However, in recent years a great deal of work has made strides in that direction: the specific example Bates and MacWhinney (1982) give is the arbitrary seeming assignment of German nouns to the three gender classes of the language. In fact, Futrell (2010) gives evidence that this assignment serves to

space out information across a larger number of words, reducing the online processing load required to deal with incoming information at each point, and allowing German speakers to use a richer set of nouns in context while not increasing the comprehender's uncertainty about upcoming nouns significantly. This said, it would be very premature (if not Panglossian) to claim that all such apparently arbitrary seeming distinctions are in fact adaptive. In fact, one of Chomsky's (1980) justifications for an arbitrary UG was that there is a functional benefit associated with merely having the same system as everyone else. That benefit holds whether it is obtained by innate specification or by learning the same arbitrary conventions that everyone else does (Pinker and Bloom, 1990).

For the time being, the discussion throughout this dissertation has tried to steer away from the question of the specific linguistic representations that are at stake, and the question of whether those representations are mere cultural conventions or hardcoded in the human genotype. However, I have also noted that there is increasing dissatisfaction with the nativist account of language acquisition and typology (e.g. Scholz and Pullum, 2006; Haspelmath, 2008b; Evans and Levinson, 2009), and I think the time is ripe for research on language evolution and language processing to probe these questions more directly.

6.4 Conclusion

Linguists, particularly those working on typology, can find plenty of material to inform and inspire new work among the modern psycholinguistic theories that have been proposed to account for human language use (Jaeger and Tily, in press). Psycholinguists, too, can benefit from an understanding of the differences and similarities between the world's languages, without which it will be hard or impossible to build models of language processing that are universal to all languages, and perhaps even informative about why those differences and similarities exist (Hawkins, 2007). These fields, which have in common a much stronger emphasis on empirical observation than many other subfields of linguistics, are now linked by an increasing amount of cross-disciplinary research. This dissertation is intended as a small contribution in that

intersection, offering some proposals and models for ways in which observations made in one field may be linked to the other. These will almost certainly be superseded very shortly, but I hope they will at least provide a point of departure for future research.

Bibliography

- Aissen, J. (1999). Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory*, 17:673–711.
- Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Natural Language and Linguistic Theory*, 21:435–483.
- Allen, C. (1999). *Case marking and reanalysis: Grammatical relations from Old to Early Modern English*. Oxford University Press, Oxford, UK.
- Allen, C. (2000). Obsolescence and sudden death in syntax: The decline of verb-final order in Early Middle English. In Bermudez-Ortero, R., Denison, D., Hogg, R. M., and McCully, C., editors, *Generative Theory and Corpus Studies: A Dialogue from 10 ICEHL*, pages 3–25. Mouton de Gruyter, Berlin.
- Altmann, G. and Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, 30:191–238.
- Anderson, S. R. (1970). A little light on the role of deep structure in semantic interpretation. *NSF Report*, 26(II.1-II.13).
- Anderson, S. R. (1976). On the notion of subject in ergative languages. In Li, C., editor, *Subject and topic*, pages 1–24. Academic Press, New York.
- Anttila, A., Adams, M., and Speriosu, M. (2010). The role of prosody in the English dative alternation. *Language and Cognitive Processes*.
- Ariel, M. (1990). *Accessing Noun Phrase antecedents*. Routledge, London, UK.

- Ariel, M. (2001). Accessibility theory: An overview. In *Text representation: Linguistic and psycholinguistic aspects*. John Benjamins, Amsterdam, The Netherlands.
- Arnold, J. E., Losongco, A., Wasow, T., and Ginstrom, R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language*, 76(1):28–55.
- Bader, M. and Häussler, J. (2010). Word order in German: a corpus study. *Lingua*, 120(3):717–762.
- Bates, E. and MacWhinney, B. (1982). Functionalist approaches to grammar. In Wanner, E. and Gleitman, L., editors, *Language acquisition: The state of the art*. Cambridge University Press, New York, NY.
- Bates, E. and MacWhinney, B. (1989). Functionalism and the competition model. In MacWhinney, B. and Bates, E., editors, *The crosslinguistic study of sentence processing*. Cambridge University Press, New York, NY.
- Behaghel, O. (1909). Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen*, 25:110–142.
- Behaghel, O. (1932). *Deutsche Syntax. Eine geschichtliche Darstellung Vol 4*. Winter, Heidelberg, Germany.
- Benor, S. B. and Levy, R. (2006). The chicken or the egg? A probabilistic analysis of English binomials. *Language*, 82(2):233–278.
- Biberauer, T. and Roberts, I. (2005). Changing EPP parameters in the history of English: accounting for variation and change. *English Language and Linguistics*, 9(1):5–46.
- Bickerton, D. (2009). Recursion: Core of complexity or artifact of analysis? In Givón, T. and Shibatani, M., editors, *Syntactic Complexity: Diachrony, acquisition, neuro-cognition, evolution*, pages 531–544. John Benjamin, Amsterdam, The Netherlands.

- Blevins, J. (2003). *Evolutionary Phonology: The emergence of sound patterns*. Cambridge University Press, Cambridge, UK.
- Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review*, 89:1–47.
- Bock, J. K. (1986). Meaning, sound and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12:575–586.
- Bock, J. K. and Griffin, Z. M. (2000). The persistence of structural priming: Transient activation or implicit learning? *Journal of Experimental Psychology: General*, 129(2):177–192.
- Bock, J. K. and Irwin, D. E. (1980). Syntactic effects of information availability in sentence production. *Journal of Verbal Learning and Verbal Behavior*, 19:467–84.
- Bock, J. K., Loebell, H., and Morey, R. (1992). From conceptual roles to structural relations: bridging the syntactic cleft. *Psychological Review*, 99:150–171.
- Bock, J. K. and Warren, R. K. (1985). Conceptual accessibility and syntactic structure in sentence formulation. *Cognition*, 21(1):47–67.
- Bouma, G. (2011). Production and comprehension in context: The case of word order freezing. In Benz, A. and Mattausch, J., editors, *Bidirectional Optimality Theory*. John Benjamins, Amsterdam, The Netherlands.
- Branigan, H. P. and Feleki, E. (1999). Conceptual accessibility and serial order in Greek language production. In *Proceedings of the 21st Conference of the Cognitive Science Society*, pages 96–102.
- Branigan, H. P., Pickering, M. J., and Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, 118:172–189.

- Bresnan, J., Cueni, A., Nikitina, T., and Baayen, R. H. (2007). Predicting the dative alternation. In Boume, G., Kraemer, I., and Zwarts, J., editors, *Cognitive Foundations of Interpretation*, pages 69–94. Royal Netherlands Academy of Science, Amsterdam.
- Briscoe, T. (1998). Language as a complex adaptive system: Co-evolution of language and of the language acquisition device. In Coppen, P., van Halteren, H., and Teunissen, L., editors, *8th Meeting of Computational Linguistics in the Netherlands*, pages 75–105. Rodopi, Amsterdam, The Netherlands.
- Brysbaert, M. and Mitchell, D. C. (1996). Modifier attachment in sentence parsing: Evidence from Dutch. *The Quarterly Journal of Experimental Psychology*, 49A:664–695.
- Bush, R. R. and Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 58:313–323.
- Bybee, J. and Hopper, P. (2001). *Frequency and the emergence of linguistic structure*. John Benjamins, Amsterdam, The Netherlands.
- Bybee, J. L. and Thompson, S. (1997). Three frequency effects in syntax. In *Berkeley Linguistics Society*, volume 23, pages 378–388.
- Carstairs-McCarthy, A. (2007). Language evolution: What linguists can contribute. *Lingua*, 117(3):503–509.
- Chater, N., Reali, F., and Christiansen, M. (2009). Restrictions on biological adaptation in language evolution. *Proceedings of the National Academy of Sciences*, 106(4):1015–1020.
- Choi, H. W. (1999). *Optimizing structure in context: scrambling and information structure*. CSLI, Stanford, CA.
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge.
- Chomsky, N. (1968). Language and the mind. *Psychology Today*, 1(9).

- Chomsky, N. (1975). *The logical structure of linguistic theory*. University of Chicago Press, Chicago, IL.
- Chomsky, N. (1980). *Rules and representations*. Columbia University Press, New York, NY.
- Chomsky, N., Hauser, M., and Fitch, T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569–1579.
- Christiansen, M. and Kirby, S. (2003). *Language Evolution: The States of the Art*. Oxford University Press, Oxford, UK.
- Clark, B., Goldrick, M., and Konopka, K. (2008). Language change as a source of word order generalizations. In Eckardt, R., Jäger, G., and Veenstra, T., editors, *Variation, Selection, Development: Probing the evolutionary model of language change*, pages 75–102. Mouton de Gruyter, Berlin, Germany.
- Clements, J. C. (2006). Primary and secondary object marking in Spanish. In Clements, J. C. and Yoon, J., editors, *Functional approaches to Spanish syntax: Lexical semantics, discourse, and transitivity*, pages 115–133. Palgrave MacMillan, London.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7:249–253.
- Collins, M. (1999). *Head-driven models for statistical natural language parsing*. PhD thesis, University of Pennsylvania.
- Comrie, B. (1989). *Linguistic universals and language typology*. Blackwell, Oxford, UK, 2nd edition.
- Costa, A., Alario, F.-X., and Sebastián-Gallés, N. (2007). Cross-linguistic research on sentence production. In Gaskell, G. and Altmann, G., editors, *The Oxford handbook of psycholinguistics*, pages 531–565. Oxford University Press, Oxford, UK.

- Crain, S. and Steedman, M. (1985). On not being led up the garden path: The use of context by the psychological parser. In Dowty, D., Karttunen, L., and Zwicky, A., editors, *Natural Language Parsing*. Cambridge University Press, Cambridge, UK.
- Croft, W. (1988). Agreement vs. case marking and direct objects. In Barlow, M. and Ferguson, C. A., editors, *Agreement in Natural Language: Approaches, Theories, Descriptions*, pages 159–180. CSLI, Stanford, CA.
- Croft, W. (2000). *Explaining language change*. Longman, London, UK.
- Croft, W. (2002). *Typology and Universals*. Cambridge University Press, Cambridge, UK.
- Croft, W. (2006). The relevance of an evolutionary model to historical linguistics. In Thomsen, O. N., editor, *Different models of linguistic change*. John Benjamins, Amsterdam, The Netherlands.
- Croft, W. (2010). The origins of grammaticalization in the verbalization of experience. *Linguistics*, 48:1–48.
- Cuetos, F. and Mitchell, D. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in spanish. *Cognition*, 3:73–105.
- Cutler, A. (1997). The comparative perspective on spoken-language processing. *Speech Communication*, 21(1-2):3–15.
- Dawkins, R. (1986). *The Blind Watchmaker*. Longman, London, UK.
- de Hoop, H. and Malchukov, A. (2008). Case marking strategies. *Linguistic Inquiry*, 39:565–587.
- DeLancey, S. (1981). An interpretation of split ergativity and related patterns. *Language*, 57:626–57.
- Dell, G., Burger, L., and Svec, W. (1997). Language production and serial order: A functional analysis and model. *Psychological Review*, 104:123–147.

- Diessel, H. (2007). Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology*, 25(2):108–127.
- Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, 67(3):547–619.
- Dryer, M. S. (1992). The Greenbergian word order correlations. *Language*, 68:81–138.
- Dryer, M. S. (1998). Why statistical universals are better than absolute universals. In *Proceedings of the Chicago Linguistic Society 33*, pages 123–145.
- Evans, N. and Levinson, S. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.
- Ferreira, F. (1994). Choice of passive voice is affected by verb type and animacy. *Journal of Memory and Language*, 33:715–736.
- Ferreira, V. (1996). Is it better to give than donate? syntactic flexibility in language production. *Journal of Memory and Language*, 35:724–755.
- Ferreira, V. and Dell, G. (2000). Effects of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology*, 40:296–340.
- Ferrer i Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters*, 76(6):1228–1234.
- Fine, A. (2006). Ordering preferences of successive prepositional phrases in German. Unpublished BA Thesis, University of North Carolina.
- Fischer, O., van Kemenade, A., Koopman, W., and van der Wurff, W. (2000). *The Syntax of Early English*. Cambridge University Press, Cambridge, UK.
- Flack Potts, K. (2007). Ambiguity avoidance as contrast preservation: Case and word order freezing in Japanese. In Bateman, L., Werle, A., O’Keefe, M., and Reilly, E., editors, *UMass Occasional Papers in Linguistics 32: Papers in Optimality Theory III*, pages 57–88. University of Massachusetts, Amherst, MA.

- Fodor, J., Bever, G., and Garrett, M. (1974). *The Psychology of Language*. McGraw-Hill, New York, NY.
- Ford, M. (1983). A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior*, 22:203–218.
- Frazier, L. and Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6:291–326.
- Fry, J. (2003). *Ellipsis and wa-marking in Japanese conversation*. Routledge, New York, NY.
- Futrell, R. (2010). German noun class as a nominal protection device. Unpublished BA Thesis, Stanford University.
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the interpretation of temporarily ambiguous sentences. *Journal of Memory and Language*, 37:58–93.
- Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27:2865–2873.
- Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multi-level/Hierarchical Models*. Cambridge University Press, Cambridge, UK.
- Gennari, S. P. and MacDonald, M. C. (2008). Semantic indeterminacy in relative clauses. *Journal of Memory and Language*, 58:161–187.
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68:1–76.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Marantz, A., Miyashita, Y., and O’Neil, W., editors, *Image, Language, Brain. Papers from the first Mind Articulation Project Symposium*. MIT Press, Cambridge, MA.

- Gibson, E. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, 54:363–38.
- Gildea, D. and Temperley, D. (2007). Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191.
- Gildea, D. and Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, 34:286–310.
- Givón, T. (2001). *Syntax: an introduction*. John Benjamins, Amsterdam, The Netherlands.
- Goldin-Meadow, S. (2005). What language creation in the manual modality tells us about the foundations of language. *The Linguistic Review*, 22:199–225.
- Goldin-Meadow, S. and Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, 391:279–281.
- Goldin-Meadow, S., So, W. C., Özyürek, A., and Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences*, 105:9163–9168.
- Gordon, P. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, 51(1):97–114.
- Gordon, P., Hendrick, R., and Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27:1411–1423.
- Görlach, M. (1997). *The Linguistic History of English: An Introduction*. Macmillan, Basingstoke, UK.
- Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In Greenberg, J. H., editor, *Universals of Language*, pages 73–113. MIT Press, London.

- Griffiths, T. L., Christian, B. R., and Kalish, M. L. (2008). Using category structures to test iterated learning as a method for revealing inductive biases. *Cognitive Science*, 32:68–107.
- Griffiths, T. L. and Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.
- Griffiths, T. L. and Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31:441–480.
- Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29(2):261–291.
- Gundel, J. K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307.
- Haerberli, E. (2002). Inflectional morphology and the loss of verb-second in english. In Lightfoot, D., editor, *Syntactic Effects of Morphological Change*. Oxford University Press, Oxford, UK.
- Hagiwara, H., Soshi, T., Ishihara, M., and Imanaka, K. (2007). A topographical study on the ERP correlates of scrambled word order in Japanese complex sentences. *Journal of Cognitive Neuroscience*, 19:175–193.
- Hale, J. T. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Haspelmath, M. (2008a). Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics*, 19(1):1–33.
- Haspelmath, M. (2008b). Parametric versus functional explanations of syntactic universals. In Biberauer, T., editor, *The limits of syntactic variation*, pages 75–107. John Benjamins, Amsterdam, The Netherlands.

- Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B. (2005). *The World Atlas of Language Structures*. Oxford University Press, Oxford, UK.
- Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, UK.
- Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, 75:244–285.
- Hawkins, J. A. (2001). Why are categories adjacent? *Journal of Linguistics*, 37:1–34.
- Hawkins, J. A. (2004). *Efficiency and Complexity in Grammars*. Oxford University Press, Oxford, UK.
- Hawkins, J. A. (2007). Processing typology and why psychologists need to know about it. *New Ideas in Psychology*, 25:87–107.
- Hemforth, B., Konieczny, L., Scheepers, C., and Strube, G. (1998). Syntactic ambiguity resolution in German. In Hillert, D., editor, *Syntax and Semantics: A Cross-Linguistic Perspective*. Academic Press, San Diego.
- Hinterhölzl, R. and Petrova, S. (2010). From V1 to V2 in West Germanic. *Lingua*, 120:315–328.
- Hogg, R. (2006). Old English dialectology. In van Kemenade, A. and Los, B., editors, *The Handbook of the History of English*, pages 395–416. Oxford University Press, Oxford, UK.
- Holmes, V. M. and O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative clause sentences. *Journal of Verbal Learning and Verbal Behavior*, 20:417–430.
- Hopper, P. (1987). Emergent grammar. In *Proceedings of the 13th Berkeley Linguistics Conference*, pages 139–157.
- Hopper, P. and Thompson, S. (1980). Transitivity in grammar and discourse. *Language*, 56:251–299.

- Hudson, R. (1984). *Word Grammar*. Blackwell.
- Ito, T., Tahara, S., and Park, W. (1993). *Bun no rikai ni hatasu josi no hataraki: Nihongo to kankokugo wo chuusin ni*. Kazama Shobo, Tokyo, Japan.
- Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. PhD thesis, Stanford University.
- Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, 59(4):434 – 446.
- Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*.
- Jaeger, T. F. and Norcliffe, E. (2009). The cross-linguistic study of sentence production. *Language and Linguistics Compass*, 3:866–887.
- Jaeger, T. F. and Tily, H. (in press). On language ‘utility’: Processing complexity and communicative efficiency. *WIRE: Cognitive Science*.
- Jaeger, T. F. and Wasow, T. (2006). Processing as a source of accessibility effects on variation. In Cover, R. and Kim, Y., editors, *Proceedings of the 31st Annual Meeting of the Berkeley Linguistic Society*, pages 169–180, Ann Arbor, MN. Sheridan Books.
- Jäger, G. and Rosenbach, A. (2006). The winner takes it all — almost: Cumulativity in grammatical variation. *Linguistics*, 44:937–971.
- Jäger, G. and Rosenbach, A. (2008). Priming and unidirectional language change. *Theoretical Linguistics*, 34(2):85–113.
- Jakobson, R. (1936). Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutungen der russischen Kasus. *Travaux du Cercle Linguistique de Prague*, 6:240–288.
- Järvinen, T. and Tapanainen, P. (1998). Towards an implementable dependency grammar. In *Proceedings of the Workshop on Processing of Dependency-Based Grammars*, pages 1–10.

- Johnson, D. E. (2009). Getting off the goldvarb standard: Introducing Rbrul for mixed-effects variable rule analysis. *Language and Linguistics Compass*, 3:359–383.
- Jurafsky, D. (1992). An on-line computational model of human sentence interpretation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-92)*, pages 302–308.
- Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod, R., Hay, J., and Jannedy, S., editors, *Probabilistic Linguistics*. MIT Press, Cambridge, MA.
- Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes and in reading comprehension. *Journal of Experimental Psychology: General*, 3:228–223.
- Kaiser, E. and Trueswell, J. (2004). The role of discourse context in the processing of a flexible word-order language. *Cognition*, pages 113–147.
- Keenan, E. L. (1975). Variation in universal grammar. In Shuy, R. W. and Fasal, R. W., editors, *Analyzing Variation in Language*, pages 136–148. Georgetown University Press, Washington, D.C.
- Keenan, E. L. and Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8(1):63–99.
- Keenan, E. L. and Hawkins, S. (1987). The psychological validity of the Accessibility Hierarchy. In Keenan, E. L., editor, *Universal Grammar: 15 Essays*, pages 60–85. Croom Helm, London, UK.
- Keller, F., Gunasekharan, S., Mayo, N., and Corley, M. (2009). Timing accuracy of web experiments: A case study using the WebExp software package. *Behavior Research Methods*, 41(1):1–12.
- Keller, R. (1994). *On language change: the invisible hand in language*. Routledge, London, UK.

- Kempen, G. and Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: animacy affects linearization independently of grammatical function assignment. In Pechmann, T. and Habel, C., editors, *Multidisciplinary Approaches to Language Production*, pages 173–181. Mouton de Gruyter, Berlin, Germany.
- Ker, N. (1957). *Catalogue of Texts containing Anglo-Saxon*. Oxford University Press, Oxford, UK.
- King, J. and Just, A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, 30:580–602.
- Kiparksy, P. (2008). Universals constrain change, change results in typological generalizations. In Good, J., editor, *Linguistic universals and language change*. Oxford University Press, Oxford, UK.
- Kirby, S. (1999). *Function, Selection and Innateness: the Emergence of Language Universals*. Oxford University Press, Oxford, UK. PhD thesis completed in 1996.
- Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.
- Kohonen, V. (1978). *On the development of English Word Order in Religious Prose around 1000 AD*. Research Institute of the Åbo Akademi Foundation, Turku, Finland.
- Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29:627–645.
- Koopman, W. (1997). Another look at clitics in Old English. *Transactions of the Philological Society*, 95:73–93.
- Koopman, W. (2005). Transitional syntax: postverbal pronouns and particles in Old English. *English Language and Linguistics*, 9(1):47–62.

- Kroch, A. (1989). Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.
- Kroch, A. and Taylor, A. (2000a). *Penn-Helsinki Parsed Corpus of Middle English*, 2nd edition.
- Kroch, A. and Taylor, A. (2000b). Verb-object order in Early Middle English. In Pintzuk, S., Tsoulas, G., and Warner, A., editors, *Diachronic Syntax: Models and Mechanisms*, pages 132–163. Oxford University Press, Oxford, UK.
- Kroch, A., Taylor, A., and Ringe, D. (2000). The Middle English verb-second constraint: A case study in language contact and language change. In Herring, S., Schoesler, L., and van Reenen, P., editors, *Textual parameters in older language*, pages 353–391. John Benjamins, Philadelphia.
- Kytö, M. (1996). Manual to the diachronic part of the Helsinki corpus of English texts. coding conventions and lists of source texts. Helsinki: Department of English, University of Helsinki.
- Labov, W. (1982). Building on empirical foundations. In Lehmann, W. and Malkiel, Y., editors, *Perspectives on Historical Linguistics*. John Benjamins, Amsterdam, The Netherlands.
- Langus, A. and Nespors, M. (2010). Cognitive systems struggling for word order. *Cognitive Psychology*.
- Lee, H. (2001). Markedness and word order freezing. In Sells, P., editor, *Formal and Empirical Issues in Optimality Theoretic Syntax*. CSLI, Stanford, CA.
- Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity*. PhD thesis, Stanford University.
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

- Lewis, R. L. and Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:1–45.
- Lewis, R. L., Vasishth, S., and Van Dyke, J. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, 10(10):447–454.
- Li, C. N. (1977). *Mechanisms of syntactic change*. University of Texas Press, Austin, TX.
- Lieven, E., Tomasello, M., Behrens, H., and Speares, J. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30:333–370.
- Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. MIT Press, Cambridge, MA.
- Los, B. (2002). The loss of the indefinite pronoun *man*: syntactic change and information structure. In Fanego, T., López-Couso, M. J., and Pérez-Guerra, J., editors, *English Historical Syntax and Morphology*, pages 181–202. John Benjamins, Amsterdam, The Netherlands.
- Los, B. (2009). The consequences of the loss of verb-second in English: Information structure and syntax in interaction. *English Language and Linguistics*, 13:97–125.
- Love, T. and Swinney, D. (1998). The influence of canonical word order on structural processing; real-time processing from a cross-linguistic perspective. In Hillert, D., editor, *Psycholinguistics: a Cross-Linguistic Perspective*, pages 153–166. Academic Press, New York, N.Y.
- MacDonald, M. C., Pearlmutter, N. J., and Seidenberg, M. S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676–703.
- MacDonald, M. C. and Seidenberg, M. S. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23:569–588.

- MacWhinney, B., Bates, E., and Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23:127–150.
- Malchukov, A. (2008). Animacy and asymmetries in differential case marking. *Lingua*, 118(2):203–221.
- Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- Mazuka, R., Itoh, K., and Kondo, T. (2002). Cost of scrambling in Japanese sentence processing. In Nakayama, M., editor, *Papers from International East Asian Psycholinguistics Workshop*. CSLI Publications, Stanford, CA.
- McDonald, J., Bock, J., and Kelly, M. (1993). Word and world order: semantic, phonological and metrical determinants of serial position. *Cognitive Psychology*, 25:188–230.
- McDonald, J. L. (1986). The development of sentence comprehension strategies in English and Dutch. *Journal of Experimental Child Psychology*, 41:317–335.
- McDonald, J. L. (1987). Assigning linguistic roles: The influence of conflicting cues. *Journal of Memory and Language*, 26:100–117.
- McDonald, J. L. and MacWhinney, B. (1989). Maximum likelihood models for sentence processing research. In MacWhinney, B. and Bates, E., editors, *The crosslinguistic study of sentence processing*, pages 397–421. Cambridge University Press, New York, NY.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63:81–97.
- Miller, G. A. and Chomsky, N. (1963). Finitary models of language users. In Luce, R. D., Bush, R. R., and Galanter, E., editors, *Handbook of mathematical psychology, volume 2*, pages 419–491. Wiley, New York.

- Miller, G. A. and McKean, K. O. (1964). A chronometric study of some relations between sentences. *Quarterly Journal of Experimental Psychology*, 16:267–308.
- Mitchell, D. C., Cuetos, F., and Corley, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research*, 24:469–488.
- Miyamoto, E. T. and Takahashi, S. (2002). Sources of difficulty in processing scrambling in Japanese. In Nakayama, M., editor, *Sentence Processing in East Asian Languages*. CSLI Publications, Stanford, CA.
- Munro, R., Bethard, S., Kuperman, V., Tzuyin Lai, V., Melnick, R., Potts, C., Schnoebelen, T., and Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, NAACL*, Los Angeles, CA.
- Nakayama, M. (1995). Scrambling and probe recognition. In Mazuka, R. and Nagai, N., editors, *Japanese Sentence Processing*. Erlbaum, Hillsdale, NJ.
- Nakayama, M., Vasishth, S., and Lewis, R. (2006). Difficulty of certain sentence constructions in comprehension. In *Handbook of East Asian Psycholinguistics, volume 2*. Cambridge University Press, Cambridge, UK.
- Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the Cognitive Science Society annual meeting*.
- Narayanan, S. and Jurafsky, D. (2002). A Bayesian model predicts human parse preference and reading time in sentence processing. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 59–65. MIT Press, Cambridge, MA.
- Newmeyer, F. (2005). *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press, Oxford, UK.

- Niyogi, P. (2006). *The Computational Nature of Language Learning and Evolution*. MIT Press, Cambridge, MA.
- Niyogi, P. and Berwick, R. C. (1998). The logical problem of language change: A case study of european portuguese. *Syntax: A Journal of Theoretical, Experimental, and Interdisciplinary Research*, 1(2):192–205.
- Nordlinger, R. (1998). *Constructive Case: Evidence from Australian Languages*. CSLI Publications, Stanford, CA.
- Osgood, C. E. (1971). Where do sentences come from? In Steinberg, D. D. and Jakobovits, L. A., editors, *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*. Cambridge University Press, London, UK.
- Oudeyer, P.-Y. and Kaplan, F. (2007). Language evolution as a Darwinian process: Computational studies. *Cognitive Processes*, 8:21–25.
- Park, Y. A. and Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT) conference*.
- Perfors, A. (2002). Simulated evolution of language: a review of the field. *Journal of Artificial Societies and Social Simulation*, 5(2).
- Petrova, S. (2006). A discourse-based approach to verb placement in early West Germanic. *Interdisciplinary Studies on Information Structure*, 5:153–185.
- Petrova, S. and Solf, M. (2008). Rhetorical relations and verb placement in early Germanic. a cross linguistic study. In Fabricius-Hansen, C. and Ramm, W., editors, *‘Subordination’ vs. ‘Coordination’ in Sentence and Text. A Cross-linguistic Perspective*, pages 333–351. John Benjamins, Amsterdam, The Netherlands.
- Pickering, M. and Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

- Pinker, S. and Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–784.
- Pinker, S. and Jackendoff, R. (2005). The faculty of language: what’s special about it? *Cognition*, 95(2):201–236.
- Pintzuk, S. (1999). *Phrase structures in competition: variation and change in Old English word order*. Garland, New York.
- Pintzuk, S. (2002). Verb-object order in old english: Variation as grammatical competition. In Lightfoot, D. W., editor, *Syntactic effects of morphological change*, pages 276–299. Oxford University Press, Oxford, UK.
- Pintzuk, S. (2005). Arguments against a universal base: evidence from Old English. *English Language and Linguistics*, 9(1):115–138.
- Pintzuk, S. and Kroch, A. (1989). The rightward movement of complements and adjuncts in the Old English of Beowulf. *Language Variation and Change*, 1:115–143.
- Pintzuk, S. and Taylor, A. (2004). Objects in Old English: why and how early English isn’t Icelandic. *York Papers in Linguistics*, pages 137–150.
- Pintzuk, S. and Taylor, A. (2006). The loss of OV order in the history of English. In van Kemenade, A. and Los, B., editors, *The Handbook of the History of English*, pages 249–278. Oxford University Press, Oxford, UK.
- Prat-Sala, M. (1997). *The production of different word orders: a psycholinguistic and developmental approach*. PhD thesis, University of Edinburgh.
- Prat-Sala, M. and Branigan, H. (2000). Discourse constraints on syntactic processing in language production: a crosslinguistic study in English and Spanish. *Journal of Memory and Language*, 42:168–182.
- Pullum, G. K. and Gazdar, G. (1982). Natural languages and context-free languages. *Linguistics and Philosophy*, 4:471–504.

- Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1972). *A Grammar of Contemporary English*. Seminar Press, New York, NY.
- Real, F. and Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language*, 53:1–23.
- Reber, A. S. (1987). The rise and (surprisingly rapid) fall of psycholinguistics. *Synthese*, 72:325–339.
- Roland, D., Dick, F., and Elman, J. L. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57:348–379.
- Rosenbach, A. (2005). Animacy versus weight as determinants of grammatical variation in English. *Language*, 81:613–644.
- Rosenbach, A. (2008). Language change as cultural evolution: Evolutionary approaches to language change. In Eckardt, R., Jäger, G., and Veenstra, T., editors, *Variation, Selection, Development: Probing the Evolutionary Model of Language Change*, pages 23–74. Mouton de Gruyter, Berlin, Germany.
- Sakamoto, T. (2001). Bun-no rikai: kakimaze bun [Understanding of sentences: Scrambling]. *Gekkan Gengo*, 30(10):106–111.
- Sasaki, Y. and MacWhinney, B. (2006). The competition model. In Nakayama, M. and Li, P., editors, *The Handbook of East Asian Psycholinguistics*, pages 307–314. Cambridge University Press, Cambridge, UK.
- Sawyer, P. H. (1968). *Anglo-Saxon Charters, An Annotated List and Bibliography*. Royal Historical Society, London.
- Schleewsky, M. and Bornkessel, I. (2003). On incremental interpretation: degrees of meaning accessed during sentence comprehension. *Lingua*, 114(9-10):1213–1234.
- Scholz, B. C. and Pullum, G. K. (2006). Irrational nativist exuberance. In Stainton, R., editor, *Contemporary Debates in Cognitive Science*, pages 59–80. Basil Blackwell, Oxford, UK.

- Seoane, E. (2006). Information structure and word order change: The passive as an information-rearranging strategy in the history of English. In van Kemenade, A. and Los, B., editors, *The Handbook of the History of English*, pages 360–391. Oxford University Press, Oxford, UK.
- Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence in its Pragmatic Aspects*. Reidel.
- Shackle, C. (1972). *Punjabi*. English Universities Press, London, UK.
- Siewierska, A. and Bakker, D. (2008). Case and alternative strategies. In Malchukov, A. M. and Spencer, A., editors, *Handbook of Case*, pages 290–303. Oxford University Press, Oxford, UK.
- Sleator, D. and Temperley, D. (1993). Parsing English with a Link Grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*.
- Snider, N. and Zaenen, A. (2006). Animacy and syntactic structure: Fronted NPs in English. In Butt, M., Dalrymple, M., and King, T. H., editors, *Intelligent Linguistic Architectures: Variations on Themes by Ron Kaplan*. CSLI Publications, Stanford, CA.
- Snow, R., O’Connor, B., Jurafsky, D., and Ng, A. Y. (2008). Cheap and fast but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-2008)*.
- Spencer, A. and Otoguro (2005). Limits to case: A critical survey of the notion. In Amberber, M. and de Hoop, H., editors, *Competition and variation in natural languages: the case for case*. Elsevier, Amsterdam, The Netherlands.
- Speyer, A. (2008). *Topicalization and Clash Avoidance*. PhD thesis, University of Pennsylvania.
- Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1999). *WinBUGS Version 1.2 User Manual*. MRC Biostatistics Unit.

- Spivey-Knowlton, M. and Sedivy, J. (1995). Parsing attachment ambiguities with multiple constraints. *Cognition*, 55:227–267.
- Sridhar, S. N. (1989). Cognitive structures in language production: A crosslinguistic study. In MacWhinney, B. and Bates, E., editors, *The Crosslinguistic study of Sentence Processing*. Cambridge University Press, Cambridge, UK.
- Stallings, L., MacDonald, M., and O’Seaghdha, P. (1998). Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-NP shift. *Journal of Memory and Language*, 39:392–417.
- Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332.
- Steels, L. (2001). Grounding symbols through evolutionary language games. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, pages 211–226. Springer Verlag, London, UK.
- Steels, L. (2002). Iterated learning versus language games. Two models for cultural language evolution. In Hemelrijk, C., editor, *Proceedings of the International Workshop of Self-Organization and Evolution of Social Behaviour*, Monte Verità, Switzerland.
- Steels, L. (2006). How to do experiments in artificial language evolution and why. In *Proceedings of the 6th International Conference on the Evolution of Language*, pages 323–332.
- Szmrecsányi, B. M. (2004). On operationalizing syntactic complexity. In Purnelle, G., Fairon, C., and Dister, A., editors, *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, volume 2, pages 1032–1039. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.
- Tanaka, M., Branigan, H., and Pickering, M. (2005). The role of animacy in Japanese sentence production. Paper presented at the CUNY Sentence Processing Conference, Tucson, AZ.

- Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268:1632–1634.
- Taylor, A., Warner, A., Pintzuk, S., and Beths, F. (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Oxford Text Archive.
- Temperley, D. (2005). The dependency structure of coordinate phrases: A corpus approach. *Journal of Psycholinguistic Research*, 34:577–601.
- Temperley, D. (2007). Minimization of dependency lengths in written English. *Cognition*, 105:300–333.
- Temperley, D. (2008). Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, 15(3):256–282.
- Tesnière, L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris, France.
- Tippets, I. and Schwenter, S. (2007). Relative animacy and differential object marking in Spanish. Paper presented at the NWAV Conference, Pennsylvania, PA.
- Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA.
- Tomlin, R. S. (1986). *Basic word order: functional principles*. CroonHelm, London, UK.
- Tomlin, R. S. (1995). Focal attention, voice, and word order. In Dowing, P. and Noonan, M., editors, *Word Order in Discourse*, pages 517–552. John Benjamins, Amsterdam, The Netherlands.
- Traugott, E. (2007). Old English left-dislocations: their structure and information status. *Folia Linguistica*, 41(3-4):405–441.
- Trueswell, J., Tanenhaus, M., and Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33:285–318.

- Trueswell, J., Tanenhaus, M., and Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden paths. *Journal of Experimental Psychology: Learning, Memory and Language*, 19(3):528–553.
- Tsunoda, T. (1985). Remarks on transitivity. *Journal of Linguistics*, 21:385–396.
- Ueno, M. and Kluender, R. (2003). Event-related indices of Japanese scrambling. *Brain and Language*, 86:243–271.
- Ueno, M. and Polinsky, M. (2009). Does headedness affect processing? a new look at the vo-ov contrast. *Journal of Linguistics*, 45:675–710.
- van Kemenade, A. (1987). *Syntactic Case and Morphological Case in the History of English*. Foris, Dordrecht, The Netherlands.
- van Nice, K. Y. and Dietrich, R. (2003). Task sensitivity of animacy effects: evidence from German picture descriptions. *Linguistics*, 41:825–849.
- Van Valin, R. D. (2004). Semantic macroroles in role and reference grammar. In Kailuweit, R. and Hummel, M., editors, *Semantische Rollen*, pages 62–82. Narr, Tübingen, Germany.
- Vasisth, S. and Lewis, R. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.
- Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature*, 150:563–565.
- Wanner, E. and Maratsos, M. (1978). An ATN approach to comprehension. In Halle, M., Bresnan, J., and Miller, G. A., editors, *Linguistic Theory and Psychological Reality*, chapter 3, pages 119–161. MIT Press, Cambridge, MA.
- Warren, T. and Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85:79–112.

- Wasow, T. (1997). Remarks on grammatical weight. *Language Variation and Change*, 9:81–105.
- Wasow, T. (2002). *Postverbal Behavior*. CSLI Publications, Stanford, CA. Distributed by University of Chicago Press.
- Wedel, A. (2003). Self-organization and categorical behavior in phonology. In *Proceedings of the Berkeley Linguistics Society*, volume 29, pages 611–622.
- Weinreich, U., Labov, W., and Herzog, M. (1968). Empirical foundations for a theory of language change. In Lehmann, W. and Malkiel, Y., editors, *Directions for historical linguistics: a symposium*, pages 95–188. University of Texas Press, Austin, TX.
- Wolff, S., Schlesewsky, M., Hirotani, M., and Bornkessel-Schlesewsky, I. (2008). The neural mechanisms of word order processing revisited: Electrophysiological evidence from Japanese. *Brain and Language*, 107:133–157.
- Yamashita, H. (1997). The effects of word-order and case marking information on the processing of Japanese. *Journal of Psycholinguistic Research*, 26:163–188.
- Yamashita, H. (2002). Scrambled sentences in Japanese: Linguistic properties and motivations for production. *Text*, 22(4):597–633.
- Yamashita, H. and Chang, F. (2001). “Long before short” preference in the production of a head-final language. *Cognition*, 81(2):B45–B55.
- Yang, C. (2000). Internal and external forces in language change. *Language Variation and Change*, 12(3):231–250.
- Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford, UK.
- Yngve, V. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104:444–466.

- Zagar, D., Pynte, J., and Rativeau, S. (1997). Evidence for early-closure attachment on first-pass reading times in French. *Quarterly Journal of Experimental Psychology*, 50A:421–38.