Natural Language Processing

The Lexical Basis of Sentence Processing

Formal, computational and experimental issues

Edited by Paola Merlo
and Suzanne Stevenson

The Lexical Basis of Sentence Processing

Natural Language Processing

Editor

Prof. Ruslan Mitkov
School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford St.
Wolverhampton WV1 1SB, United Kingdom
Email: R.Mitkov@wlv.ac.uk

Advisory Board

Christian Boitet (University of Grenoble)
John Carroll (University of Sussex, Brighton)
Eugene Charniak (Brown University, Providence)
Eduard Hovy (Information Sciences Institute, USC)
Richard Kittredge (University of Montreal)
Geoffrey Leech (Lancaster University)
Carlos Martin-Vide (Rovira i Virgili Un., Tarragona)
Andrei Mikheev (University of Edinburgh)
John Nerbonne (University of Groningen)
Nicolas Nicolov (IBM, T. J. Watson Research Center)
Kemal Oflazer (Sabanci University)
Allan Ramsey (UMIST, Manchester)
Monique Rolbert (Université de Marseille)
Richard Sproat (AT&T Labs Research, Florham Park)
Keh-Yih Su (Behaviour Design Corp.)
Isabelle Trancoso (INESC, Lisbon)
Benjamin Tsou (City University of Hong Kong)
Jun-ichi Tsujii (University of Tokyo)
Evelyne Tzoukermann (Bell Laboratories, Murray Hill)
Yorick Wilks (University of Sheffield)

Volume 4

The Lexical Basis of Sentence Processing: Formal, computational and experimental issues
Edited by Paola Merlo and Suzanne Stevenson

The Lexical Basis of Sentence Processing

Formal, computational and experimental issues

Edited by

Paola Merlo
University of Geneva

Suzanne Stevenson
University of Toronto

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

The Lexical basis of sentence processing : formal, computational and experimental issues / edited by Paola Merlo and Suzanne Stevenson.

p. cm. (Natural Language Processing, ISSN 1567–8202 ; v. 4)

"This volume derives from the special conference session...held in conjunction with the 11th Annual CUNY Conference on Human Sentence Processing, March 19-21, 1998"–Pref.

Includes bibliographical references and index.

1. Grammar, Comparative and general--Syntax--Congresses. 2. Lexicology--Congresses. 3. Phycholiguistics--Congresses. 4. Computational linguistics--Congresses. I. Merlo, Paola, 1949- II. Stevenson, Suzanne. III. Natural Language Processing (Amsterdam, Netherlands); v.4.

P295.L488 2002
415--dc21 2001058313
ISBN 90 272 4987 3 (Eur.) / 1 58811 156 3 (US) (Hb; alk. paper)

© 2002 – John Benjamins B.V.
No part of this book may be reproduced in any form, by print, photoprint, microfilm, or any other means, without written permission from the publisher.

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Part I: Fundamental issues

Part II: Division of labour between syntax and the lexicon

Part III: Details of lexical entries

Preface

This volume derives from the special conference session entitled "The Lexical Basis of Sentence Processing: Formal and Computational Issues," held in conjunction with the 11th Annual CUNY Conference on Human Sentence Processing, March 19–21, 1998. The special session highlighted talks and posters on current theories of the lexicon from the perspective of its use in sentence understanding. Lexical influences on processing are currently a major focus of attention in psycholinguistic studies of sentence comprehension; however, much of the work remains isolated from the vast amount of scientific activity on the topic of the lexicon in other subdisciplines. In organising the special session, we felt that the time was ripe to bring together researchers from these different perspectives to exchange ideas and information that can help to inform each others' work. Participants included a multi-disciplinary slate from theoretical linguistics, computational linguistics, and psycholinguistics, representing various theoretical frameworks within each of these disciplines. The special session was quite successful with about 250 registrants, many of whom do not belong to the usual CUNY crowd.

A primary motivation for the special session was that a focus of attention on the lexicon has the potential to bring the structural and probabilistic approaches to sentence processing closer together. To gain a deeper understanding of the lexicon and its impact on processing, we need to elaborate both the structure and the probabilistic content of lexical representations, which together influence sentence interpretation. These questions bring up general issues in the architecture of the mind, and meet up with work in computer science, linguistics, and philosophy on the relation between conceptual knowledge, grammatical knowledge, and statistical knowledge.

The contents of this volume reflect this range of issues from various perspectives in the multidisciplinary study of the lexicon in language processing. A selection of the contributors to the special session were invited to prepare written versions of their presentations to be included in the volume. The preparation of the final version of the papers was assisted by very careful and detailed multiple peer reviewing. Very few of the reviewers were also contributors to the volume. We would like to acknowledge here the fundamental role of the anonymous reviewers to the success of this volume.

The editors have written an introductory chapter intended to guide the reader to the content of the volume and to state our view of the connections between approaches to the lexicon across the disciplines. Our intention is to provide an integrated cross-disciplinary viewpoint on many of the issues that arise in developing lexicalist theories of sentence processing. The introduction highlights areas of overlap across the fields and the potential for the differing perspectives to be mutually informing. We hope that this effort will spark further interest in cross-disciplinary work both in psycholinguistics and in computational linguistics.

This volume owes much to a large number of people to whom the conference organisers and the editors of this volume are indebted. The special session could not have taken place without the generous support of the following organisations: the National Science Foundation, Rutgers University, City University of New York, The Ohio State University, the University of Pennsylvania, the University of Rochester, and the University of Southern California. The organisers were greatly assisted by the conference administrator at the Rutgers Center for Cognitive Science, Trish Anderson, and her support staff, Kevin Keating and Carol Butler-Henry.

Finally, the editors would like to thank all the contributors for their patience and helpfulness, and the editors at John Benjamins: Dan Jurafsky, Ruslan Mitkov, and especially Kees Vaes. We are also grateful for the support from our current and past institutions: the University of Geneva, the University of Pennsylvania, Rutgers University, and the University of Toronto.

Paola Merlo
Geneva, Switzerland
Suzanne Stevenson
Toronto, Canada

Words, numbers and all that

The lexicon in sentence understanding

Suzanne Stevenson and Paola Merlo
University of Toronto and University of Geneva

1. Cross-disciplinary issues in lexical theories

It is hardly a controversial statement that the acquisition and processing of language require knowledge of its words. Yet, the type and use of information encoded in a lexical entry, the relation of words to each other in the lexicon, and the relationship of the lexicon to the grammar, are complex and unsettled issues, on which researchers hold very different views. While there is a recent consensus that mechanisms operating in the lexicon are not substantially different from those operating in the syntax, there are differences on whether the syntax is in the lexicon, or the lexicon is in the syntax. On the one view, the lexicon is a static repository of very rich representations, which regulate the composition of words to an extent that goes beyond the phrase, while on the other view the lexicon is dynamically generated as a result of composition and competition mechanisms, largely syntactic in nature. Roughly speaking, computational linguistics and psycholinguistics in general follow the first view, and the two fields are converging on some similar lexicalised, probabilistic models of grammars. Theoretical linguistics has recently proposed models of the latter type.

In computational linguistics, work on the lexicon has stemmed from two different areas of research: parsing and grammar formalisms, and construction of electronic databases (lexicography). In the area of parsing, the interest in probabilistic models and lexicalised grammars did not develop simultaneously. Parsers based on probabilistic context-free grammars were motivated by the difficulties in building robust, large-scale systems using the explicit representation of linguistic knowledge. Large corpus annotation efforts and the creation of tree-banks (text corpora annotated with syntactic structures) enabled researchers to develop and automatically train probabilistic models of syntactic disambiguation (Marcus, Santorini, and Marcinkiewicz 1993). In an attempt to take advantage of the insights gained in the area of statistical speech processing, computational linguists initially adopted very simplified statistical models of grammar and parsing, abandoning the more sophisticated lexicalised feature-based formalisms (Magerman and Marcus 1991; Magerman and Weir 1992; Resnik 1992; Schabes 1992).

However, it soon became apparent that the success of probabilistic context-free grammars was limited by the strong (and incorrect) assumption of probabilistic independence of rule applications (Black, Jelinek, Lafferty, Magerman, Mercer, and Roukos 1992; Charniak 1996; Johnson 1998). The search for more context-sensitive models of disambiguation led to the development of probabilistic models that rely heavily on lexical heads. Current models of probability assume a lexicalised grammar, in which a syntactic rule is conditioned on its lexical head, as well as on the heads of its dependent constituents. In this way, a probability model is defined that takes into account lexical dependencies that go beyond a single context-free rule (Brew 1995; Collins 1996; Abney 1997; Charniak 1997; Collins 1997; Ratnaparkhi 1997; Johnson, Geman, Canon, Chi, and Riezler 1999; Srinivas and Joshi 1999). These models derive their formal specifications from (non-statistical) lexicalised grammar formalisms, where the needs for empirical coverage have similarly led to precise definitions of such grammars (e.g., Bresnan 1982; Pollard and Sag 1987; Mel'cuk 1988; Joshi and Schabes 1997).

The lexicalisation of computational models of grammar and parsing, and the emphasis on robust systems, has brought to the forefront one of the major practical and scientific problems in large-scale linguistic applications, namely the difficulty in acquiring and encoding lexical information. Manually building the rich lexical representations that are a central component of linguistic knowledge is time-consuming, error prone, and difficult, as shown by the effort required to produce electronic databases such as Wordnet (Miller, Beckwith, Fellbaum, Gross, and Miller 1990; Fellbaum 1998) and verb classifications such as Levin's (Levin 1993). The complexity of the task of hand-building lexical entries has sparked interest in extending the learning approaches developed for parsing to the inductive learning of fine-grained lexical classifications. By exploiting the implicit syntactic and lexical information encoded in annotated corpora, lexical knowledge can be induced from the statistical analysis of distributional data (Brent 1993; Briscoe and Carroll 1997; Dorr 1997; Mc-

Carthy 2000; Lapata and Brew 1999; Schulte im Walde 2000; Stevenson and Merlo 2000; Merlo and Stevenson 2001; Siegel and McKeown 2001).

In sentence processing, the path of development of lexical, probabilistic theories has been different, but reaching similar conclusions. In this context, a key question is the degree to which lexical information underlies distinctions in on-line processing difficulty, especially in the process of ambiguity resolution. Early work in sentence processing mirrored the emphasis on the syntactic component in grammatical theory, by focusing on the large-grained structural properties of interpretations as they are developed incrementally. For example, the most widely known statements of preference, Minimal Attachment and Late Closure (Frazier 1978), rely solely on general properties of the size and locality of incremental additions to a partial syntactic structure. Beginning with the work of Ford, Bresnan, and Kaplan (1982), there has been a gradual shift to emphasizing more and more the influence of individual lexical information within such structure-based accounts. For example, Pritchett (1992) examines the role of the lexical head of a phrase in determining its basic properties, and the impact of the argument structure of a verb on preferred structural representations. However, while some lexical properties are assumed to influence processing decisions in this type of structure-based account, they are viewed as secondary to structural information.

The recent lexicalist constraint-based approach in sentence processing takes the lexical basis of language comprehension much further, suggesting that perhaps all of the relevant processing distinctions can be traced to distinctions in lexical information (e.g., MacDonald, Pearlmutter, and Seidenberg 1994; Spivey-Knowlton and Sedivy 1995; Spivey-Knowlton, Trueswell, and Tanenhaus 1993; Trueswell 1996). On this view, interpretation is an incremental process of satisfying constraints associated with lexical entries. Dynamical models inspired by the connectionist literature are used to describe the integration of multiple numerically-weighted constraints (e.g., Spivey and Tanenhaus 1998; Tabor and Tanenhaus 1999). Typically, the frequency of lexical features and their co-occurrence is believed to be the basis of the weights, and therefore frequency plays a primary role in determining sentence processing behaviour. These models, analogously to the corpus-based work in computational linguistics, tie together processing and learning, as much of the lexical frequency information needed for parsing is learned from exposure.

In both disciplines, increased lexicalisation raises interesting issues concerning the role of frequency (or probability) in parsing, the conception of parsing itself, and the issue of incrementality. In computational linguistics, the automatic acquisition of lexical knowledge through statistical analysis entails an emphasis on frequency information, as corpus counts are used to estimate probabilities. This brings up the important issue of what are the relevant entities or features of entities to count in order to achieve accurate probabilistic parsers. In lexicalist sentence processing as well, this issue of what to count has been a focus of much attention – i.e., for exactly what types of lexical information do humans keep track of frequencies (morphological, syntactic, thematic, semantic, etc.), and what is the grain of frequency information that influences human parsing (Mitchell, Cuetos, Corley, and Brysbaert 1995; Gibson, Schütze, and Salomon 1996)?

One also sees an influence of lexicalisation on how parsing is conceived. Both in computational linguistics and in lexically based models of sentence processing, there is consensus that the processing primitives are rather large units specifying a lexical head and all its argument relations. In this way, the primitive grammatical objects correspond to a lexicalised specification of the non-recursive set of substructures that come into play in building a parse. This notion can be realized in one of two ways. One option is an explicit representation of grammar-in-the-lexicon, such as in tree-adjoining grammars or lexicalist constraint-based models (MacDonald, Pearlmutter, and Seidenberg 1994; Srinivas and Joshi 1999), in which the lexical entries are themselves tree structures. Lexicalised probabilistic models of parsing adopt an alternative approach which implicitly associates grammatical rules with lexical entries, by conditioning each rule on the lexical heads of it and its children. In conjunction with the emphasis on frequencies and probabilities, parsing is viewed within both these formulations as a competition of structures, whose respective probabilities depend both on the substructures composing them and on the words that they contain. In connectionist models, this competition is represented directly in the component processors, through their levels of activation. In a probabilistic parser, the competition is encoded as the ranking given by the probability of each parse.

Lexicalisation raises further questions related to incremental parsing. First, lexicalisation tests the limits of incrementality, as it assumes that each word can be directly integrated into the previously determined structure. Also, the related assumption that the domain of influence – the domain of locality – of each individual word extends to units larger than the immediate phrase raises interesting technical issues of how exactly the integration of each word in the parse tree occurs. In this view, each incoming word gives rise to a complex set of interactions with the existing structure, where several mutual constraints – determined by both the word and by the structure – have to be satisfied. The proper description of these mechanisms has consequences for the choice of both the grammatical representations encoded in the lexicon and the parsing architectures.

While computational linguistics and sentence processing have focused on the locus of lexical statistics and the process of lexically-guided parsing, theoretical linguistics has focused on the organization of the lexicon. The increased complexity of lexical entries in computational linguistics and sentence processing appears in contrast with the final goal in theoretical linguistics, which consists in reducing the lexicon to its primitive components and describing its regularities. Much linguistic knowledge is situated in how lexical entries are organized with respect to each other. On this view, lexical entries tend to be simplified, while the organisation of words in the lexicon and the integration of words with sentence construction becomes more complex.

For example, Levin's (1993) work on verb classes aims at reducing the information in a lexical entry to its primitive meaning components (see also Levin 1985; Pinker 1989). Under the hypothesis that semantic properties of verbs largely determine their syntactic behaviour, the linguistic knowledge about a verb consists in its specific set of meaning components along with general relations between each meaning component and its possible syntactic expressions. The architecture of Wordnet (Miller et al. 1990), an electronic lexicon organised on psycholinguistic principles, provides another example of the complexity of organisation in the mental lexicon. Components of meaning do not always appear in the actual definition of the word, rather they are sometimes part of the organisational structure of the lexicon, as relational notions between entities. For example, the notion of causation does not appear as part of the meaning component of certain verbs, such as *melt*, but as the relation connecting two senses of the verb *melt*, the intransitive (not causative) and the transitive (causative).

Thus, as lexical entries are simplified, the organizational knowledge itself becomes more complex, in many cases appearing syntactic in its formal nature. Furthermore, as the lexical entries become impoverished, the notion of "projecting" syntactic structure from the lexicon becomes less tenable. In fact, taking this form of reasoning to its logical conclusion, current conceptions in Optimality Theory view the lexicon no longer as the input to the sentence level but as the result of syntactic competition (Prince and Smolensky 1993). Constructionist approaches to the lexicon also propose to simplify the lexical entries and devise more complex relations to link them. For instance, some have argued that the lexical entries themselves are syntactic – that is, lexical structure is not only predictive of syntactic structure, but is itself subject to syntactic processes within the lexicon, such as the Head Movement Constraint (Hale and

Keyser 1993). Others have proposed that syntactic structures themselves are the bearers of meaning (Fillmore 1988; Goldberg 1995).

Thus, there seems to be consensus across disciplines that there is more structure in the lexicon than previously assumed, and that similar mechanisms operate in the lexicon and in the syntax. One area of disagreement is in the conception of lexical entries themselves. In theoretical linguistics, these are generally assumed to be minimal, while in computational linguistics and sentence processing, they are by contrast entire subtrees. A second issue on which the fields differ is the nature of the competitive processes at work in the lexicon. In sentence processing and in computational linguistics, lexical processes are fundamentally a numeric competition, while in Optimality Theory the lexicon can be viewed as the result of a non-probabilistic syntactic competition process. In evaluating the arguments on the differing sides of these two issues, it is important to broaden one's view beyond a particular discipline, as the evidence for an encompassing theory of language will come from many sources.

For example, the lack of consensus on what constitutes a lexical entry is often resolved in favour of the most succinct description, on the basis of Occam's razor. However, this is appropriate only if the sole criterion is simplicity of representation. Given that the language faculty has a computational component, simplification in the description of knowledge might lead to more complex computations. A more appropriate metric would be one in the spirit of the Minimum Description Length principle, which characterizes simplicity as a function of both the complexity of the data and the complexity of the computation (Rissanen 1989). This points to the inherent multidisciplinarity of the lexical enterprise, as this principle clearly cannot be applied, and consequently the preferable approach cannot be decided on, in the absence of a precise model of both the representations and the computations involved.

Another example arises concerning the nature of lexical competition. The existence of competition effects in sentence processing has been assumed to argue in favour of a probabilistic, connectionist architecture. But frequency effects can be incorporated in a natural way in other types of models, such as probabilistic parsers that are not connectionist in nature, and competition can be naturally expressed in a symbolic model (as in Optimality Theory). As some of the contributions in this volume argue convincingly (e.g., Spivey et al.), it is premature to infer architectural organisations from currently available empirical evidence. Rather, a broader range of both human and computational experiments must be pursued to provide mutually constraining evidence about plausible language processing architectures.

The resolution of these open questions in lexical theorising thus requires accumulation of evidence from all three disciplines of theoretical linguistics, computational linguistics, and sentence processing. Very precise theories must be developed, along with more elaborated computational models and new experimental methods, that, as evidence accrues, will be mutually constraining. This volume represents an attempt to bring together leading research on the lexicon across these three fields, in particular focusing on work that influences computational views of the human sentence processor. In the remainder of this introduction, we outline brief summaries of the contributions to the volume, and discuss current issues in theories of lexical information and processing that cut across the groups of papers collected here.

2. Tour of the volume

This volume derives from the special conference session entitled "The Lexical Basis of Sentence Processing: Formal and Computational Issues," held in conjunction with the 11th Annual CUNY Conference on Human Sentence Processing, March 19–21, 1998. The special session focused on current theories of the lexicon from the perspective of its use in sentence understanding. Participants included a multidisciplinary slate from theoretical linguistics, computational linguistics, and psycholinguistics, representing various theoretical frameworks within each of these disciplines.

By analogy with the conference special session, the volume is organized into three parts: Part I presents theoretical proposals on the lexicon and discusses their relation to sentence processing; Part II explores the relationship between syntactic and lexical processing; and Part III investigates more specific issues about the content of lexical entries. Here we briefly discuss the contributions within each part; we return in Sections 3 and 4 to issues that cut across these broad topic areas.

2.1 Part I: Fundamental issues

Part I of the volume contains papers that elaborate on foundational issues concerning the nature of the lexicon and its connection to processing. Bresnan and Fodor address two fundamental aspects of a lexical item: its lexico-syntactic paradigm (i.e., its relation to other items in the lexicon), and its combinatory properties (i.e., its relation to phrases in which it can occur). Both chapters bring up issues surrounding the key topics discussed earlier, of the kind of structure in the lexicon and the type of operations used within it. Bresnan espouses a view of language as a system of contrasts, where the meaning and use of a word are determined by competition with the other members of its paradigm (which includes both words and phrases), rather than by the intrinsic features of the word. Fodor on the other hand is concerned with the syntagmatic relations of a word – the hierarchical and compositional properties of words and phrases that mutually determine their semantic content. Johnson and Weinberg are direct commentaries, respectively, on how each of these views relates to parsing and sentence processing – in particular, to current probabilistic proposals. In the final chapter of Part I, Steedman tackles similar issues from a computational perspective, exploring the complex problem of how to integrate a competitive-ranking view of language processing with the structure-building necessary to derive adequate semantic representations.

Bresnan, "The Lexicon in Optimality Theory". In her chapter, Bresnan elaborates on the proposal that the lexicon is not the source but rather the result of syntactic variation. She focusses on markedness facts, providing an account of the lexical forms that surface in instances of variation, specifically explaining the gap in the paradigm of negation in English. In Bresnan's view, words and phrases are elementary units that can compete with each other to be the optimal expression of an underlying form; specific lexical items result from syntactic competition induced by a language-specific ranking of universal constraints.

The proposed account views language as a system of inter-related and competing levels. In particular, Bresnan elaborates a generalised view of paradigmatic competition, within a framework that combines the violable constraint optimization of Optimality Theory with the rich feature specifications of Lexical-Functional Grammar. The classical markedness view of contrasting words is generalised to a competition of elements that are not necessarily lexical items, but also phrases or even larger fragments. For example, periphrastic and lexical forms of negation can compete with each other within a uniform competitive process. All forms are subjected to the very same constraints, whose ranking gives rise to the surfacing of the observed words.

Johnson, "Optimality-theoretic Lexical Functional Grammar". Johnson's paper focusses on the relevance of Bresnan's proposal for processing. The explanation of linguistic universals and markedness in terms of optimisation of well-formedness makes contact with current views of grammaticality and parsing in computational linguistics, brought about by the recent interest in probabilistic language models. Both Optimality Theory and maximum likelihood parsing involve selecting a parse of the input string over an ordinal scale. In this view, grammaticality becomes a relative or comparative notion. Some importants details differ across the two fields – for instance, Optimality Theory does not use continuous scoring functions as in statistical parsing approaches. However, there are also several similarities, especially with recent probabilistic exponential models.

Johnson also notes that Bresnan's general integration of Optimality Theoretic optimization with Lexical-Functional Grammar feature mechanisms may increase the formal complexity of the framework. The view of lexicosyntactic processes as competitive satisfaction of constraints departs from previous Lexical-Functional Grammar theory in a way that might affect its decidability. Specifically, it may not be decidable whether a given string belongs to this kind of grammar formalism, since competitors that are unboundedly different in structure need be compared. This is a topic of on-going research.

Fodor, "The Lexicon and the Laundromat". Fodor's chapter also concerns what constitutes the lexicon. But while Bresnan's paper focuses on the explanation of markedness effects in a paradigm, Fodor's paper concentrates on relations that are purely syntagmatic. The relationship of the words to each other in the lexicon are not of concern here, but rather the relationship of each word to its host – the sentence or phrase it can occur in.

The theoretical decision of what is in the lexicon, and (as importantly) what is not, is guided, according to Fodor, by two necessary principles for lexical well-formedness and interpretability – compositionality and reverse compositionality. The principle of compositionality states that the linguistic structural description of a host is entirely determined by the linguistic structural description of its constituents (plus principles of construction). Reverse compositionality states that the grammar of the constituents is exhausted by what they contribute to their host.

Under this view, lexical entries contain the minimal information that supports compositionality without violating reverse compositionality – that is, lexical entries cannot contain more than what they contribute to the interpretation of the phrases they occur in. Fodor claims as a consequence of this view that frequency information cannot be associated with lexical entries, because it would violate reverse compositionality. Specifically, according to Fodor, since the relative frequency of a host does not depend on the relative frequency of its components, frequency cannot be a lexical property. (We discuss alternative views on this particular assumption in Section 3.2 below.)

Weinberg, "Semantics in the Spin Cycle: Competence and Performance Criteria for the Creation of Lexical Entries". Weinberg's commentary focuses on the apparent mismatch between the constraints imposed by Fodor's reverse compositionality principle, and the recent success of probabilistic (or frequency-based) lexicalised approaches to parsing in computational linguistics and sentence processing. Weinberg reconciles these differing views by noting that reverse compositionality is most appropriately seen as part of the competence theory. Theories that assume a competence/performance distinction do not enforce a one-to-one mapping for representations at those two levels. Thus, lexical representations for processing that extend the competence theory are not precluded provided they are learnable.

To illustrate the necessary competence/performance distinction, Weinberg shows that speakers can keep track of frequencies of properties that are either not distinguished in the competence theory, or are distinguished in a discrete, as opposed to graded, manner. Weinberg also notes that current lexicalist constraint-based approaches to sentence processing are not incompatible with a lexicon organised around compositionality and reverse compositionality. In the lexicalist constraint-based view, lexical entries are dynamically constructed, putting features together that are strongly associated. Highly correlated features (almost always) correspond to lexical entries. This is compatible with a view in which the definitional features are both compositional and reverse compositional, with principles of construction that include performance notions such as frequency or plausibility.

Steedman, "Connectionist and Symbolist Sentence Processing". In his chapter, Steedman notes that competition-based views of processing typically leave unspecified the mechanisms for structure building (see Bresnan's chapter, for instance). Steedman addresses the issue of whether such theories can be integrated with compositional approaches to language and parsing, whose output is a complete structural description, which he assumes to be necessary for semantic interpretation.

Steedman notes that the connectionist models known as simple recurrent networks have been claimed to be models of syntactic parsing; in fact, since their memory fades with time and space, they are effectively equivalent to finite-state devices, such as n-gram part-of-speech taggers. Instead of structured representations, their output is the prediction of the next grammatical element in the input. Such devices are very effective, and can be used to resolve a large amount of ambiguity, in lexicalised, sense disambiguated grammars (see the chapter by Kim et al., described below). However, neither n-gram part-of-speech taggers or simple recurrent networks produce structural interpretations that can support semantic processing; for example, they cannot disambiguate structural ambiguities such as PP attachment.

Steedman proposes that associative memory devices are more promising as a basis for structured lexical representations that support semantic interpretation. If these kinds of devices were used to learn lexical entries, in particular verbs, in a highly lexicalised grammar (such as combinatory categorial grammar, or lexicalised tree-adjoining grammar), then acquisition of these lexical items would carry a large amount of structural information.

2.2 Part II: Division of labour between syntax and the lexicon

The chapters in Part II address the general question of what is the basic architecture of the human sentence processor. The contributions use a range of computational methodologies, such as modelling and corpus analysis, and experimental methodologies, both behavioural and neuro-imaging, often offering complementary insights.

All of the papers focus on the particular issue of how lexical and syntactic information are integrated, or kept apart, in sentence processing. Kim et al., Crocker and Corley, and Stowe address the problem globally, by presenting computational or functional models. Questions that arise in this kind of enterprise are whether the processor manipulates different types of information at different levels, and whether these information types, defined functionally and computationally, belong to different modules of the processor or not. Spivey et al. and Lombardo and Sturt focus on a more restricted question related to the interaction of lexical and syntactic information. In both cases, they focus on a specific property of an existing model, exploring the conceptual and empirical consequences. Spivey et al. ask what kind of evidence is needed to argue for different stages of processing, while Lombardo and Sturt provide the corpus data to quantify the feasibility of a fully incremental lexicalised processor.

On one hand, the proposed models are increasingly sophisticated and make fine-grained predictions. On the other hand, the recent introduction of new conceptual elements (such as studying frequency of usage) or new methodologies (such as neuro-imaging) have not yet given rise to a convergence on a range of models that is more restricted than previously.

Kim, Srinivas, and Trueswell, "A Computational Model of the Grammatical Aspects of Word Recognition as SuperTagging". Kim et al. investigate whether current lexicalist constraint-based approaches can be precisely defined at the grammatical level and implemented on a large scale. They illustrate the implementation of a model of lexico-syntactic processing (Srinivas and Joshi 1999) based on the formalism of lexicalised tree-adjoining grammar, in which the grammar is entirely stored in the lexicon, as a forest of trees (Schabes 1991). The lexicon thus contains explicit calculation of (at least some) syntactic operations, leading to very rich lexical descriptions. By contrast, the grammatical operations for creating structure are minimal. The lexical trees are known as "supertags", by analogy with the simpler category-based part-of-speech tags typically associated with a lexical entry, and assigning the appropriate trees to lexical items during the parsing process is known as "supertagging". In this approach, the balance of work is shifted from purely syntactic combinatory operations to choosing the best lexical tree for a word.

In a representation of this kind, many potential syntactic ambiguities are grounded in a lexical ambiguity – i.e., the choice of lexical tree or supertag. Kim et al. therefore argue for a model in which the lexical disambiguator accomplishes most of the work of syntactic disambiguation. This view of the lexicon/grammar is then encoded in a distributed representation of the supertags, within a connectionist architecture that directly reflects the lexicalist constraint-based approach, with frequencies influencing the likelihood of choice of lexical tree. They proceed then to show that this general purpose implementation presents some of the behavioural effects that have been documented in the psycholinguistic literature, such as the frequency-by-regularity interaction and the bias-by-contextual-cues interaction.

Lombardo and Sturt, "Incrementality and Lexicalism: A Tree-Bank Study". Purely lexicalist theories propose that large amounts of the information necessary to parse a sentence must be prestored together with a given lexical item and activated during parsing. In these kinds of models, as with the proposal by Kim et al., the determination of how structure building exactly works is left under-specified. Lombardo and Sturt explore a crucial computational issue that arises in such an approach: how much non-lexically driven structure building is required in a parsing model that is fully lexicalised, but that also respects one of the basic assumptions of current sentence processing, namely that interpretation is incremental.

They investigate this issue through an analysis of a structurally annotated corpus – that is, a tree-bank. The use of a tree-bank enables them to provide an operational definition of incrementality in terms of structure: incrementality is the requirement of building a fully connected tree at all times. Their results show that 80% of the words can be incrementally integrated into a parse without the use of a headless projection (i.e., non-lexically-based structure), and that very few words require more than one headless projection. Moreover, in all the cases requiring a headless projection, there are systematic patterns related to the current word and the left context that can help in building structure. This result is important because it shows that a large amount, but not all, of incremental syntactic structure building can be accomplished through lexical projection.

Crocker and Corley, "Modular Architectures and Statistical Mechanisms: The Case from Lexical Category Disambiguation". In contrast to the lexicalist investigations above, Crocker and Corley argue in favour of a more traditional modular and pipe-lined architecture, along the lines of several large scale applications in computational linguistics (Brants 1999; Ratnaparkhi 1999). Like Kim et al., they propose a model in which lexical frequencies play a critical role in disambiguation, but in contrast to the supertagging approach, they claim that the distinction between lexical processing and syntactic processing is clearly demarcated. Conceptually, they base their assumption on the observation that having frequency information for the units at each level of grammar increases the amount of information available in each module. (We can remark that this intuition is supported by formal results about the greater power of statistical formal mechanisms, compared to their non-probabilistic counterparts (Cortes and Mohri 2000).) Consequently, they argue, the use of statistics supports a better encapsulation of the modules.

The chapter presents empirical evidence in support of their view. On one hand, they argue that the lexical level of processing makes use of a statistical model, based on data from an experiment showing that lexical statistics are used in disambiguation. On the other hand, they show that the lexical processor does not have access to syntactic information in its statistical calculation, through an experiment revealing that lexical statistical constraints are stronger than syntactic constraints in lexical disambiguation. Together, these results favour a statistical, but modular, language processing architecture.

Stowe, Withaar, Wijers, Broere, and Paans, "Encoding and Storage in Working Memory during Sentence Comprehension". Stowe et al. turn to the methodology of neuro-imaging to address the issue of the division between syntax and the lexicon, by attempting to localize different functions involved in sentence comprehension. Specifically, they claim that the following three language processing functions are associated with distinct areas of the brain: the encoding of lexical information, the storage of lexical and phrasal information in memory, and structural processing. Since theories of sentence processing typically appeal to notions of memory load and processing complexity as the basis for observed behaviour, Stowe et al. argue that their results have important ramifications, as they show that working memory and structural processing can be dissociated. For example, they note that a straightforward interpretation of theories that equate memory load with processing (such as Just and Carpenter 1992; Gibson 1998) cannot be strictly maintained.

In addition, of high relevance to the ideas in this part of the volume, we think that the evidence in the chapter by Stowe et al. challenges the degree to which (and the manner in which) syntax and the lexicon can be unified in models of human sentence processing. While they show that lexical and phrasal memory are not distinct – drawing on the same resources, in the same area of the brain – they also reveal distinctions between lexical encoding and memory on one hand, and structural processing on the other. These results indicate that proposals in which lexical and syntactic processing are the same (e.g., MacDonald, Pearlmutter, and Seidenberg 1994; Kim et al., this volume) may be oversimplified when taken at face value. The Kim et al. chapter downplays the degree of processing involved in attachment ambiguities, but Stowe et al.'s neuro-imaging results suggest that the structuring of phrases, rather than just the encoding and storage of lexical information, is a central aspect of sentence processing.

Spivey, Fitneva, Tabor, and Ajmani, "The Time-Course of Information Integration in Sentence Processing". The papers by Crocker and Corley and Stowe et al. bring new kinds of evidence in favour of a modular organisation of processing, on functional or computational grounds. Spivey et al., on the other hand, question the very notion of stages of processing. Moreover, they present human and computational experimental results supporting the view that previous evidence taken to indicate stages of processing can be equally well explained within an interactive, non-modular constraint-based system.

Specifically, Spivey et al. compare the results and conclusions of McElree and Griffith (1995, 1998) to the conclusions that can be drawn on the basis of a dynamical model of parsing. McElree and Griffith had argued that subcategorization (as syntax) and thematic role information (as lexical semantics) come into play at different times in sentence processing. Spivey et al. re-interpret the results of the experiments to show that a different explanation can be found within Spivey's normalized recurrence model (Spivey-Knowlton 1996). In their computational system, all the factors influencing ambiguity resolution are lexicalized constraints, which, regardless of information content, are simultaneously activated at the moment of lexical access. However, the influence of various factors is felt at different points in time, due to differential weightings of the types of information. Thus, constraints that are simultaneously available may still have a "staged" influence on behaviour.

Spivey et al. conclude that it is important to distinguish differences in timing versus differences in strength of information sources, in both computational and human experiments, in order to differentiate current theories of human sentence processing.

2.3 Part III: Details of lexical entries

Part III of the volume contains papers that elaborate, through human experimental studies as well as corpus analysis, details concerning the information that is stored within a lexical entry. Whereas papers in Part II focus on a larger grain of analysis of a lexical entry – how "syntactic" is it? – here the emphasis is on finer-grained details of individual pieces of the stored information.

All of the papers focus particularly on argument structure properties of verbs – their representation and role in processing – reflecting the importance of verbs in guiding language understanding, both in computational linguistics and in theories of sentence processing. A theme that runs through the papers is the need to elaborate more clearly what constitutes the representation of argument structure for a verb, and how this representation relates to frequency biases that influence human behaviour. Some of the questions being raised are the following: Is the information conceptual in nature (i.e., real-world knowledge) or a set of more circumscribed formal properties (i.e., a more "linguistic" view)? How fine-grained is the representation of differences in argument structure across verbs? What is the relation between verb sense, verb argument structure, and frequency biases for verb/argument relations?

Three of the papers (Mauner et al., Filip et al., and Altmann) focus on determining the precise nature of the information that influences sentence processing, using psycholinguistic experiments to build evidence for the early use of more finely grained thematic and/or semantic information. The corpus-based work of Argaman and Pearlmutter, and Roland and Jurafsky addresses a somewhat more general issue of the appropriate level of representation of frequency biases – focusing less on exactly how frequencies influence human processing and more on which aspects of lexical information they are stored with.

Mauner, Koenig, Melinger, and Bienvenue, "The Lexical Source of Unexpressed Participants and their Role in Sentence and Discourse Understanding".

The main idea investigated by Mauner et al. is that, as part of the argument structure of a verb, implicit (i.e., unexpressed) arguments are made available to and used by the human sentence processor immediately at the verb. They show that, by contrast, this is not the case for entities that are merely implied by general conceptual knowledge. They conclude that the participants in an action that are *linguistically* licensed can influence processing earlier than those that are only inferrable.

This proposal is supported by experiments on two structures that have the same logical necessity of an Agent, but have different linguistic representations – a short passive in which the unexpressed Agent is nonetheless part of the argument structure of the verb, and an intransitive in which an unexpressed Agent is implied by the semantics of the situation, but is not part of the representation of the verb. Evidence from eye-tracking experiments reveals differences in processing between conditions with linguistically and conceptually derived implicit agents as early as at the verb, and in first pass reading times.

The authors further argue that their results indicate that thematic roles in processing must be finer-grained than the traditional labels such as Agent or Patient commonly adopted in linguistic theory. For example, they compare volitional and non-volitional agents of causation and show that "volitionality" influences the availability of an implicit agent in processing (precisely, it affects the suitability of an implicit agent as an antecedent for a volitional anaphor).

Filip, Tanenhaus, Carlson, Allopenna, and Blatt, "Reduced Relatives Judged Hard Require Constraint-Based Analyses". Filip et al. develop the role of fine-grained thematic relations in sentence processing even further, arguing that the linguistic differences between verbs in different lexical semantic classes can be captured with semantic features based on Dowty's (1991) featural decomposition of Proto-Agent and Proto-Patient roles. Using a questionnaire study, they verify Stevenson and Merlo's (1997) observation that class-based distinctions between verbs arise in processing, but argue for modelling these results with verb frequency values and thematic fit biases. Filip et al. further note that a fine-grained analysis of thematic roles (in terms of Dowty's linguistic properties) is key to a full understanding of the relation between verb class and behaviour.

A key novel observation in their work is that the difficulty of processing a reduced relative construction – such as *The horse raced past the barn* in *The horse raced past the barn fell* – is influenced by the main verb (i.e., *fell* in the example). They account for this effect in terms of the compatibility between the two sets of Proto role features assigned to the subject of the sentence by the two verbs – the verb in the reduced relative (*raced*) and the main verb (*fell*).

They suggest that on-line processing of this type of construction is facilitated when the two verbs assign compatible features to the initial noun phrase, and made more difficult when the two verbs assign incompatible features.

The Filip et al. proposal draws on converging evidence from linguistics, psycholinguistics, and computational modelling, by embedding Dowty's Proto features within a constraint-based linguistic theory (HPSG), and modelling the ambiguity resolution data using Spivey's normalized recurrence algorithm (see Spivey et al., above).

Altmann, "Predicting Thematic Role Assignments in Context". Altmann also develops a proposal involving finer-grained semantic information to define thematic relations, and an emphasis on probabilistic use of this information. In contrast to Filip et al., however, his proposal for thematic roles is that they are verb specific selectional restrictions based on real world knowledge, thus moving even further from the standard linguistic theoretic notion of thematic roles. His primary claim is that discourse context establishes probabilistic relationships that serve as predictions concerning the entities which will play a role in the predicate of a subsequent verb. That is, the sentence processor *at the verb* (and therefore, in English, before the object position) tries to assign the verb's object thematic role to an entity already in the discourse. In support of this, Altmann finds that people experience an anomaly at a verb such as "injured" when it follows a context in which nothing is "injurable".

Altmann offers two hypotheses concerning how the sentence processor encodes "probabilistic contingencies" that represent relations particular to a lexical item. One suggestion is that the processor projects empty structure at the verb corresponding to its object thematic role, and then attempts to link this, following the attendant thematic (selectional) restrictions, with a prior entity in the discourse. Under this view, anomaly detection occurs when the processor attempts to anaphorically link the empty object position to a prior entity and finds no suitable antecedent. An alternative hypothesis is that the prior nouns in the discourse restrict the range of expected verbs to those which can fill a role with one of the entities. Anomaly detection in this scenario occurs when earlier entities have activated possible verbs of which they could be arguments and the current verb is not one of them. Altmann discusses the implications of the experimental data and these possible processing explanations within a connectionist framework for sentence processing.

Argaman and Pearlmutter, "Lexical Semantics as a Basis for Argument Structure Frequency Biases". In their chapter, Argaman and Pearlmutter shift the focus from what is the precise nature of argument structure relations in a verb's lexical entry, to what determines the argument structure frequency biases. Following Pinker (1989) and Levin (1993), they assume that primitive meaning components licence particular argument-taking properties of predicates. Since, under this view, argument structures map to "partial semantic representations", they note that argument structure selection is essentially sense disambiguation. Then, by analogy with frequency effects in word sense disambiguation, argument structure disambiguation involves frequency biases that are associated with the lexical semantic meaning components which license the argument structures. Just as meaning frequencies for ambiguous words are determined by properties of the world (the frequency of the things they refer to), so too are argument structure frequencies under Argaman and Pearlmutter's view. They claim then that argument structure biases can be independently determined through a theory of real world semantics.

The specific hypothesis investigated in the chapter is that words closely related in meaning will have similar argument structure frequencies. To determine this, Argaman and Pearlmutter compare a set of verbs and their derived nouns, and find a highly correlated bias between the two for a particular complement type (the sentential complement), in both corpus and completion studies. Conversely, a comparison of the sentential complement bias across a group of Levin verb classes reveals a significant difference. The question remains whether finer-grained differences (within-class semantic distinctions) also influence frequency biases. They found partial support for this hypothesis in the form of marginally significant correlations between verbs and their derived nouns within a class.

Argaman and Pearlmutter conclude that, while preliminary, their results provide clear evidence for a connection between frequency biases of argument structures and their underlying semantic basis.

Roland and Jurafsky, "Verb Sense and Verb Subcategorization Probabilities". Roland and Jurafsky set out to explore the issue of frequency bias for lexically determined structural information, such as subcategorization, from a different and complementary point of view to that of Argaman and Pearlmutter. Specifically, they explore the extra-linguistic factors that underlie variations in subcategorization frequencies, in an attempt to distinguish differences attributable to the sense of a verb, from differences due to the modality or style of usage.

Using a comparative method of corpus analysis, they examine the differential frequencies of subcategorization frames of selected verbs across modalities, in written and oral corpora. They further compare elicited to spontaneous production for written corpora, and also balanced corpora to simple text collections (hence unbalanced). Interestingly, they notice that while the difference between experimental data (elicited language) and corpora (spontaneous language) is large, the difference between corpora in the same modality, whether balanced or not, is not significant, once controlled for sense.

By factoring out extra-linguistics properties, Roland and Jurafsky confirm the hypothesis that subcategorization probabilities vary according to word senses, once the context effects have been taken into account. This conclusion converges with that of Argaman and Pearlmutter. Together, these two papers provide evidence that the sense of a word is the grain at which frequency is expressed, and that therefore the underlying events that give rise to frequency differentials must be found in lexical semantic primitives.

3. Discussion of cross-cutting issues

Although we have divided the contributions into sections of the book that emphasize the primary focus of the individual chapters, many of the papers address issues across these boundaries, and there are numerous points of contact between papers in different sections. Here we highlight three major topics that run through the papers, which capture the primary issues in current lexicalized theories of language and language processing: the organizational basis of the lexicon, the role of statistical information in lexical entries and processing, and the impact of lexicalization on incremental processing algorithms.

3.1 Lexical organisation

A number of issues cut across the chapters concerning the fundamental nature of the lexicon: how it is organized and how that organization influences processing. Here we first discuss the nature and role of lexical classes, and their basis in semantics. Then we turn to the timing of structural versus semantic information in on-line interpretation, and how they interact. Finally, we address the basic character of the lexicon itself – as a static set of pre-stored tree structures versus a dynamically generated set of structures created from simple universal primitives.

3.1.1 *The role of lexical classes*

The lexicon is not simply a list of irregularities. Each word in the lexicon is the bearer of unique, idiosyncratic information, but also of information similar to that contained in the lexical entries of many other words. If the regularities and underlying organisation of the lexicon are not taken into account, lexicalist approaches are prone to be redundant. To avoid such redundancies, lexical theories have relied on the notion of a lexical class as a means for capturing regularities. In the current volume this issue is reflected in several chapters that address the problems of determining precisely how (or even whether) classes are defined, and what role they serve in processing.

Filip et al. focus on the issue of whether verb classes are organized on the basis of structural or semantic properties. Not only do they claim that the structural distinctions between classes proposed by Stevenson and Merlo (1997) are unnecessary, they further argue that a categorical view of argument structure representations is insufficient to fully account for data on ambiguity resolution. Thus, they elaborate a view of verb class distinctions as semantic, not syntactic, and as graded, not categorical. Spivey et al. reinforce this latter point, in stressing that behaviour that appears categorical may result from the interaction of continuously weighted constraints. From this perspective, verb classes are (possibly overlapping) fuzzy sets.

Like Filip et al., Argaman and Pearlmutter also view verb classes as semantically defined, but assume a more discrete view of classes. Specifically, they adopt the approach of Pinker and Levin, in which verbs in a class share primitive meaning components that license particular argument-taking properties. Since, under Argaman and Pearlmutter's proposal, these meaning primitives serve as a site for frequency counts, the class of a verb plays a direct and observable role in processing phenomena. Argaman and Pearlmutter also put forth some preliminary evidence suggesting that frequency differentials arise from within-class meaning distinctions as well. Thus, under their view, semantic regularities lead to coarse-grained commonalities in behaviour across verbs within a class, while finer-grained distinctions in meaning lead to correspondingly finer-grained differences in behaviour among those verbs.

3.1.2 *Timing of different information types*

In addition to the issue of which type of lexical information (syntactic or semantic) underlies particular processing effects, there is also the issue of the relative timing of the information in on-line interpretation. In contrast to previous sentence processing models, lexicalist approaches raise the possibility of having both syntactic and semantic information arise from a common source, with both playing an immediate role in processing. This view is different from more traditional structure-based models, in which syntactic information is maintained separately from semantics, and comes into play much faster. The lexicalist view is also different from interactive, staged models in which syntactic and semantic information can influence each other, but the separation of different types of information into distinct levels is clear.

With respect to this issue, Filip et al. and Spivey et al. take an approach similar to a number of lexicalist constraint-based theories, in which both types of information are available from the outset. The general approach accounts for effects in ambiguity resolution with separate but interacting frequencies corresponding to either syntactic or semantic information.

Two other chapters attempt a finer grained elucidation of the relationship between linguistic and real world information, and appear to come to contradictory conclusions. Mauner et al. claim to find an earlier influence of linguistic information (knowledge about argument structure), compared to the later influence of general semantic plausibility. Altmann, on the other hand, finds that real world constraints come into play immediately. However, the two approaches are probing somewhat different aspects of the use of fine-grained lexical information. Mauner et al. test for the immediate use of linguistically-specified arguments, as opposed to plausible (non-argument) participants. Altmann, by contrast, tests whether plausibility plays an immediate role when processing a linguistically-specified argument, and finds that it does so. In combining these two insights, it seems that a more complex relationship must be elaborated, in which real world knowledge may have an early influence if it is closely associated with a linguistic argument (as is the case with selectional restrictions).

Argaman and Pearlmutter, from a very different perspective, provide additional evidence that linguistic and real world information are closely linked through the relation between semantics and argument structure. Thus, it may be the case that syntactic and semantic information not only arise from a common source (a lexical item or its class), but are tightly connected to each other in a way that directly influences processing.

What all of these chapters demonstrate more generally is that lexicalist theories of processing may involve very fine-grained interactions of different levels of information, requiring more sophisticated experimental methodologies and analyses to elucidate them. Spivey et al. point to one particular methodological difficulty, that of using claims about timing to determine the primacy of some particular organizational principle, such as structural information. Specifically, they propose that observed differences between syntactic and semantic factors in processing depend on their initial and accrued strength, and not on a distinction in availability. Thus, evidence for differences in staging of information during processing must be carefully examined, and new methods developed for distinguishing between availability of information and its strength.

3.1.3 *Static vs. dynamic organization of the lexicon*

Under a lexicalist constraint-based account, much observable behaviour is viewed as the result of competition among constraints of varying strengths during on-line processing. Recently, lexical theories in theoretical linguistics have extended the role of competition to determining not only on-line behaviour, but the content of the lexicon itself. What we think of as the lexicon of words or morphemes for a particular language may not be a pre-existing database of linguistic facts (whether organized by classes or not), but rather the result of a competitive process over universal primitives.

This view is espoused in Bresnan's chapter, in which the lexicon is proposed to be the result of syntactic competition. In her Optimality Theory account, words can compete with entire phrases and structures to be the optimal expression of an underlying input, so that words and phrases lie within the same level of representation and are subject to the very same (syntactic) operations. At least at the level of the competence theory, it is not only syntactic structure that is generated when a sentence is formed, but the lexical entries themselves as well. The pre-existing lexicon consists of a universal set of primitives, from which the competitive process for an input selects the optimal combination for a particular language (Prince and Smolensky 1993).

At first blush, this contrasts starkly with the view in sentence processing, where the notion of competition as central to interpretation has led to more elaborated structures in the lexicon (MacDonald, Pearlmutter, and Seidenberg 1994 and others). The paper by Kim et al. exemplifies this view, in which lexical competition occurs between syntactic trees within the lexicon which are activated in response to an input. In other words, in sentence processing, the notion of competition has led to extensive pre-compiled structure in the lexicon, while in theoretical linguistics, it has conversely led to less pre-existing structure – to the point where even words or morphemes result from syntactic competition, rather than being the input to it.

The differences between these views may not be as great as first appears, however. In his chapter, Johnson points out the modifications that would have to be applied to Bresnan's competence theory to render it computable (and therefore able to form the basis of a model of sentence processing). The GEN function in Optimality Theory raises similar issues to those introduced by other theories that presuppose an infinite generating function: a computational algorithm depends on at least some of the filtering constraints being interleaved with the generation of possible alternatives (compare Fong 1991; Tesar 1996). Johnson proposes to apply the phonological string as the first filter, to guarantee that the computation will be possible (see also Johnson 1989). Perhaps the difference then between the lexicalist sentence processing view and the view proposed by Bresnan is not so much one of kind as of degree – it remains for future work to determine how much information can, and must, be precompiled into lexical entries.

3.2 Frequencies and statistics

An emphasis on lexical influences and on frequency effects have gone hand-in-hand in sentence processing research. Frequency is thought to be a prototypical example of lexical information due to standard word-based frequency effects. Furthermore, the apparent ease of associating frequencies with pre-stored information in the lexicon, rather than with syntactic constructions, argues for a lexicalist view of frequency effects. An emphasis on frequency has also led to a corresponding emphasis on semantics, as we saw in the previous discussion, since frequency provides an observable encoding of the exposure to the external world and its influence on the sentence processing mechanism.

Before turning to the issues involving the precise specification of lexical frequencies, their impact on processing, and their origin, we first discuss the possibility that frequencies are not an integral part of the lexicon.

3.2.1 *Are there frequencies in the lexicon?*

One possible view is that frequencies (of any granularity) simply do not occur in the lexicon. In his chapter, Fodor claims that such is the case, based on the assumption that lexical entries project the entirety of their content to their host. According to Fodor, then, frequency cannot be in the lexicon, as this would require that the relative frequency of the host be determined by the relative frequency of its parts, an assumption Fodor disavows.

However, probability models in computational linguistics hold precisely this assumption: any complex event is assumed to be decomposable into smaller, independent events, and the probability of the complex event is the product of the probability of the independent subevents. The independence assumptions are acknowledged to be too strong, and violated in practice, but the response is to develop more sophisticated and accurate probability models, rather than to abandon the approach. Contingent frequencies may be seen as rules of composition over the raw frequencies stored with an individual lexical item; what is required is to determine the appropriate combination algorithm.

Another view on this issue, reflected in sentence processing work, is that contingent frequencies are strengths of association *between* lexical items – i.e., contingent frequencies are not part of a lexical entry, but rather are part of the organization of the lexicon. From both of these perspectives, it is clear that Fodor's assumption that reverse compositionality cannot apply to frequencies is not a given, but rather highlights a known challenge of determining an appropriate representation and processing algorithm for lexical frequencies.

Weinberg replies to Fodor's arguments in a different vein, calling on the competence/performance distinction. Weinberg notes that frequencies do not need to be part of the representation of a lexical entry to be relevant to parsing, rather they can be part of the performance system. As such, the principles constraining lexical entries proposed by Fodor do not rule out current lexicalist theories of processing. However, Weinberg proposes a view that is not what many lexicalist proponents would endorse, we think. The distinction between competence and performance is not very clear in those approaches that are crucially based on a continuous as opposed to discrete representation. Hence, lexicalist approaches generally appear to say that frequency is part of the lexical entry. If a representation is distributed, the strength of the association of certain features is crucial to the representation itself.

If we accept that frequencies are associated to lexical items at some level of representation, then several issues must be addressed when studying the influence of frequency on processing. The primary one seems to be what the frequencies are associated with – i.e., what level and type of information carries frequency information (sense of a word, lemmas, phrases, constructions, among the many possible candidates). It is also important to study how to determine which frequency information comes into play at different points in processing, and where frequency differentials come from.

3.2.2 *What do we count?*

Roland and Jurafsky address the first question above of determining the lexical unit with which frequencies are associated, and they argue convincingly that the indexing unit is the individual word sense. Argaman and Pearlmutter, from a very different perspective, reinforce this view with their evidence that frequencies are associated with argument structures, which are themselves linked to semantic primitives. This result raises a practical and a theoretical problem. Practically, many experiments and data collections on lexical frequency are not based on the sense of the word, but on the lexical string, thus potentially confounding frequencies of very different word meanings and uses. Theoretically, the senses of a word are often not clearly defined; they may not even be enumerable, but rather the result of the interaction with the sentence context (Pustejovsky 1995).

If we combine these two observations to their logical conclusions, we might envisage a picture of lexical frequency as being the result of a process of syntactic analysis. If a word sense is determined by a compositional operation, and lexical frequencies are associated with senses, then lexical frequencies are the result of a compositional process. This view is similar to the one espoused by Bresnan, discussed above, in which the lexicon itself is the result of a compositional process. It also impacts on the discussion of Fodor's claim above, reinforcing the view that it is the combination of frequencies that must be addressed in a model of interpretation.

3.2.3 *Frequencies in processing*

If lexical frequencies are sensitive to syntactic structure and, more generally, the influence of context, then one might envision that several types of lexical frequencies are associated with the same lexical item. Moreover, the various types of frequencies may refer to increasingly larger domains or classes of information within a lexical hierarchy. Questions then arise concerning how these different levels of lexical frequencies are used in the time course of processing. When several levels of statistics come into play, it must be determined whether they do so all at the same time, or according to some predetermined ordering procedure. The first method is envisaged by proponents of distributed representations and architectures in psycholinguistics, in which interacting syntactic and semantic constraints, weighted by frequency, simultaneously determine the overall activation of an interpretation. In computational linguistics, this simultaneous, non-linear combination of frequencies is only rarely used (Henderson 2000), and simpler techniques based on linear interpolation are more current. Backing-off techniques are also used, which impose explicit ordering on the use of different grains of frequencies. These techniques have been developed to handle the problems of sparse data by supplementing the core fine-grained probabilities with coarser grained ones that play a secondary role (Katz 1987; Collins and Brooks 1995; Jelinek 1997).

Several positions concerning the timing of frequency information are illustrated in the current volume. Kim et al. propose a model where several levels of frequency are activated at the same time, and the appropriate level of specificity is found automatically. Crocker and Corley argue against this view and propose a more traditional architecture where frequency information is exclusively related to lexical items and encapsulated in a lexical preprocessor. Argaman and Pearlmutter argue that frequency is a property of partial semantic components that make up the meaning of a word and that hold across boundaries of syntactic categories. However, while Argaman and Pearlmutter assert that processing behaviour is correlated with argument structure frequencies, they leave open the precise role and timing of such information. Specifically they note that it isn't yet known whether it is stored frequencies that influence sentence processing, or the underlying semantic representations themselves that are directly at work on-line.

3.2.4 *The origin of frequencies*

By grounding frequency in real world semantics, the proposal by Argaman and Pearlmutter also addresses the problem of the origin of frequencies. They delimit the space of possibilities as including non-causal and causal explanations. In one case, frequency is an accident, a random variation that has reached larger proportions over time (as in Tabor 1995). Argaman and Pearlmutter adopt a different view, in which frequency is the effect of an underlying cause that accounts for similarities and differences in individual frequencies. According to Argaman and Pearlmutter, this underlying cause is the salience of objects in the world that the words refer to.

A connection between real-world salience and frequency is not a new idea for the explanation of word sense frequencies (e.g., the use of the word *bank* as an institution is more frequent than *bank* as the edge of a river because the former are more commonly talked about today). What is novel in the Argaman and Pearlmutter proposal is that frequencies of more structural notions such as argument structure also directly reflect differences in the world. This is an interesting proposition, that connects some features of the world, or our knowledge of the world, directly to our linguistic behaviour. Given the connection between argument structure and syntax, the position is rather radical.

A more indirect relation is usually assumed in order to explain a certain arbitrariness in the lexicalisation and grammaticalisation of real world knowledge. In particular, a relation between (low) complexity on the one hand, and (high) frequency and (wide) typological distribution on the other, is a widely attested phenomenon, captured by the notion of linguistic *markedness* (Moravcsik and Wirth 1983). An instantiation, and in part an explanation, for the relationship betwen complexity, frequency, and cross-linguistic variation is illustrated in Bresnan's paper. In her account, more complex structures violate more grammatical constraints, and therefore surface less frequently. Thus Bresnan's account includes an intra-linguistic component to frequency differentials, which are not exclusively a function of salience of referents in the external world.

3.3 Incrementality

The increasing emphasis on lexical information brings a corresponding emphasis on a-word-at-a-time processing. In sentence processing, this is due to the observed rapidity of interpretation (which in standard views requires a connected parse); in computational linguistics, this is due to the need for efficiency and support for semantic interpretation (Charniak 1997; Roark and Johnson 1999; Brants and Crocker 2000; Sturt, Lombardo, Costa, and Frasconi 2001). The chapters here investigate the degree to which incremental word-based processing is possible and conducive to interpretation, and the effects on incremental processing of having enriched lexical entries.

3.3.1 *The limits of incrementality*

Lombardo and Sturt show that while a large amount of lexically projected syntactic structure-building can be performed incrementally, some attachments cannot be resolved in a fully lexically projected and incremental approach. In this context, Steedman's implicit reminder of the difference between lexical semantics and sentential semantics help situate the claims of lexicalist approaches with respect to parsing proper. Steedman's remarks apply both to simple recurrent network architectures and to recent lexicalist approaches to parsing (Kim et al., this volume). Any approach equivalent to part-of-speech tagging will leave some structural attachments undone, and is therefore not supportive of full semantic interpretation, for which a fully connected structure is required. Note too the empirical support from Stowe et al. that suggests the existence of a functional area of the brain for structure-building as opposed to lexical processing.

These points indicate that structure building beyond lexicalist projection must be accomplished if one is to build a complete interpretation. However, recent research, both in psycholinguistics and computational linguistics, has called into question the completeness assumption in parsing. Work on reanalysis of garden-path sentences has shown that comprehenders often end up with an interpretation that is based on both the initially incorrect and ultimately correct structures (Christianson, Hollingworth, Halliwell, and Ferreira 2001). These results are interpreted as suggesting that alternative syntactic analyses may not be completely constructed; rather, the reanalysis mechanism may be satisfied with a "good enough" parse. A different way of relaxing the requirements for a full parse is exemplified in the paper by Kim et al. in this volume, and similar approaches have been suggested in the computational parsing literature (Abney 1996). Here a large amount of lexical projection is performed, which gives rise to a partial parse, where fragments of the entire structure are constructed. Some difficult structure building decisions are left undone (PP attachment is a typical case), under the assumption that they will be resolved at later stages by knowledge based on pragmatics, or lexical associations. Both these lines of proposals raise interesting questions regarding exactly what constitutes "an interpretation," and the precise interaction required between lexically-driven and discourse-driven processes.

3.3.2 *The influence of rich lexical information*

The chapters by Mauner et al. and Altmann expand the view of incremental interpretation, by proposing rich lexical information that licenses early postulation of hypothesized entities. These empty elements play a central role in on-line interpretation. Mauner et al. provide evidence that linguistically-licensed entities (e.g., arguments to a verb) influence interpretation even when those entities are not expressed in the sentence. Altmann suggests further that real-world properties of such arguments can affect subsequent integration of words in the input. Both of these proposals go far beyond the early establishment of an empty element in the syntactic representation of an input, as proposed in previous models of sentence processing (e.g., Crocker 1995; Gibson and Hickok 1993; Stevenson 1993; Stevenson 1995). In the Mauner et al. view, fine-grained thematic properties of empty discourse elements play an early role in interpretation. According to Altmann, this role extends to the incremental computation of expectations concerning the real-world properties of entities and events in the input. In both cases, the representation of empty elements is more sophisticated than previously assumed, and thus their role in incremental processing is potentially more complex and influential.

4. Methodological concerns

Many of the chapters use computational methods to investigate psycholinguistic questions, either by modelling experimental results or by investigating corpus data. Both types of approaches raise important methodological issues in the study of human language processing, and increase the connections between work in psycholinguistics and computational linguistics. Other chapters take a more traditional human experimental approach, but expand the repertoire of experimental methodologies in order to address the novel questions raised by a lexicalist view. We discuss the import of each of these methodological insights here.

4.1 Computational modelling

Building a computational model is a complex business, as computer programs can be interpreted at two levels of abstraction, usually referred to as the representational level and the algorithmic level (Marr 1982; see also the discussion in Brent 1996). These different perspectives are concerned with what is computed and how it is computed, respectively. For a particular set of observational data from human experiments, there could be many underlying representational schemes, each of which has many possible ways of being computed. The consequence is that experimental data generally underspecify the set of possible computational models. The chapter by Spivey et al. addresses this issue in detail, with a concrete demonstration that certain experimental evidence can be explained equally well by two different kinds of models founded on very different representational schemes and algorithms.

The under-specification of models by the data leads to a range of approaches to computational modeling illustrated in the chapters here, which lie along a gradient of specificity, or degree of abstractness. Starting from a full implementation, one can propose a very detailed model, such as the "almost" parsing model of Kim et al. By assuming a specific computational architecture, this approach provides theoretical justification for hypotheses for which the empirical evidence is insufficient. One can also propose models more limited in scope, such as that of Crocker and Corley. They develop a detailed model of disambiguation, situated in a well-understood, although not implemented, parsing framework. An even greater degree of abstraction can also profitably be adopted, as seen in the chapters by Filip et al. and Spivey et al. Each of these papers uses a model with a completely underspecified parsing framework, in order to highlight very specific aspects of the competition process, and the influential constraints, in disambiguation.

4.2 Corpus-based investigations

The availability of large collections of annotated (part-of-speech-tagged or parsed) text has recently introduced a new opportunity to explore even more abstract computational models. The paper by Lombardo and Sturt exemplifies this methodology, by developing a very abstract model of "possible parsers," founded on representations derived from corpora. Corpora also support the investigation of models of lexical information without any explicit relation to parsing, as seen in the papers by Roland and Jurafsky, and Argaman and Pearlmutter.

More generally, these papers underscore the fact that annotated corpora are implicit repositories of grammars (Merlo and Stevenson 1998). In this regard, corpora go beyond idealized grammatical knowledge, and serve as an approximation to a speaker's linguistic experience, containing important frequency information. It has become common practice in psycholinguistic approaches, illustrated in numerous chapters here (e.g., Argaman and Pearlmutter, Filip et al., Kim et al., Spivey et al.), to use frequency data collected from corpora as representative of a speaker's knowledge. More helpful still will be the sophisticated lexicalized grammars currently being developed by computational linguists through automatic extraction from parsed corpora (e.g., Xia, Palmer, and Joshi 2000). The result of such efforts would be grammars that incorporate statistics over usage, providing the integrated grammatical and statistical knowledge needed to evaluate sentence processing proposals that rely on lexicalized frequency effects.

It is important to note, though, that as an approximation to an actual linguistic experience, data from a corpus must be confirmed through analysis and comparison to actual behavioural and linguistic studies (Gibson and Pearlmutter 1994; Merlo 1994; Roland and Jurafsky 1998; Lapata, Keller, and Schulte im Walde 2001). The chapter here by Roland and Jurafsky shows that such studies can both evaluate properties of experimental stimuli, and lead to preliminary hypotheses which can then be tested experimentally. From a practical point of view, the paper by Roland and Jurafsky provides very useful evidence on the materials and methods needed to estimate subcategorization and argument structure frequencies, a crucial problem in developing sophisticated probability models in both psycholinguistics and computational linguistics.

4.3 Experimental advances

An emphasis on lexical information has similarly led to innovations in experimental methodologies as well, illustrated in several of the chapters here. For example, motivated by fine-grained predictions from a lexicalist constraint-based perspective, Spivey et al. introduce a very interesting new methodology, called speeded sentence completion. This technique allows them to access representations at different points in on-line processing, by eliciting completions of a sentence fragment after differing time delays. This type of data is needed to support or disprove the claims that multiple interpretations are simultaneously, but differentially, activated, as factors of differing weights compete over time.

Both the Mauner et al. and Altmann chapters extend experimental approaches to determine the role of lexically-specified information about arguments that are not (or not yet) explicitly present in the input. Both add to the repertoire of experimental methods for eliciting information about the early use of argument structure in sentence processing. Mauner et al. detail methods for detecting elements of argument structure (the semantic arguments of a verb) independently of subcategorization (the syntactic expression of the arguments). Their techniques are also independent of the plausibility of the argument, allowing for arguments to be detected at the verb itself. Altmann also seeks to elucidate the immediate role of argument expectations in processing, using a method of anomaly detection at the verb to elicit responses when no previously introduced entities are compatible with the selectional restrictions on its object (what he terms the object's "thematic role"). These selectional restrictions involve general semantic "fit" with the verb, and thus go beyond the purely linguistic information suggested by Mauner et al.

The chapter by Stowe et al. demonstrates the need for additional neuro-imaging data and techniques to help constrain possible models of sentence processing. Their results indicate a complex relationship between functional areas of the brain, and the division of sentence processing labor into encoding, storage and processing. Approaches that equate lexical and syntactic (phrase-level) processing, a common assumption in lexicalist theories, initially appear incompatible with the evidence from Stowe et al. As we have noted at several key points of discussion above, a lexicalist approach raises many new interesting questions concerning the precise representation of different information types, and the nature of their interaction. Advances in neuro-imaging studies will ultimately be required to elicit the fine-grained data needed to distinguish among the logical possibilities those that are compatible with the functional architecture of the brain.

5. Conclusions

The contributions to this volume illustrate the wide range of issues that arise as a consequence of the increased role that the lexicon plays in current theories of syntax, of parsing and of sentence processing. The findings here reveal the complexity of both representations and algorithms required in theories of language that capture the richness of lexical information and its interaction with syntax and structure-building. In dealing with this complexity, researchers face a tension between the descriptive need to represent the full range of lexical variability in language, and the conceptual need to explain the regularities and organisation of the lexicon. We think that the accumulated evidence that lexical effects are strong ("it is all in the words") can be reconciled with the theoretical needs for generalisation and succinctness by further exploring the notion of *classes* of words. The investigation of the notion of class promises to be informative to some of the common concerns that have appeared across many of the papers in the volume.

A class structure for lexical items implicitly assumes that the lexicon is organised, since it imposes regularity on lexical variability. Studying the potential principles that underlie lexical organisation is important practically, as a means of reducing redundancy and rendering lexicalised approaches manageable, and is also important conceptually, as it highlights regularities and generalisations (Daelemans, De Smedt, and Gazdar 1992; Briscoe, Copestake, and de Paiva 1994). Recent proposals in computational linguistics for automatic verb classification have investigated classifications based on both syntactic and semantic information, such as subcategorisation (Xia, Palmer, and Joshi 2000), argument structure (Merlo and Stevenson 2001), Levin's classes (Lapata and Brew 1999), and finer-grained classes than Levin's (Dang, Kipper, Palmer, and Rosenzweig 1998). A possibility unifying these approaches is that these different types of classes correspond to different levels in a hierarchical lexicon, which simultaneously captures generalizations at different levels of abstraction (Palmer 2000; Merlo and Stevenson 2001).

The notion of lexical class further provides the conceptual locus to integrate symbolic linguistic notions and probabilistic concepts. For example, the recent work in computational linguistics on the verbal lexicon has shown that there are pervasive regularities in statistical distributions of verbs belonging to the same semantic class. These statistical regularities are attested at several levels of granularity of lexical organisation (Lapata and Brew 1999; Merlo and Stevenson 2001). These findings lend further support to the idea that the lexicon is hierarchically organised along several levels at the same time, and extends this view by suggesting an organisation that is sensitive to frequency. This type of rich lexical organization, in terms of a frequency-informed hierarchy, supports the sophisticated probabilistic modeling techniques (using back-off and smoothing) that have been so useful in computational linguistics.

Many of the profitable definitions of classes in computational linguistics have been based on structural notions, such as subcategorisation, alternations in the expression of arguments, and argument structure. If words are systematically grouped in classes organised around structural notions, then the integration of each word in the sentence representation during parsing requires the integration of a little piece of structure. Thus, the notion of class is directly relevant to issues of incrementality in parsing, concerning what kind of information is immediately available, and the projection or prediction of such information. Furthermore, a structure-based notion of classes has the consequence that the relation between structure and frequency can be productively studied, with each investigated as the predictor of sentence processing complexity.

Finally, the systematic relationship between class and frequency can support the integration of research on processing and acquisition. Recent work on the automatic acquisition of properties of verbs has shown that even in cases where the surface syntactic representation (subcategoristion) does not distinguish between classes, the statistical differentials related to verb class properties are strong enough to generalise and enable the semantic classification of previously unseen verbs (Merlo and Stevenson 2001). If words are systematically grouped into classes organised around structural notions, then learning words is already in part learning structure. This means that the structural notions related to classes can be learnt by exposure to their frequency differentials, and then both the structures and frequencies can be used in processing (Merlo and Stevenson 1999).

Thus, the idea of a hierarchical lexicon organized according to different classes of information yields a unified framework for further exploration of the important issues raised in this volume concerning: lexical organization; the interaction between lexical, syntactic and semantic information and processing; and the role of lexical statistics in guiding the acquisition and interpretation of language. While this is just one of the possible developments in the study of the interface between the lexicon and sentence understanding, the notion of class provides a fruitful ground from which both variability and regularity in language and processing can be successfully investigated.

References

Abney, S. (1996). Partial parsing via finite-state cascades. In Carroll, J. (Ed.), *Proceedings of the Workshop on Robust Parsing at the 8th Summer School on Logic, Language and Information*, Number 435 in CSRP, pp. 8–15. University of Sussex, Brighton.

Abney, S. (1997). Stochastic attribute-value grammars. *Computational Linguistics 23*(4), 597–618.

Black, E., Jelinek, F., Lafferty, J., Magerman, D., Mercer, R. and Roukos, S. (1992). Towards history-based grammars. In *Proceedings of the 30th Meeting of the Association for Computational Linguistics*, Newark, DE, pp. 31–38.

Brants, T. (1999). Cascaded Markov models. In *Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*.

Brants, T. and Crocker, M. (2000). Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, pp. 111–117.

Brent, M. (1993). From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics 19*(2), 243–262.

Brent, M. (1996). Advances in the computational studies of language acquisition. In Brent, M. (Ed.), *Computational Approaches to Language Acquisition*, pp. 1–38. Cambridge, MA: MIT Press.

Bresnan, J. (1982). *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Brew, C. (1995). Stochastic HPSG. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics (EACL'95)*, Dublin, Ireland, pp. 83–89.

Briscoe, T. and Carroll, J. (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, Washington, D.C., pp. 356–363.

Briscoe, T., Copestake, A. and de Paiva, V. (Eds.) (1994). *Inheritance, Defaults and the Lexicon*. Cambridge University Press.

Charniak, E. (1996). Tree-bank grammars. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'96)*, Portland, Oregon, pp. 1031–1036.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference on AI (AAAI'97)*, Providence, Rhode Island, pp. 598–603.

Christianson, K., Hollingworth, A., Halliwell, J. and Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Pychology 42*, 368–407.

Collins, M. and Brooks, J. (1995). Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, MA, pp. 27–38.

Collins, M.J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 184–191.

Collins, M.J. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 16–23.

Cortes, C. and Mohri, M. (2000). Context-free recognition with weighted automata. *Grammars 3*(2/3), 133–150.

Crocker, M. (1995). *Computational Psycholinguistics: An Interdisciplinary Perspective*. Dordrecht: Kluwer Academic Publishers.

Daelemans, W., De Smedt, K. and Gazdar, G. (Eds.) (1992). *Computational Linguistics, Special Issue on Inheritance*, Volume 18. MIT Press.

Dang, H.T., Kipper, K., Palmer, M. and Rosenzweig, J. (1998). Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, Montreal, Canada, pp. 293–299.

Dorr, B. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation 12*(4), 1–55.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language 67*(3), 547–619.

Fellbaum, C. (1998). *Wordnet: an Electronic Lexical Database*. MIT Press.

Fillmore, C. (1988). The mechanisms of "construction grammar." In *Berkeley Linguistics Society*, Volume 14, pp. 35–55.

Fong, S. (1991). *Computational Properties of Principle-based Grammatical Theories*. Ph.D. thesis, MIT, Cambridge, MA.

Ford, M., Bresnan, J. and Kaplan, R. (1982). A competence-based theory of syntactic closure. In Bresnan, J. (Ed.) *The Mental Representation of Grammatical Relations*, pp. 727–796. Cambridge: MIT Press.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph. D. thesis, University of Connecticut. Available through the Indiana University Linguistics Club, Bloomington, IN.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition 68*, 1–76.

Gibson, E. and Hickok, G. (1993). Sentence processing with empty categories. *Language and Cognitive Processes 8*, 147–161.

Gibson, E. and Pearlmutter, N.J. (1994). A corpus-based analysis of psycholinguistic constraints on prepositional phrase attachment. In Clifton, C., Frazier, L. and Rayner, K. (Eds.) *Perspectives on Sentence Processing*, pp. 181–198. Laurence Erlbaum.

Gibson, E., Schütze, C. and Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research 25*(1), 59–92.

Goldberg, A. (1995). *A Construction Grammar Approach to Argument Structure*. Chicago: The University of Chicago Press.

Hale, K. and Keyser, J. (1993). On argument structure and the lexical representation of syntactic relations. In Hale, K. and Keyser, J. (Eds.) *The View from Building 20*, pp. 53–110. MIT Press.

Henderson, J.B. (2000). A neural network parser that handles sparse data. In *Proceedings of the 6th International Workshop on Parsing Technologies (IWPT2000)*, Trento, Italy, pp. 123–134.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press.

Johnson, M. (1989). The use of knowledge of language. *Journal of Psycholinguistic Research 18*(1), 105–129.

Johnson, M. (1998). PCFG models of linguistic tree representations. *Computational Linguistics 24*(4), 613–632.

Johnson, M., Geman, S., Canon, S., Chi, Z. and Riezler, S. (1999). Estimators for stochastic unification-based grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, pp. 535–541.

Joshi, A. and Schabes, Y. (1997). Tree-adjoining grammars. In Rozenberg, G. and Salomaa, A. (Eds.) *Handbook of Formal Languages*, Volume 3, pp. 69–124. Berlin, New York: Springer Verlag.

Just, M.A. and Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review 99*(1), 122–149.

Katz, S. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustic, Speech and Signal Processing 35*(3), 400–401.

Lapata, M. and Brew, C. (1999). Using subcategorization to resolve verb class ambiguity. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, pp. 266–274.

Lapata, M., Keller, F. and Schulte im Walde, S. (2001). Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research 30*(4), 419–435.

Levin, B. (1985). Introduction. In Levin, B. (Ed.) *Lexical Semantics in Review*, Number 1 in Lexicon Project Working Papers, pp. 1–62. Cambridge, MA: Centre for Cognitive Science, MIT.

Levin, B. (1993). *English Verb Classes and Alternations*. Chicago, IL: University of Chicago Press.

MacDonald, M., Pearlmutter, N. and Seidenberg, M. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review 101*(4), 676–703.

Magerman, D. and Marcus, M. (1991). Pearl: A probabilistic parser. In *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany, pp. 15–20.

Magerman, D. and Weir, C. (1992). Efficiency, robustness, and accuracy in Picky chart parsing. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics*, Newark, DE, pp. 40–47.

Marcus, M., Santorini, B. and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics 19*, 313–330.

Marr, D. (1982). *Vision*. S.Francisco, CA: W.H.Freeman.

McCarthy, D. (2000). Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of ANLP-NAACL 2000*, Seattle, WA, pp. 256–263.

Mel'cuk, I. (1988). *Dependency Syntay: Theory and Practice*. Albany, NY: State University of New York Press.

Merlo, P. (1994). A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research 23*(6), 435–457.

Merlo, P. and Stevenson, S. (1998). What grammars tell us about corpora: the case of reduced relative clauses. In *Proceedings of the Sixth Workshop on Very Large Corpora*, Montreal, CA, pp. 134–142.

Merlo, P. and Stevenson, S. (1999). Language acquisition and ambiguity resolution: the role of frequency distributions. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society (CogSci'99)*, Vancouver, Canada, pp. 399–404.

Merlo, P. and Stevenson, S. (2001). Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics 27*(3), 373–408.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. (1990). Five papers on Wordnet. Technical report, Cognitive Science Lab, Princeton University.

Mitchell, D., Cuetos, F., Corley, M. and Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research 24*(6), 469–488.

Moravcsik, E. and Wirth, J. (1983). Markedness – an Overview. In Eckman, F., Moravcsik, E. and Wirth, J. (Eds.) *Markedness*, pp. 1–13. New York, NY: Plenum Press.

Palmer, M. (2000). Consistent criteria for sense distinctions. *Special Issue of Computers and the Humanities, SENSEVAL98: Evaluating Word Sense Disambiguation Systems 34*(1–2), 217–222.

Pinker, S. (1989). *Learnability and Cognition: the Acquisition of Argument Structure*. MIT Press.

Pollard, C. and Sag, I. (1987). *An Information-based Syntax and Semantics*, Volume 13. Stanford University: CSLI lecture Notes.

Prince, A. and Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University.

Pritchett, B. (1992). *Grammatical Competence and Parsing Performance*. Chicago, IL: University of Chicago Press.

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

Ratnaparkhi, A. (1997). A linear observed time statistical parser based on maximum entropy models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*, Providence, RI, pp. 1–10.

Ratnaparkhi, A. (1999). Learning to parse natural language with maximum entropy models. *Machine Learning 34*, 151–175.

Resnik, P. (1992). Probabilistic tree-adjoining grammars as a framework for statistical natural language processing. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France, pp. 418–424.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. New Jersey: World Scientific.

Roark, B. and Johnson, M. (1999). Efficient probabilistic top-down and left-corner parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MD, pp. 421–428.

Roland, D. and Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of the 36th Annual Meeting of the ACL and the 17th International Conference on Computational Linguistics (COLING-ACL '98)*, Montreal, Canada, pp. 1122–1128.

Schabes, Y. (1991). *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph. D. thesis, University of Pennsylvania.

Schabes, Y. (1992). Stochastic lexicalized tree-adjoining grammars. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, Nantes, France, pp. 424–432.

Schulte im Walde, S. (2000). Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbruecken, Germany, pp. 747–753.

Siegel, E. and McKeown, K. (2001). Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics 26*(4), 595–628.

Spivey, M. and Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition 24*, 1521–1543.

Spivey-Knowlton, M. (1996). *Integration of visual and linguistic information: Human data and model simulations*. Ph. D. thesis, University of Rochester.

Spivey-Knowlton, M. and Sedivy, J.C. (1995). Resolving attachment ambiguity with multiple constraints. *Cognition 55*(3), 227–267.

Spivey-Knowlton, M., Trueswell, J. and Tanenhaus, M. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology 47*(2), 276–309.

Srinivas, B. and Joshi, A.K. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics 25*(2), 237–265.

Stevenson, S. (1993). Establishing long-distance dependencies in a hybrid network model of human parsing. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 982–987.

Stevenson, S. (1995). Arguments and adjuncts: A computational explanation of asymmetries in attachment preferences. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, pp. 748–753.

Stevenson, S. and Merlo, P. (1997). Lexical structure and processing complexity. *Language and Cognitive Processes 12*(1–2), 349–399.

Stevenson, S. and Merlo, P. (2000). Automatic lexical acquisition based on statistical distributions. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pp. 815–821. Saarbruecken, Germany.

Sturt, P., Lombardo, V., Costa, F. and Frasconi, P. (2001). A wide-coverage model of first-pass structural preferences in human parsing. In *Paper presented at the 14th CUNY Conference on Human Sentence Processing*, University of Pennsylvania, pp. 23.

Tabor, W. (1995). Lexical change as nonlinear interpolation. In Moore, J.D. and Lehman, J.F. (Eds.), *Proceedings of the 17th Annual Cognitive Science Conference*, Hillsdale, NJ. Lawrence Erlbaum Associates.

Tabor, W. and Tanenhaus, M.K. (1999). Dynamical models of sentence processing. *Cognitive Science 23*(4), 491–515.

Tesar, B. (1996). Computing optimal descriptions for Optimality Theory grammars with context-free position structures. In *Proceedings of 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA, pp. 101–107.

Trueswell, J. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language 35*, 566–585.

Xia, F., Palmer, M. and Joshi, A. (2000). A uniform method of grammar extraction and its applications. In *Proceedings of the Fifth Conference on Empirical Methods in NLP (EMNLP-2000)*, pp. 52–59, Hong Kong.

The lexicon in Optimality Theory[1]

Joan Bresnan
Department of Linguistics,
Stanford University

The view that the lexicon is the source of syntactic variation is widely accepted in various theoretical frameworks, but lexical approaches have not illuminated dialect variation in negative *be* inversion: *Aren't I?* vs. **I aren't* in Standard English, *Amn't I?* vs. **I amn't* in Scots, and *Amn't I?, I amn't* in Hiberno-English. In Optimality Theory (OT), in contrast, the lexicon is not the source but the result of syntactic variation, via the reranking of violable universal constraints. An OT analysis of this dialect variation can successfully relate the inventories of verb forms to other properties of the dialects, such as the use of *are* as a general form in Standard English and the competition between Standard and Scots forms for negation (*nae* vs. *-n't*).

1. The lexicon as the source of syntactic variation

The view that the lexicon is the source of syntactic variation is widely accepted in various theoretical frameworks, and seems to be supported by movement paradoxes such as (1) found in spoken Standard English (Langendoen 1970; Hudson 1977; Dixon 1982; Gazdar et al. 1982; Kim and Sag 1996; Bresnan 2000):

(1) **Standard English negative auxiliary inversion:**
  a. *Aren't you/we/they going? ~ You/we/they aren't going.*
  b. *Isn't she/he going? ~ She/he isn't going.*
  c. *Aren't/*ain't/*amn't I going? ~ *I aren't going.*

In (1a, b) the inverted auxiliary in the interrogative sentence appears to have been moved from an underlying position following the subject – a position in which it overtly appears in the corresponding declarative sentence. In (1c) however, there is no such source for a moved form *aren't*, yielding a movement paradox. Lexicalist constraint-based theories such as Generalized Phrase

Structure Grammar (GPSG, Gazdar et al. 1982), Head-Driven Phrase Structure Grammar (HPSG, Pollard and Sag 1994; Kim and Sag 1996), and Lexical-Functional Grammar (LFG, Kaplan and Bresnan 1982), which generate the overt structures without movements, can simply postulate as an addition to the plural and second person *are*, a specific first person singular negative lexical form of *aren't* that can only be inserted into the inverted position:[2]

(2)

$$\textit{aren't}_1\text{:}\ \begin{bmatrix}\text{NEG} & + \\ \dots & \end{bmatrix}\ \textit{aren't}_2\text{:}\ \begin{bmatrix}\text{PERS} & 1 \\ \text{NUM} & \text{SG} \\ \text{NEG} & + \\ \text{INV} & + \end{bmatrix}$$

A similar approach can be adopted in a transformational framework which allows post-movement lexical feature checking, as does the Minimalist Program (MP, Chomsky 1995). The features of *aren't*$_2$ in (2) could be checked against derived positions; the feature Inverted being a special feature which must be checked in C (the inverted position).

Yet such language-particular lexical feature specifications, whether they are implemented in frameworks with or without movement, are unsatisfying because they fail to relate the specified forms to the rest of the syntactic system. Why, for example, does *aren't* appear in the inverted position in (1) rather than *isn't*? Why does a movement paradox occur in Scots (3) but not in Hiberno-English (4)? These questions remain unanswered.

(3) Scots: *Amn't I going? *I amn't going.*

(4) Hiberno-English: *Amn't I going? I amn't going.*

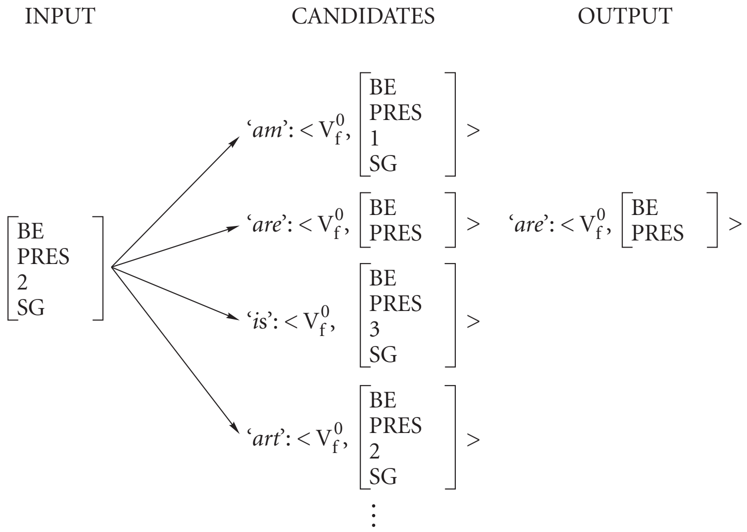
2. The lexicon as the result of syntactic variation (reranking)

A very different conception of the lexicon is found in Optimality Theory (OT): the lexicon is not the source but the result of syntactic variation, via the reranking of violable universal constraints (Prince and Smolensky 1993; Grimshaw 1997a, b; Legendre, Smolensky, and Wilson 1998; Grimshaw and Samek-Lodovici 1998; Samek-Lodovici 1996; Bresnan 2000, 2001a). As we will see, the lexical inventories of present tense *be* forms and their asymmetrical syntactic distribution can be derived from the reranking of general structural markedness and faithfulness constraints in OT.

The particular OT framework assumed here is shown in (5) (Bresnan 2001a).[3]

(5) **OT Morphosyntactic Framework:**

a.


INPUT
CANDIDATES
OUTPUT
[BE PRES 2 SG]
'am': < V^0_f, [BE PRES 1 SG] >
'are': < V^0_f, [BE PRES] >
'is': < V^0_f, [BE PRES 3 SG] >
'art': < V^0_f, [BE PRES 2 SG] >
⋮
'are': < V^0_f, [BE PRES] >

b. GEN: INPUT → CANDIDATES

c. EVAL: CANDIDATES → OUTPUT

A generator GEN produces candidate structural analyses or realizations of the input, as indicated in (5b). Following Jakobson (1984) and Andrews (1990), we may assume that morphosyntactic candidates may have general (nonspecific or vague) meanings. Generality is represented by fewer feature specifications; so general forms express fewer featural distinctions.[4] The candidates are evaluated according to a function EVAL, indicated in (5c). EVAL refers to a Constraint Set, consisting of a hierarchy of (largely) universal, violable constraints:

(6) **Evaluation of candidates:**
Given a language-particular strict dominance ranking of the Constraint Set, the optimal/most harmonic/least marked candidate (= the output for a given input) is one that best satisfies the top ranked constraint on which it differs from its competitors.

Two fundamental conditions hold of the OT framework. First, GEN must be universal. That is, the input and the candidate set are the same for all languages. Systematic differences between languages arise from different constraint rankings, which affect how the candidates are evaluated (Prince and Smolensky 1993; Smolensky 1996), and not from language-particular specifications of differences in input or lexical inventory. This condition is called 'richness of the base'. Secondly, to ensure learnability the input must be recoverable from the output and the output itself must contain the overt perceptible data (Tesar and Smolensky 1998).

Richness of the base is captured in (5) by viewing the morphosyntactic input as arbitrary points in an abstract multidimensional space of dimensions of contrast, formally modelled by complex feature structures. Recoverability of the input from the overt perceptible output is ensured by a well-defined correspondence between feature structures and the types of overt forms of expression which may realize them. Both of these requirements are met by taking the morphosyntactic gen to be a lexical-functional grammar, lfg (Bresnan 2000; Kuhn 2001).[5] In lfg feature structures (f-structures) represent morphosyntactic content in a language-independent format, while categorial structures (c-structures) represent overt forms of syntactic expression.

If both the input and the candidate set are universal, where is the lexicon? In this framework, systematic lexical properties, such as whether there are auxiliary verbs in the inventory of word classes or whether person and number distinctions are neutralized, are derived by constraint ranking. Unsystematic properties must be specified as language-particular properties. Given the constraint ranking for English, then, the lexicon of English is a sampling of the (systematic) inventory (Smolensky 1996), with which unsystematic properties such as language-particular form-meaning correspondences are associated. In (5a) the orthographic labels in single quotes (*'am'*, *'are'*, etc.) represent the pronunciations of various auxiliaries, which are English-particular lexical associations.

The morphosyntactic inventories of English auxiliaries can be derived from the relative ranking of the two types of constraints shown in (7) – constraints on faithfulness to the input ('Faith') and constraints on the structural markedness or wellformedness of forms ('Struct') (Prince and Smolensky 1993; Smolensky 1996).

(7) Constraints:
Faith: Faith$^{P \,\&\, N}$
Struct: *pl, *sg and *2, *1, *3

'Faith$^{P \,\&\, N}$' is violated by any candidate which fails to match the input in both person and number.[6] Struct constraints *2, *1, *3 are respectively violated by candidates specified for second, first, and third person values. Different faithfulness constraints may be instantiated for various morphosyntactically defined domains (Urbanczyk 1995; Benua 1995). In Standard English the three present-tense verbal paradigms (*be*, modal verbs, and other verbs) are thus represented by three different $\textsc{Faith}^{\textsc{P \& N}}$ constraints, of which we will be concerned here only with faithfulness in the domain of the copula (*be*), $\textsc{Faith}^{\textsc{P \& N}}_{be}$.

If all of the structural markedness constraints dominate the faithfulness constraints in the constraint hierarchy, as in (8), then by (6) candidates which violate them will be less optimal than those which do not. ('$c_1 \gg c_2$' means that constraint c_1 outranks constraint c_2 in the constraint hierarchy. The ranking relations of constraints separated by commas are not specified here.)

(8) $^*\textsc{pl},^*\textsc{sg},^*2,^*1,^*3 \gg \textsc{Faith}^{\textsc{P \& N}}_{be}$

Hence, it will be worse for candidates to express number and person contrasts than it will be for them to fail to faithfully preserve the input content. The result will be complete neutralization of person-number contrasts. While most English dialects preserve some contrasts in the present tense of *be*, there are non-Standard English dialects spoken in the West and East Midlands (Cheshire, Edwards, and Whittle 1993: 80) in which complete neutralization has occurred in the past tense, as shown in (9):

(9) West and East Midlands (Cheshire, Edwards, and Whittle 1993: 80):

	sg	pl
1	were	were
2	were	were
3	were	were

I were singing. So were John. Mary weren't singing.

Suppose now that the structural markedness constraints are ranked with respect to the faithfulness constraints as in (10).

(10) $^*\textsc{pl},^*2 \gg \textsc{Faith}^{\textsc{P \& N}}_{be} \gg {}^*\textsc{sg},^*1,^*3$

Standard English:

	sg	pl
1	am	are
2	are	are
3	is	are

The ranking of the markedness constraints for second person and plural above the faithfulness constraint means that violations of the former are worse than violations of the latter. Thus it is worse to express these features than to be unfaithful to the input by failing to preserve them. Hence a general form unmarked for second person or plural number will be preferred over candidates specifically marked for these features. On the other hand, the ranking of faithfulness above the other markedness constraints means that it is worse to fail to express the input features of singular number and first or third person than to bear the complexity penality against marking them. The end result of these rankings will be that specific forms for first or third person singular will be optimal when they match the input, as we see in (11), and the general unmarked form will be optimal elsewhere, as we see in (12).

In these tableaux the constraints are ordered from left to right according to their relative ranking. Violations of constraints are indicated by a *, and the ! denotes a fatal violation, rendering a candidate nonoptimal. The optimal candidate(s) are designated by ☞. Constraint evaluations which have no effect in determining the outcome are shaded gray. Thus the marks incurred in (11) by '*am*', which violates *1 and *SG by bearing the features 1 and SG, are nevertheless overridden by the fatal marks incurred by its unfaithful competitor candidates and have no role here in determining the outcome:

(11) input: [BE PRES 1 SG]

	*PL,*2	$\text{Faith}^{\text{P \& N}}_{be}$	*SG,*1,*3
☞ 'am': [BE PRES 1 SG]			**
'is': [BE PRES 3 SG]		*!	**
'are': [BE PRES]		*!	
'art': [BE PRES 2 SG]	*!	*	*

Similarly, in (12), the perfect faithfulness of second person singular *art* to the input is overridden by its high markedness, and of the remaining unfaithful forms, the least marked candidate wins:

(12) input: [BE PRES 2 SG]

	*PL,*2	FAITH$^{P\,\&\,N}_{be}$	*SG,*1,*3
‘am’: [BE PRES 1 SG]		*	*!*
‘is’: [BE PRES 3 SG]		*	*!*
☞ ‘are’: [BE PRES]		*	
‘art’: [BE PRES 2 SG]	*!		*

By the logic of this theory if the markedness constraint against first person in (12) were promoted above faithfulness, *are* would be generalized to first singular. This possibility is realized in the Southern and East Midland dialects, where both *I are* and *Are I?* are heard (Orton et al. 1962–1971), as shown in (13):

(13) *PL,*2,*1 ≫ FAITH$^{P\,\&\,N}_{be}$ ≫ *SG,*3

Southern and East Midland Counties (Orton et al. 1962–1971)

	sg	pl
1	are	are
2	are	are
3	is	are

I are. Are I?

Conversely, if the markedness constraint against second person were to be demoted below faithfulness, the second person form would now become optimal, as in the older Somerset dialects studied by Ihalainen (1991: 107–108):

(14) *PL,*1 ≫ FAITH$^{P\,\&\,N}_{be}$ ≫ *SG,*2,*3

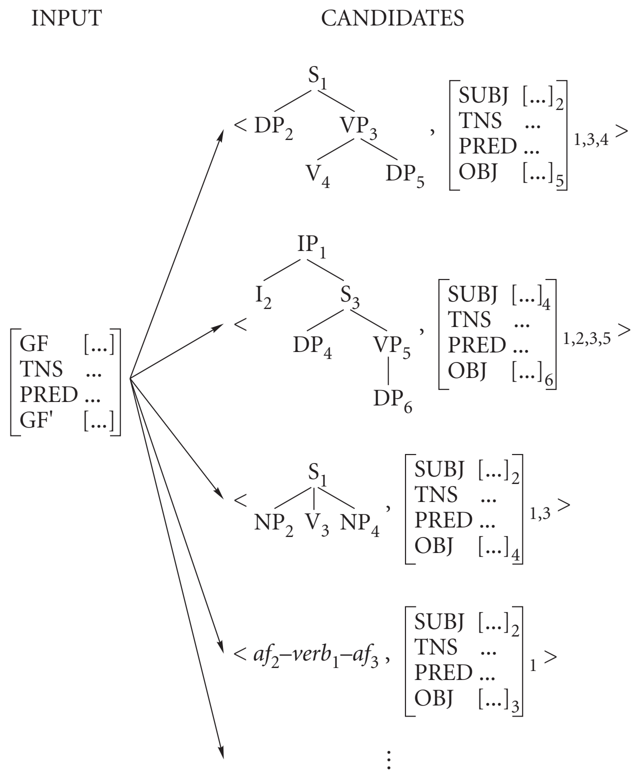
Somerset (Ihalainen 1991: 107–108):

	sg	pl
1	be	be
2	art	be
3	is	be

Note that the orthography and pronunciation of the general form here (*be*) differs from that of the previous dialects (*are*), but this difference in form-meaning correspondences is an unsystematic language-particular property, from the point of view of our constraint ranking.

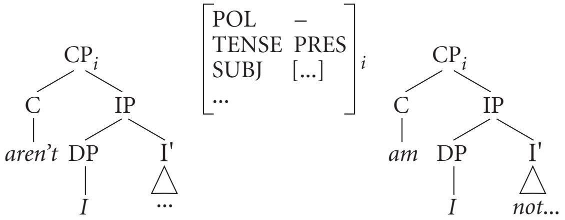
Finally, we observe that the overall structure of this framework for morphosyntax (5) applies as well to larger syntactic structures (Bresnan 2000; Choi 1999; Kuhn 2001; Lee 2001; Sells 2001; Asudeh 2001):

(15) **OT-LFG Syntactic Framework**



The inputs are again f-structures (with undifferentiated argument function types GF, GF′), and the candidates are again pairs of expressions and their corresponding f-structures, but this time at the level of sentence structure (as in LFG and similar syntactic frameworks).[7] In such a framework, words and phrases may be close competitors in the candidate set, expressing essentially the same content, as informally illustrated in (16). Here the synthetic word *aren't* and the analytic phrase *am … not* specify exactly the same f-structure content, consisting of negative clausal polarity [POL –], present tense [TENSE PRES], subject person and number attributes (not shown), and the like:

(16) Words compete with phrases:



It is precisely this kind of competition that explains that movement paradox in negative auxiliary inversion in colloquial Standard English (1) and Scots English (3), as we will see.

3. Negative auxiliary inversion

Let us now address the problem of dialectal variation in inventories of negative auxiliary structures. We limit ourselves to forms used for sentential negation in basic sentences – *standard negation* as defined by Payne (1985). Crosslinguistically, standard negation is overwhelmingly a verbal category (Payne 1985): it occurs as an invariant negative adverb, clitic, or particle associated with VPs and verbs in various clausal positions, as a negative verbal inflection, or as a negative verb root which negates its complement. We set aside discussion of constituent negation here for reasons of space (see Bresnan 2001a).

In what follows 'V' denotes the normal main verb position in English, 'I' is the VP-external position for finite auxiliary verbs and modals in English, and 'C' is the position of the inverted (pre-subject) auxiliary verb. All of these categories denote word class positions, not empty categories representing abstract features or bound morphemes. As in other constraint-based, output-oriented syntactic approaches, the present framework for GEN assumes no derivational operations such as syntactic movements (Bresnan 2000, 2001a, b). Auxiliaries and modal verbs share common categorical features of I and C, which allows them to be base generated in either position (King 1995).

Let us assume that the polarity of clauses in standard negation is represented in the input, and again take partially indeterminate f-structures as our formal model of the input. The output – a syntactic structure and its specific interpretation – will again be formally represented by a corresponding c-structure/f-structure pair.

The inventories of negative auxiliary structures can be derived from the relative ranking of faithfulness and markedness constraints, shown in (17) and (18).

(17) Faith$^{\text{NEG}}$: Sentence scope negation in the input should be preserved in the output.

By definition, all forms of standard negation can express sentence scope negation.[8] As with the markedness constraints on person values (7), Struct constraints penalize the structural complexity associated with the expression of negation:

(18) Struct:

i. Avoid an analytic negator associated with verb phrases or verbs in various positions (VP, V, I, C):
*neg-vp, *neg-v, *neg-i, *neg-c.

ii. Avoid negative inflections of verbs (auxiliaries and modals, or lexical verbs):
*ninfl-aux, *ninfl-v.

iii. Avoid negative lexical verb roots: *neg-vroot.

The relative ranking of these markedness and faithfulness constraints determines the inventory of negative structures for expressing standard negation. For example, if all of the structural markedness constraints for negation are ranked above the faithfulness constraint Faith$^{\text{NEG}}$, the markedness of negative expressions will be worse than the failure to express negation. The resulting grammar would define a hypothetical language severely limited in its expressibility by the absence of specialized expressions for negation. Demotion of one or another markedness constraint below faithfulness will admit the corresponding marked form into the inventory (Bresnan 2001a).

English dialects have several different forms of negation, each of which can be used to express sentential negation under certain circumstances. In Hawick Scots (Brown 1991), the negative clitic *nae* is preferred in unstressed negation of declarative sentences. The contradiction in (19a) shows that *nae* has wide scope over the first clause, while the absence of contradiction in (19b) shows that *no* has narrow scope (constituent) negation:[9]

(19) a. $^{?}$*She couldnae have told him, but she did.*
('It was impossible for her to have told him, but she did tell him.')

b. *She could no have told him, but she did.*
('It was possible for her not to have told him, but she did tell him.')

In questions, however, *nae* cannot be used, and *no* can be used for wide-scope negation:

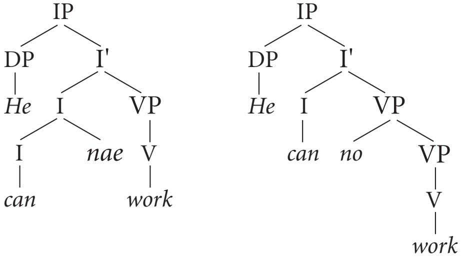
(20) a. **Isnae he coming?* (Hawick Scots – Brown 1991: 80)
b. **Couldnae he work?*
c. **Could he nae work?*
d. *Could he no work?*

Finally, the contracted form *-n't* may be used for sentence negation in interrogatives, but not in declaratives, as we already saw in (3):

(21) a. *Couldn't he work?*
b. **He couldn't work.*

We can explain these facts very straightforwardly in the following way. As shown in (22), the clitic *nae* adjoins to a finite auxiliary or modal in declaratives (in the 'I' position), while *no* adjoins to the VP:[10]

(22) **nae, no in Hawick Scots** (Brown, 1991):



In other words, the clitic *nae* is the Scots pronunciation of NEG-I, while *no* is the Scots pronunciation of NEG-VP. The contracted form *-n't* is a morphological suffix to individual finite auxiliaries or modal verbs, as in Standard English (Zwicky and Pullum 1983). It is thus an instance of NINFL-AUX. The following constraint ranking admits just these three forms into the inventory of standard negation expressions:[11]

(23) Scots:
. . . , *NEG-C ≫ FAITH$^{\text{NEG}}$ ≫ *NEG-VP, *NINFL-AUX ≫ *NEG-I

The relative ranking of these constraints explains the distribution of the various forms of negation. The lowest ranked constraint *NEG-I applies to the least-marked form, *nae*, which is optimal in basic (unstressed) declaratives (19). The highest ranked constraint shown, *NEG-C, being ranked above the faithfulness constraint, eliminates analytic negators in the inverted auxiliary position (C) from the inventory of negation structures altogether. This accounts for (20a, b). The ranking of the remaining markedness constraints makes the forms *-n't* (NINFL-AUX) and *no* (NEG-VP) available in the inventory, but they are optimal for sentence negation only where the less marked analytic negator *nae* is unavailable. (This is so because these forms violate higher-ranking markedness constraints and therefore to minimize violations these forms must be avoided wherever possible.) This explains their appearance in the negative auxiliary inversions (20), (21) and their exclusion from the declarative (19).[12] The logic of this analysis is summarized in the following tables (adapted from Bresnan 2001a):[13]

(24) Scots:

	*NEG-C	FAITHNEG	*NEG-VP, *NINFL-AUX	*NEG-I
input: ¬(POSS(work(he)))				
he couldn't work			*!	
☞ *he couldnae work*				*
he could no work			*!	
input: Q(¬(POSS(work(he))))				
☞ *couldn't he work?*			*	
couldnae he work?	*!			
☞ *could he no work?*			*	

Note the violation of *NEG-C by *couldnae*: *nae* is here the analytic negator associated with a verb in the inverted (C) position. In contrast, the negative suffix in *couldn't* is part of the morphological structure of the verb itself, and therefore does not violate *NEG-C.

On this account what explains the movement paradox of (3), repeated here –

(25) Scots:
*Amn't I going? *I amn't going.*
**Amnae I going? I amnae going.*

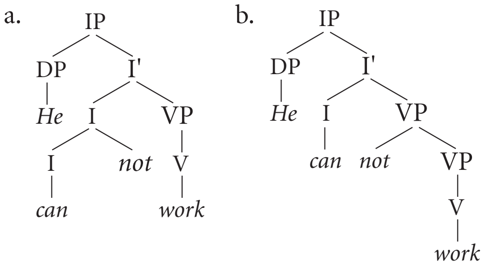
– is the relative markedness of the negative auxiliary inflection *-n't*, compared to the syntactic I negator *nae*. There is independent evidence for such a difference in markedness. The form *nae* is native to Scots, but the auxiliary suffix *-n't* is a Standard English form having restricted use in Scots both socially and lexically. According to Miller (1993), the contracted form *-n't* is preferred by educated speakers of Scots in formal contexts. In Scots it is also lexically restricted compared to *nae*, as shown in (26) from Brown's (1991: 93) study:

(26) *cannae, mustnae, willnae, couldnae, . . .*
**can't, *mustn't, *won't, couldn't, . . .*

The relative markedness of this form is captured in the constraint ranking in (23).

How does Standard English differ from Scots on this theory? Where Scots pronounces neg-i (*nae*) differently from neg-vp (*no*), English pronounces both as *not*:

(27)



The ambiguity of *not* in English has been noted by various researchers (Payne 1985). In Standard English only *not* adjoined to VP can express VP scope in the absence of focus operators, as we see in (28b) and can be separated by adverbs from the modal, as we see in (29c), while only *not* adjoined to I can express sentence scope in declaratives, as we see in (28a), and can form an orthographic word with the modal *can*, as we see in (29a):

(28) a. *He [could not] have been working.* neg-i
b. *He could [not have been working].* neg-vp

(29) a. *He cannot have been working.* ¬(poss(work(he)))
b. *He can (just/simply) not have been working.* poss(¬(work(he)))

Where the Scots ninfl-aux *-n't* is a relatively marked form, the same form in Standard English is among the least marked expressions of sentence negation, an alternative to neg-i in declaratives –

(30) a. *He can't have been working.* ¬(poss(work(he)))
b. *He cannot have been working.* ¬(poss(work(he)))

c. *He can not have been working.* POSS(¬work(he))

– and strongly preferred to NEG-VP in interrogatives. In spoken Standard English examples like (31c) sound very formal (they are termed “stilted and unnatural” by Palmer and Blandford 1969: 293). The more natural expression of wide-scope negation in interrogatives is *-n’t* (31a):

(31) a. *Can’t he have been working?* Q(¬(POSS(work(he))))
b. *Can he not have been working?* Q(POSS(¬(work(he))))
c. %*Can he not have been working?* Q(¬(POSS(work(he))))

All of these differences follow from the constraint ranking shown in (32), in which *NINFL-AUX is ranked below *NEG-VP, in contrast to the Scots ranking (23):[14]

(32) Standard English:
. . . , *NEG-C ≫ Faith$^{\text{NEG}}$ ≫ *NEG-VP ≫ *NEG-I, *NINFL-AUX

The consequences of this ranking are summarized in (33):

(33) Spoken Standard English:

	*NEG-C	Faith$^{\text{NEG}}$	*NEG-VP	*NEG-I, *NINFL-AUX
input: ¬(POSS(work(he)))				
☞ *he can’t have been working*				*
☞ *he cannot have been working*				*
he can not have been working			*!	
input: Q(¬(POSS(work(he))))				
☞ *can’t he have been working?*				*
cannot he have been working?	*!			
can he not have been working?			*!	

The present theory explains why it is in Scots that *-n’t* appears only where *nae* cannot appear, and why there is a contrast in the scope of NEG-VP in Scots and Standard English. It can also easily explain the movement-paradox contrast between Scots and Hiberno-English noted in (3) and (4): Scots rejects the use of *-n’t* in declaratives, while Hiberno-English allows it. The solution is simply that Hiberno-English has the same constraint ranking as Standard English (32). This is a quite plausible approach because in Hiberno-English, unlike Scots, both the NEG-I and NINFL-AUX forms of negation are shared with Standard English.

Despite its similarity to Hiberno-English, Standard English differs conspicuously in one respect: it lacks a negative inflected form for first person singular present tense *be* (1): **I amn't; *Amn't I?* Various explanations for this lexical gap have been proposed; Dixon (1982), for example, proposes avoidance of the phonologically marked *mn* sequence. Here we will simply assume a high-ranking constraint **amn't* which penalizes this candidate, for whatever reason. (In Bresnan 2001a, lexical gaps are analyzed by means of a universal constraint LEX against unpronouncible candidates, which penalizes those candidates idiosyncratically associated with no pronunciation in a language-particular lexicon.)

If no other changes are made to the constraint ranking for Standard English, the consequences of eliminating this candidate are that syntactic constructions with *am … not* replace the missing first person singular negative inflected form of *be* in both declaratives and interrogatives expressing sentential negation:

(34) Possible effect of a lexical gap (I):

	**amn't*	*NEG-C	FAITH-NEG	*NEG-VP	*NEG-I, *NINFL-AUX
(declarative input)					
I amn't working	*!				*
☞ *I [am not] working*					*
I am [not working]				*!	

	**amn't*	*NEG-C	FAITH-NEG	*NEG-VP	*NEG-I, *NINFL-AUX
(interrogative input)					
Amn't I working?	*!				*
Am not I working?		*!			
☞ *Am I [not working]?*				*	

Though some speakers may avoid the lexical gap in this way, it is much more common (certainly in informal spoken Standard American English) to use

Aren't I . . . ?, the apparent "first person" *aren't* of (1), (2). What is happening is that faithfulness to person and number is sacrificed in order to avoid the very marked use of NEG-VP with wide scope. For these speakers, *NEG-VP dominates FAITH$_{be}^{\text{P \& N}}$ in the constraint hierarchy, as shown in (35):

(35) *NEG-VP ≫ FAITH$_{be}^{\text{P \& N}}$ and FAITH$_{be}^{\text{P \& N}}$ ≫ *NEG-I, *NINFL-AUX

With all other constraint rankings the same as before, this means that it is a worse violation to use VP negation (for wide-scope input) than to violate faithfulness to number and person. The main result is shown in (36):

(36) Possible effect of a lexical gap (II):

	**amn't*	*NEG-C	FAITH$^{\text{NEG}}$	*NEG-VP	FAITH$_{be}^{\text{P \& N}}$	*NEG-I, *NINFL-AUX
(declarative input)						
I amn't working	*!					*
I aren't working					*!	*
☞ *I [am not] working*						*
I am [not working]				*!		
(interrogative input)						
Amn't I working?	*!					*
☞ *Aren't I working?*					*	*
Am not I working?		*!				
Am I [not working]?				*!		

The reason that *aren't* is the optimal form here is that the constraints against more faithful analytic expressions of negation such as **Am not I?, *Am I not?* – namely *NEG-C and *NEG-VP – outrank faithfulness to person and number (FAITH$_{be}^{\text{P \& N}}$). According to our analysis of person/number neutralization in Section 2, *are* is the most general form in the present tense paradigm of *be*. Hence, when faithfulness to the input is overridden, *are* will emerge as the least marked form, generalizing further into the paradigm (see Bresnan 2001a).

In conclusion, we see that the movement paradoxes in (1) and (3) are not matters of brute lexical stipulation, but can be derived from more general properties of the grammatical systems of these English dialects: the unmarkedness of *are* in the Standard English paradigm for present *be*, the relative markedness of Standard *-n't* in Scots compared to the non-Standard native form *nae*, and the competition between morphological and syntactic forms of negation across dialects, which follows from the feature-logic based theory of GEN for morphosyntax provided by OT-LFG (Bresnan 2000, 2001a).

Notes

1. This work is based in part on work supported by the National Science Foundation under Grant No. BCS-9818077.

2. The square brackets in (2) employ attribute-value notation, in which + *feature* is rendered [*feature* +] (Johnson 1988).

3. PRES may be regarded as an abbreviation for [TENSE PRES], [2] for [PERS 2], [SG] for [NUM SG], and [BE] for [PRED 'BE'] in the customary attribute-value notation of n. 2. (In LFG values of PRED such as 'BE' stand for an index to lexical semantics. Languages of course vary as to precisely which complexes of semantic primes they lexicalize, a topic beyond the scope of the present study. 'BE' stands for one such complex.) Alternatively, PRES etc. may be interpreted as monovalent (privative) features, which are represented uniquely by their values. The choice of feature interpretations is independent of the main issues addressed here.

4. Output indeterminacy of this sort must not be confused with underspecification in the phonological sense (Steriade 1995). The latter involves the omission of features in underlying structures which are required at the overt level.

5. Kuhn (2001) proves the decidability of the universal parsing problem for the present framework (OT-LFG), raised by Johnson (this volume).

6. As in Bresnan (2001a), faithfulness in fusional morphology is assumed to respect *sets* of values, such as person and number combined in FAITH$^{P\ \&\ N}$. This property in turn may be derived from finer-grained morphological constraints such as 'FUSE-PERS-NUM', which morphemes will satisfy by marking person if and only if they mark number.

7. Expressions of syntax are actually composite, consisting of c-structures and their lexical instantiations. Hence, the candidates are more accurately thought of as quadruples of lexical strings, trees, feature structures and their correspondence functions.

8. However, only forms associated with constituent phrases can express constitent negation of that phrase. Hence, both NEG-VP and NEG-I can express sentence negation, but only NEG-VP can express VP constituent negation (Bresnan 2001a).

9. The form *no* also allows a negative stressed wide scope reading (Brown 1991: 83; Bresnan 2001a).

10. We represent *nae* as adjoined to I for simplicity and clarity, but there are various other ways of associating a NEG-I form with I which could have the same effects within the present framework of assumptions. Note that *nae* cannot be separated from its host I, a fact which requirs additional constraints on clitics or X^0 adjuncts, not discussed here.

11. '…' includes all of the remaining markedness constraints in (18): *NEG-V, *NINFL-V, *NEG-VROOT. These are omitted in (23) for perspicuity.

12. The additional constraints which require auxiliary inversion in questions and its absence in declaratives are discussed in Grimshaw (1997a) and Bresnan (2000).

13. The input is succinctly shown here as a logical formula representing the wide scope of sentential negation rather than as an f-structure. The two constraints separated by commmas are treated as floating constraints having variable ranking values (Boersma 1997; Asudeh 1999). At evaluation these two constraints may be ranked in either order, allowing for variability in the occurrence of the two expressions of negation they mark.

14. The ranking may differ, of course, in more formal varieties of Standard English.

References

Andrews, Avery D. (1990). Unification and morphological blocking. *Natural Language & Linguistic Theory, 8*, 507–557.

Asudeh, Ash (2001). Linking, optionality, and ambiguity in Marathi. In Sells (ed.), 257–312.

Barbosa, Pilar, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky (eds). (1998). *Is the Best Good Enough? Optimality and Competition in Syntax*. Cambridge, Massachusetts: The MIT Press and MIT Working Papers in Linguistics.

Beckman, J., Dickey, L. and Urbanczyk, S. (eds), (1995). *Papers in Optimality Theory.* University of Massacusetts Occasional Papers 18. Amherst: University of Massachusetts.

Benua, Laura (1995). Identity effects in morphological truncation. In Beckman et al. (eds), 77–136.

Bresnan, Joan (2000). Optimal syntax. In Dekkers et al. (eds), 334–385.

Bresnan, Joan (2001a). Explaining morphosyntactic competition. In *Handbook of Contemporary Syntactic Theory*, ed. by Mark Baltin and Chris Collins, 11–44. Oxford: Blackwell.

Bresnan, Joan (2001b). *Lexical-Functional Syntax*. Oxford: Blackwell.

Brown, Keith (1991). Double modals in Hawick Scots. In Trudgill and Chambers (eds), 74–103.

Cheshire, Jenny, Viv Edwards, and Pamela Whittle (1993). Non-standard English and dialect levelling. In Milroy and Milroy (eds), 53–96.

Choi, Hye-Won (1999). *Optimizing Structure in Context: Scrambling and Information Structure*. Stanford, CA: CSLI Publications.

Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, MA: The MIT Press.

Dekkers, J., van der Leeuw, F. and van de Weijer, J. (eds), (2000). *Optimality Theory: Phonology, Syntax, and Acquisition*. Oxford: Oxford University Press.

Dixon, R.M.W. (1982). Semantic neutralisation for phonological reasons. In *Where Have All the Adjectives Gone? and other Essays in Semantics and Syntax*, 235–238. Berlin: Mouton Publishers.

Gazdar, Gerald, Geoffrey Pullum, and Ivan Sag (1982). Auxiliaries and related phenomena in a restrictive theory of grammar. *Language, 58*, 591–638.

Grimshaw, Jane (1997a). Projection, heads, and optimality. *Linguistic Inquiry, 28*, 373–422.

Grimshaw, Jane (1997b). The best clitic: Constraint conflict in morphosyntax. In Liliane Haegeman (ed.), *Elements of Grammar*, 169–196. Dordrecht: Kluwer Academic Publishers.

Grimshaw, Jane and Vieri Samek-Lodovici (1998). Optimal subjects and subject universals. In Barbosa et al. (eds), 193–219.

Hudson, Richard (1977). The power of morphological rules. *Lingua, 42*, 73–89.

Ihalainen, Ossi (1991). On grammatical diffusion in Somerset folk speech. In Trudgill and Chambers (eds), 104–119.

Jakobson, Roman (1984). Structure of the Russian verb. In *Roman Jakobson. Russian and Slavic Grammar. Studies 1931–1981*, ed. by Linda R. Waugh and Morris Halle, 1–14. Berlin: Mouton Publishers. (Original work published 1932.)

Johnson, Mark (1988). *Attribute-Value Logic and the Theory of Grammar*. Stanford, CA: CSLI Publications.

Johnson, Mark. (this volume)

Kaplan, R.M. and Bresnan, J. (1995) [1982]. Lexical-functional grammar: a formal system for grammatical representation. In Dalrymple et al. (eds), 29–130. [reprinted from Bresnan (ed.) 1982, 173–281.]

Kim, Jong-Bok and Ivan Sag (1996). French and English negation: a lexicalist alternative to head movement. Stanford, CA: Stanford University Department of Linguistics MS.

King, Tracy (1995). *Configuring Topic and Focus in Russian*. Stanford, CA: CSLI Publications.

Kuhn, Jonas (2001). Generation and parsing in Optimality Theoretic syntax. Issues in the Formalization of OT-LFG. In Sells (ed.), 313–366.

Langendoen, D. Terence (1970). *Essentials of English Grammar*. NY: Holt, Rinehart, and Winston.

Lee, Hanjung (2001). Markedness and Word Order Freezing. In Sells (ed.), 63–127.

Legendre, Géraldine, Paul Smolensky, and Colin Wilson (1998). When is less more? Faithfulness and minimal links in wh-chains. In *Is the Best Good Enough? Optimality and Competition in Syntax*. Barbosa, Fox, Hagstrom, McGinnis, and Pesetsky (Eds). 249–289. Cambridge: MIT Press.

Miller, Jim (1993). The grammar of Scottish English. In Milroy and Milroy (eds), 99–138.

Milroy, James and Lesley Milroy (1993). *Real English. The Grammar of English Dialects in the British Isles*. London: Longman.

Orton, Harold et al. (eds). (1962–1971). *Survey of English Dialects*. Leeds: Leeds, published for the University of Leeds by E.J. Arnold.

Palmer, Harold E. and F.G. Blandford. (1969). *A Grammar of Spoken English*. Third edition, revised and rewritten by Roger Kingdon. Cambridge: W. Heffer and Sons Ltd.

Payne, John R. (1985). Negation. In *Language Typology and Syntactic Description. Vol. I: Clause Structure*, ed. by Timothy Shopen, 197–242. Cambridge: Cambridge University Press.

Pollard, C. and Sag, I.A. (1994). *Head-Driven Phrase Structure Grammar*. Stanford and Chicago: CSLI Publications and University of Chicago Press.

Prince, Alan and Paul Smolensky (1993). Optimality Theory: constraint interaction in generative grammar. *RuCCS Technical Report* #2. Piscateway, NJ: Rutgers University Center for Cognitive Science.

Samek-Lodovici, Vieri (1996). *Constraints on Subjects. An Optimality Theoretic Analysis*. New Brunswick, NJ: Rutgers University Ph.D. dissertation.

Sells, Peter (2001). *Structure, Alignment and Optimality in Swedish*. Stanford, CA: CSLI Publications.

Sells, Peter (ed.). (2001). *Formal and Empirical Issues in Optimality-Theoretic Syntax*. Stanford, CA: CSLI Publications.

Smolensky, Paul (1996). The initial state and 'richness of the base' in Optimality Theory. *Technical Report* JHU-CogSci-96-4, Department of Cognitive Science, Johns Hopkins University.

Steriade, Donca (1995). Underspecification and markedness. In *Handbook of Phonological Theory*, ed. by John Goldsmith, 114–174. Oxford: Blackwell.

Tesar, Bruce and Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry, 29*, 229–268.

Trudgill, Peter and J.K. Chambers (eds). (1991). *Dialects of English. Studies in Grammatical Variation*. London: Longman.

Urbanczyk, Suzanne (1995). Double reduplications in parallel. In Beckman et al. (eds), 499–531.

Zwicky, Arnold M. and Geoffrey K. Pullum (1983). Cliticization vs. inflection: English *n't*. *Language, 59*, 502–513.

Optimality–theoretic Lexical Functional Grammar

Mark Johnson [*]

Cognitive and Linguistic Sciences, Brown University

1. Introduction

In her chapter in this volume, Bresnan describes a version of Lexical-Functional Grammar (LFG) in which Optimality Theory (OT) constraint satisfaction is used to identify well-formed linguistic structures. Bresnan shows how reranking of constraints changes the set of optimal outputs (surface forms), and uses this to elegantly account for a range of dialectal and cross-linguistic variation in the English auxiliary system.

Rather than focusing on the details of her analysis, this paper concentrates on the broader implications of the OT approach in LFG. We will see that OT LFG involves a fairly radical change to the version of LFG presented in Kaplan and Bresnan (1982) and Kaplan (1995) (called "classical LFG" below), even though that version of LFG is often considered a constraint-based theory of grammar (Shieber, 1992). This change may affect generative capacity: the parsing problem for OT LFG may be undecidable even though the corresponding problem for classical LFG is decidable. However it is not clear how relevant such a result would be, since the OT perspective suggests an alternative account in which sentence comprehension does not involve determining the grammaticality of the sentence being understood. This approach is conceptually related

[*] I would like to thank Avery Andrews and Joan Bresnan, as well as Katherine Demuth, Pauline Jacobson and my other colleagues at Brown, the participants at the 1998 CUNY conference, and the anonymous reviewers for their insightful comments about this paper. This material is partially based on research supported by the National Science Foundation under Grants 9720368, 9870676 and 9812169.

to maximum likelihood parsing, and suggests that OT is closely related to a certain kind of probabilistic language model.

2. Bresnan's analysis of auxiliary selection

This section compares the OT and classical accounts of auxiliary selection. In early versions of LFG, auxiliary selection is intimately related to agreement. Specifically, in a language with subject-verb agreement the lexical entries of tensed auxiliaries and verbs constrain the values of the subject's person and number features, so that each inflected lexical entry determines the range of subject agreement features it can appear with (Kaplan and Bresnan, 1982). This account had the advantage of formal simplicity, but the disadvantage that important substantive properties of the auxiliary system do not follow from the analysis. For example, it seems that every auxiliary or main verb can appear with the full range of subject person and number features (i.e., there are no inflectional gaps), and by and large there is exactly one inflected form of each auxiliary or main verb that agrees with each set of person and number features. This pattern follows from the architecture of Chomsky's original account of the English auxiliary system (Chomsky, 1957), but in classical LFG (and other contemporary unification-based theories) it was essentially stipulated in the lexical entries themselves.[1]

As LFG developed, structure in the lexicon came to play a more prominent role, and facts of the kind just mentioned could be directly captured via lexical redundancy rules or other devices. For example, Andrews' Morphological Blocking Principle (1982, 1990) prevents insertion of a lexical item if another lexical item from the same paradigm with more specific constraints could also be inserted, so an inflected form with no agreement constraints is blocked if another inflected form with more specific constraints is present in the lexicon. Bresnan's OT account in this volume can be viewed as a radical extension of Andrew's Morphological Blocking Principle to all of syntax. Bresnan's examples in which the realization of verbal negation alternates between an inflectional element attached to an auxiliary (e.g., *aren't*) and an independent lexical item (i.e., *not*) provide evidence for competition at the syntactic as well as the morphological level (although perhaps a phonological account could possibly be given of some of the data).

Similar points are made by Grimshaw (1997), Legendre (to appear) and others. Indeed, it seems that Bresnan's analysis of lexical selection does not depend heavily on details of LFG, and could be re-expressed in a non-LFG OT framework. Perhaps this is because Bresnan's account follows primarily from the particular constraints she posits and their ranking, neither of which are LFG-specific, rather than details of LFG's syntactic representations. It should be possible to express Bresnan's account in any OT-based syntactic framework so long as the syntactic representations permit one to identify from candidates the features Bresnan's analysis requires; the "unification-based" machinery of classical LFG seems largely superfluous.

2.1 Inflectional classes

In morphological blocking accounts such as Andrews', competing lexical forms are always ranked in terms of featural specialization, while in OT a language-particular constraint ranking determines how competing forms are ordered. Depending on the constraint-ranking, it may turn out that some candidate feature combinations are not the optimal surface forms for any input feature specification, so the constraint ranking effectively determines the range of possible candidate features and hence possible lexical entries.

For example, as Bresnan shows the presence of exactly the two specialized forms *am* and *is* in the present tense paradigm for BE follows from constraint ranking *2, *PL $\gg$ FAITH$^{\text{PERS\&NUM}}$ $\gg$ *1, *3, *SG. However, note that the regular present tense verb paradigm in English contains only one specialized form (3rd singular), which would require a different constraint ranking, namely *1, *2, *PL $\gg$ FAITH$^{\text{PERS\&NUM}}$ $\gg$ *3, *SG. Thus each inflectional class must be somehow associated with its own constraint ranking, rather than there being a single constraint ranking holding across a language.

Further, inflectional form selection in Bresnan's account seems to be fundamentally a choice between either a form that is specialized for a particular combination of input features or a general unspecialized form. However, not all inflectional patterns can be described in this way. For example, the more specialized form *was* surfaces in both the first and third person singular forms of the past tense of BE in Standard English. The constraint ranking for present tense BE given by Bresnan would permit specialized forms to appear in these two positions in the paradigm, but does not explain their homophony.

2.2 Universals in OT LFG

Moving to more general issues, it is interesting to ask whether and how the OT LFG framework Bresnan outlines is capable of expressing putative typological universals that have been proposed elsewhere. Greenberg (1966) proposes several well-known universals concerning agreement. Some of these can be straight-forwardly expressed in Bresnan's framework, although they do not seem to follow from deeper principles.

> *Greenberg's Universal 32*:
> Whenever the verb agrees with a nominal subject or nominal object in gender it also agrees in number.

This could be expressed as a substantive universal requirement that every constraint ranking must satisfy, viz.:

> If g is a gender feature and $\text{FAITH}^g \gg {*}g$, then there is a number feature n such that $\text{FAITH}^n \gg {*}n$.

However, other universals proposed by Greenberg cannot be expressed so straight-forwardly.

> *Greenberg's Universal 37*:
> A language never has more gender categories in nonsingular numbers than it does in the singular.

This universal does not seem to be easy to express as a condition on constraint rankings, although sufficient conditions which ensure that the language generated by a constraint ranking satisfies this universal seem easy to state. For example, if $*\text{SG} \not\gg *\text{PL}$ then singular forms will never be more marked than corresponding plural forms, from which Universal 37 follows.

3. Formal implications for LFG

The previous section focussed on the empirical implications of Bresnan's analysis. This section investigates the impact of Bresnan's adoption of OT competitive constraint satisfaction on the formal basis of LFG. Classical LFG as formulated in Kaplan and Bresnan (1982) and Kaplan (1995) is often described as a "constraint-based" theory of grammar (Shieber, 1992). The constraints in classical LFG are "hard" in the sense that a single constraint violation leads to ungrammaticality. Competition plays no role in classical LFG, although there have been other proposals besides Bresnan's to add it to LFG such as Frank et al. (1998).

Bresnan's OT account cannot be regarded merely as a theory of the lexicon, with the resulting lexical entries interacting syntactically via the "hard" constraint mechanisms of classical LFG: one of the major points in Bresnan's paper is that competition between ranked constraints determines a language's multi-word syntactic constructions in the same way as it determines the language's lexical inventory. Thus OT competition cannot be restricted to the lexicon, and syntactic structures must be permitted to compete. This makes the mechanisms operative in the lexicon and the syntax much more uniform in Bresnan's account than they were in earlier LFG accounts. But as subsection 4.1 discusses, such syntactic competition may make the parsing problem much harder.

3.1 Feature structure constraints in OT LFG

While Bresnan says that she intends her feature structures to be standard attribute-value structures, it is striking that in Bresnan's fragment the features associated with a candidate f-structure are merely sets of atoms, rather than the attribute-value pairs of classical LFG. For example, Bresnan's lexical entry for the candidate form *am* is merely [BE 1 SG], whereas in a comparable classical LFG lexical entry each atom would appear as the value of a unique attribute or f-structure function, i.e., [[PRED=BE],[PERSON=1],[NUMBER=SG]]. The additional structure is necessary in classical LFG and other "unification-based" theories since they rely on *functional uniqueness* in their account of agreement. Agreeing elements both specify values for the same attribute. Functional uniqueness requires that each attribute have a single value, so if the values specified by the agreeing elements differ then the construction is ill-formed. For example, in a language with subject-verb number agreement a singular subject specifies that the value of its NUMBER attribute is SG, and a plural verb specifies that the value of that same NUMBER attribute is PL. But the functional uniqueness constraint requires that the NUMBER attribute have a single value, so any syntactic structure in which a singular subject appears with a plural verb would be ungrammatical. More abstractly, one role of the attributes in classical LFG is to formally identify which feature values clash. Continuing with the example, SG and PL clash because both are the value of the same NUMBER attribute, while SG and 1 do not clash because they are values of different attributes.

Bresnan's account focuses on the possible realizations of inflectional forms within verb phrases, and does not discuss subject-verb agreement per se. Bresnan intends these atomic features as abbreviations for attribute-value pairs, and the resulting f-structures meet all of the conditions classical LFG imposes on well-formed f-structures, and it seems possible use functional uniqueness to force subject-verb agreement as in classical LFG. However, whenever a new mechanism (in this case, OT competition between syntactic structures) is added to the formal machinery of a theory, one should ask if that mechanism supplants or makes redundant other mechanisms used in the theory.

Besides its role in agreement, functional uniqueness is also used in classical LFG to ensure that the grammatical functions of a clause (e.g., SUBJ, OBJ, etc.) are not doubly filled. But recent work semantic interpretation in LFG has adopted a "resource-based" linear logic approach which enforces both functional completeness and functional uniqueness as a by-product of semantic interpretation (Dalrymple, 1999). Johnson (1999) extends this approach to provide a feature structure system that can account for agreement without any functional uniqueness constraint.

Indeed, a direct extension of Bresnan's own analysis can account for subject-verb agreement without appealing to functional uniqueness or linear logic resource mechanisms. In this extension I distinguish the subject's semantic argument-structure features appearing in the input, which I write as 'SG', '1', etc., from the corresponding superficial verbal inflection features '$\mathrm{SG_V}$', '$\mathrm{1_V}$', etc, which I take to appear in candidate representations only. (Presumably nominal inflection is encoded using similiar nominal features '$\mathrm{SG_N}$', '$\mathrm{1_N}$', although for simplicity I ignore this here.) The faithfulness constraint FAITH ensures that the input features appear in the candidates. I posit an additional constraint $\mathrm{AGR_S}$, which is violated by a candidate representation whenever a verb's person or number inflection feature differs from its subject's corresponding feature in that candidate.[2] The ranking of the $\mathrm{AGR_S}$ constraint relative to the constraints $\mathrm{{*}SG_V}$ and $\mathrm{{*}1_V}$ determines the possible inflected forms of a verb in exactly the same way that the relative ranking of $\mathrm{FAITH^{PERS\&NUM}}$, *SG and *1 determines the inflected forms in Bresnan's account (presumably object agreement inflection is determined by the relative ranking of a similiar $\mathrm{AGR_O}$ constraint).

Consider the example *I am*. Bresnan's analysis of present-tense *be*, expressed in terms of the constraints just discussed, corresponds to the constraint order FAITH $\gg$ $\mathrm{{*}2_V}$, $\mathrm{{*}PL_V}$ $\gg$ $\mathrm{AGR_S}$ $\gg$ $\mathrm{{*}1_V}$, $\mathrm{{*}3_V}$, $\mathrm{{*}SG_V}$. Just as in Bresnan's analysis, FAITH is a faithfulness constraint which is violated when an argument structure feature in the input fails to appear in a candidate: it appears undominated here because its role in Bresnan's analysis is played by $\mathrm{AGR_S}$ here.

Input:	[BE, SUBJ [PRO, 1, SG]]	FAITH	$\mathrm{{*}PL_V}$, $\mathrm{{*}2_V}$	$\mathrm{AGR_S}$	$\mathrm{{*}SG_V}$, $\mathrm{{*}1_V}$, $\mathrm{{*}3_V}$
'I are':	[BE, SUBJ [PRO, 1, SG]]			*!*	
☞ 'I am':	[BE, $\mathrm{1_V}$, $\mathrm{SG_V}$, SUBJ [PRO, 1, SG]]				**
'I is':	[BE, $\mathrm{3_V}$, $\mathrm{SG_V}$, SUBJ [PRO, 1, SG]]			*!	**
'She is':	[BE, $\mathrm{3_V}$, $\mathrm{SG_V}$, SUBJ [PRO, 3, SG]]	*!			**

It should be clear that because of the close correspondence between this approach and Bresnan's, all of Bresnan's analyses can be expressed in the manner just described. Thus using just the mechanisms Bresnan assumes, it is possible to account for subject-verb agreement without appealing to f-structure constraints such as functional uniqueness. Thus feature structure well-formedness constraints, such as functional uniqueness, that play such a central role in classical LFG may not be needed in OT LFG, leading to a radical simplification of the formal machinery of LFG.

4. Parsing in OT LFG

In computational linguistics and psycholinguistics, parsing refers to the identification of the syntactic structure of a sentence from its phonological string. In OT LFG, the universal parsing problem[3] is as follows:

> *The universal parsing problem for OT LFG*:
> Given a phonological string s and an OT LFG G as input, return the input-candidate pairs $\langle i, c\rangle$ generated by G such that the candidate c has phonological string s and c is the optimal output for i with respect to the ordered constraints defined in G.

The corresponding universal parsing problems for classical LFG and other unification-based theories are computationally difficult (NP-hard) but decidable (Barton, Berwick and Ristad, 1987).

4.1 Complexity of OT LFG parsing

One might suspect that the global optimization over syntactic structures involved in OT LFG and other optimality-theoretic grammars may make their parsing problems more difficult than than those of corresponding theories without OT-style constraint optimization. This is because the well-formedness of a candidate representation may involve a comparison with candidates whose phonological strings differ arbitrarily from the string being analyzed. Just because a candidate is higher ranked than all other candidates with the phonological string being parsed does not guarantee that it is the optimal candidate for any input, since there may be higher ranked candidates with other phonological strings. The situation is depicted abstractly in Figure 1. In this figure the phonological string s_2 appears in two candidates c_2 and c_3. However, the input-candidate pair $\langle i_1, c_2\rangle$ is not an optimal candidate since the pair $\langle i_1, c_1\rangle$ is

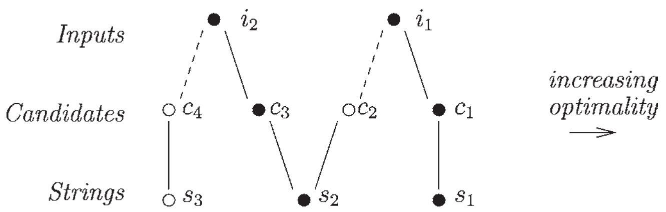


Figure 1. The highest ranked candidate (c_2) with a given phonological string (s_2) need not be an optimal candidate for any input, and an optimal candidate (c_3) for some input (i_2) need not be the highest ranked candidate for any string.

more optimal. On the other hand, the pair $\langle i_2, c_3 \rangle$ is optimal, even though the corresponding candidate c_3 is ranked lower than c_2. The phonological string s_3 is ungrammatical, since c_4, the only candidate with string s_3, is not the optimal candidate for any input.

In the mathematical study of parsing complexity it is standard to work with a simplification of the parsing problem called the recognition problem.

> *The universal recognition problem for OT LFG*:
> Given a phonological string *s* and an OT LFG *G*, answer 'yes' if there is an input *i* which has an optimal candidate with *s* as its phonological string, otherwise answer 'no'.

Because a solution to the universal parsing problem implies a solution to the universal recognition problem, the complexity of the universal recognition problem is a lower bound on the complexity of the universal parsing problem. Depending on exactly how OT LFG is ultimately formalized, it may be possible to show that the universal recognition problem for OT LFG, and hence the universal parsing problem, is undecidable. The idea is to reduce the universal recognition problem for OT LFG to the emptiness problem for classical LFG, which is known to be undecidable (Kaplan and Bresnan, 1982). It is well-known that for any Turing machine M there is a classical LFG G_M whose terminal strings are precisely the sequences of moves of M's halting computations (Johnson, 1988). In effect, the string w that G_M recognizes is the sequence of computational steps that the M performs. Using G_M to recognize w is equivalent to checking that w is in fact a legitimate sequence of computational steps for the machine M. This computation is not especially difficult, since w itself specifies exactly which steps must be checked. However, the problem of determining if any such w exists is extremely hard: indeed, there is no algorithm for determining if any such w exists, which implies that the emptiness problem for classical LFG is undecidable.

The undecidability of the emptiness problem for classical LFG might be adapted to show the undecidability of the universal recognition problem for OT LFG as follows. Suppose that OT LFG is formalized in such a way that for every Turing machine M there is a grammar G'_M whose candidate set consists of the set S_M of the syntactic structures generated by G_M plus a single extra syntactic structure s recognizable in some obvious way, say by having the unique terminal string "Doesn't Halt". (This is clearly possible if the candidate set in a OT LFG can be any set generated by a classical LFG, as Bresnan proposes.) Further, suppose the constraints can be arranged so that every syntactic structure in S_M is more optimal than s. For example, one might introduce a feature FAIL which appears only on s, and introduce *FAIL as an undominated constraint. Then G'_M generates "Doesn't Halt" if and only if there are no syntactic structures more optimal than s, i.e., if and only if S_M is empty. But this latter condition holds if and only if the Turing machine M halts. Since there is no algorithm for determining if an arbitrary Turing machine M halts, there is also no algorithm for determining if the string "Doesn't Halt" is generated by G'_M, i.e., there is no algorithm which can solve the universal recognition problem if grammars such as G'_M are OT LFGs.

Thus the question becomes: under what assumptions would grammars such as G'_M be expressible as OT LFGs? Bresnan describes an OT LFG as having its input and candidate sets generated by classical LFGs. If no further constraints are imposed, then the procedure for constructing the classical LFGs G_M described in Johnson (1988) could be straight-forwardly adapted to generate OT LFGs G'_M as described above, and the undecidability result would presumably follow.

Would reasonable restrictions on OT LFGs rule out such pathological grammars? It is certainly true that construction just sketched yields grammars quite unlike linguistically plausible ones. But this observation does not justify ignoring such complexity results; rather it challenges us to try to make precise exactly how the artificial grammars required for the complexity proof differ are linguistically implausible. Note that the construction makes no assumptions about the input set (indeed, it is systematically ignored), so assuming it to be universally specified has no effect on the construction.

Kuhn (2000a) points out that restricting Gen so it only generates candidates whose f-structures differ from the input in certain minor ways always results in a finite candidate set, from which decidability follows. While Kuhn's particular proposal has minor technical difficulties (e.g., it does not permit epenthetic pronouns), it seems a set of constraints on Gen can be formulated which ensure that the candidate set is finite, yet includes all cross-linguistically attested structures. A cynic might describe such constraints on Gen as the analog of the Offline Parsability Constraint, which ensures the decidability of classical LFG (Kaplan and Bresnan, 1982; Pereira and Warren, 1983). Still, ensuring decidability via such a constraint on Gen seems to go against the spirit of Optimality Theory, since constraints on Gen are "hard" constraints that do not interact via the standard OT mechanism.

However, it is important to note that OT LFG may be decidable even if the candidate set generated by Gen is infinite. For example, in Bresnan's OT LFGs, optimization is local to the clause, i.e., the global optimum can be obtained by optimizing each clause independently. If all optimization in OT LFG is over bounded domains, then one might be able to exploit this to prove decidability without imposing external constraints on Gen that ensure that the candidate set for any given input is always finite. Other approaches also seem possible here. Taking a different tack, in recent work Kuhn (2000b) suggests reformulating OT LFG to require *bi-directional optimization*, which implies decidability without imposing external constraints on Gen.

4.2 Alternative perspectives on parsing

If it is possible to exploit the locality of constraint optimization in OT LFG as suggested above to show that there are only a finite number of clausal input feature combinations and candidate clausal structures then it may be possible to precompute for each lexical item the range of input clauses for which it appears in the optimal candidate. Under such conditions, OT LFG parsing need not involve an explicit optimization over candidates with alternative phonological strings, but might be "compiled" into a parsing process much like one for classical LFG. (Tesar (1995) exploits similiar locality properties in his "parsing" algorithm, while Frank and Satta (1998) and Karttunen (1998) show how a different kind of OT grammar can be compiled into a finite-state transducer.) In such a system the OT constraints would serve to specify the morphosyntactic inventory of a language (i.e., account for cross-linguistic variation), but might not actually be used on-line during parsing.

A more radical approach is to reformulate the OT LFG parsing problem so that parsing only optimizes over candidates with the same phonological string, perhaps as follows:

> *The revised universal parsing problem for OT LFG*:
> Given a phonological string *s* and an OT LFG *G* (i.e., a set of ranked constraints and a lexicon), find the optimal candidates from the set of all candidates with *s* as their phonological string.

Under this revision, a parser presented with input s_2 in Figure 1 would produce c_2 as output, even though c_2 is not an optimal candidate for any input. This revised parsing problem could be computationally much simpler than the OT LFG parsing and recognition problems, as optimization over candidates with phonological strings that differ arbitrarily from the string being parsed (a crucial component of the undecidability proof sketch just presented) no longer occurs. Frank et al. (1998) have extended a classical LFG parser in exactly this way. Stevenson and Smolensky (1997), working in a slightly different framework, show how this kind of model can account for a variety of psycholinguistic phenomena. They also point out that grammatical constraints may need to be reinterpreted or reformulated if they are to be used in such a parsing framework, and this seems to be true in the OT LFG setting as well. Indeed, it is not clear how or even if Bresnan's analysis could be restated in this framework.

Smolensky (1997) points out that in general the set of phonological forms generated by an OT grammar is a subset of the set of phonological forms which receive an analysis under the revised parsing problem above. The language generated by an OT LFG can differ dramatically from the language accepted under the revised definition of the parsing problem. However, this may not be altogether bad, since humans often assign some interpretation to ungrammatical phonological strings. For example, the phonological string *I aren't tired* is interpretable, yet it is not the phonological string of any input's optimal candidates in Bresnan's OT LFG. Schematically, such a string may play the role depicted by s_3 in Figure 1; it is ungrammatical since it is not the optimal candidate for any input, but under the revised definition of the parsing problem it receives the parse c_4.

4.3 Optimality Theory and probabilistic grammars

Prince and Smolensky (1998) speculate that there is a "deep" relationship between optimality theory and connectionism. This section presents a related result, showing a close connection between the revised OT parsing problem and the maximum likelihood parsing problem, which is often adopted in probabilistic parsing. Both problems involve selecting a parse of the phonological string which is optimal on an ordinal scale, defined by ranked constraint violations in the case of OT, or a probability distribution in the case of probabilistic parsing.

Specifically, the revised OT parsing problem is closely related to a very general class of probabilistic models known as Gibbs distributions, Markov Random Fields models, or Maximum Entropy models. See Jelinek (1997) for an introduction, Abney (1997) for their application to constraint-based parsing, and Johnson et al. (1999) for a description of a stochastic version of LFG using such models. Eisner (2000) also notes the connection between OT and Maximum Entropy models. In this kind of model, the logarithm of the likelihood $P(\omega)$ of a parse ω is a linear function of real-valued properties $v_i(\omega)$ of the parse, i.e.,

$$P(\omega) \ = \ \frac{1}{Z} \exp(\sum_{i=1,\ldots,n} -\lambda_i v_i(\omega)).$$

In this class of models, $v_i(\omega)$ is the value of the ith of n properties of the parse ω, λ_i is an adjustable weight of property i, and Z is a normalization constant called the "partition function". The theory of these models imposes essentially no constraints on what the properties v_i can be, so we can take the properties to be the constraints of an OT grammar and let $v_i(\omega)$ be the number of times the ith constraint is violated by ω.

Suppose there is an upper bound c to the number of times any constraint is violated on any parse,[4] i.e., for all ω and i, $v_i(\omega) \leq c$. For simplicity assume that the OT constraint ranking is a linear order, i.e., that the ith constraint outranks the $i+1$th constraint. This implies that the OT parse ranking is the same as the lexicographic ordering of their property vectors $\tilde{v}(\omega)$. Set $\lambda_i = (c+1)^{n-i}$, which ensures that a single violation of the ith constraint will outweigh c violations of constraint $i+1$. It is straightforward to check that for all parses ω_1, ω_2, $P(\omega_1) > P(\omega_2)$ iff $\tilde{v}(\omega_1)$ lexicographically precedes $\tilde{v}(\omega_2)$, which in turn is true iff ω_1 is more optimal than ω_2 with respect to the constraints.

This result shows that if there is an upper bound on the number of times any constraint can be violated in a parse, the revised OT parsing problem can be reduced to the maximum likelihood parsing problem for a Gibbs form language model. It implies that although OT grammars are categorical (i.e., linguistic structures classified as either grammatical or ungrammatical), they are closely related to probabilistic language models; indeed, they are limiting cases of such models. This raises the possibility of applying techniques for parsing and learning for one kind of model to the other. For example, it might be interesting to compare the constraint re-ranking procedure for learning OT constraint rankings presented in Tesar and Smolensky (1998) with the statistical methods for estimating the parameters λ_i of a Gibbs distribution described in Abney (1997), Jelinek (1997) and Johnson et al. (1999).

5. Conclusion

The Optimality-theoretic version of Lexical Functional Grammar that Bresnan provides not only an interesting account of cross-linguistic variation in the lexical inventories of auxiliary verbs and negation, it also provides a framework in which linguistic universals can be systematically explored. It has implications for the formal basis of LFG and other "unification-based" grammars, as it suggests that other linguistic processes, such as agreement, can be viewed in terms of competitive constraint satisfaction. Perhaps as importantly, by recasting LFG into a ranked constraint setting, Bresnan's work suggests novel ways of approaching parsing and learning in LFG. Specifically, the fact that well-formedness in Optimality Theory is defined in terms of an optimization suggests a close connection with probabilistic language models.

As noted above, Bresnan's analysis does not depend heavily on the details of LFG's syntactic representations, and it could be re-expressed in a variety of OT-based syntactic frameworks. Indeed, it is only necessary that we be able to identify the constraint violations Bresnan posits from the candidate structures; exactly how these constraint violations are encoded in candidate structures seems to be of secondary importance. This seems to be a general property of OT-based accounts. Thus from the perspective of both parsing and learning, the details of the representations used in an OT account are less important than the kinds of constraints that the account posits.

Notes

[1] In a transformational grammar with no optional rules the architecture of the grammar guarantees that each underlying or deep structure will have at most one surface form. No similar property seems to follow from the architecture of a mono-stratal constraint-based grammar such as classical LFG.

[2] Because AGR_S only refers to candidate representations, this account does not require that agreement features appear in the input. Thus it is not necessary to assume that language-specific agreement features appear in the input.

[3] A parser is a device that returns the analyses of its inputs with respect to some fixed grammar. A *universal* parser is one in which the grammar *G* is part of the parser's input: i.e., a universal parser must be capable of parsing using any grammar.

[4] Frank and Satta (1998) and Karttunen (1998) also assume such a bound.

References

Abney, Steven P. (1997). Stochastic Attribute-Value Grammars. *Computational Linguistics, 23*(4), 597–617.

Andrews, Avery (1990). Unification and morphological blocking. *Natural Language and Linguistic Theory, 8*, 507–557.

Andrews, Avery D. (1982). The representation of Case in modern Icelandic. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, pp. 427–502. The MIT Press, Cambridge, Massachusetts.

Barton, Jr., Edward, G., Robert C. Berwick, and Eric Sven Ristad (1987). *Computational Complexity and Natural Language*. The MIT Press, Cambridge, Massachusetts.

Chomsky, Noam (1957). *Syntactic Structures*. Mouton, The Hague.

Dalrymple, Mary (1999). *Semantics and Syntax in Lexical Functional Grammar: The Resource Logic Approach*. The MIT Press, Cambridge, Massachusetts.

Eisner, Jason (2000). Review of Kager: "Optimality Theory." *Computational Linguistics, 26*(2), 286–290.

Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John Maxwell (1998). Optimality Theory style constraint ranking in large-scale LFG grammars. In Miriam Butt and Tracy Holloway King (eds), *Proceedings of LFG'98*, Stanford, California. CSLI Press.

Frank, Robert and Giorgio Satta (1998). Optimality Theory and the generative complexity of constraint violability. *Computational Linguistics, 24*(2), 307–316.

Greenberg, Joseph H. (1966). Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (ed.), *Universals of Language* chapter 5, pp. 73–113. The MIT Press, Cambridge, Massachusetts.

Grimshaw, Jane (1997). Projection, heads, and optimality. *Linguistic Inquiry, 28*(3), 373–422.

Jelinek, Frederick (1997). *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts.

Johnson, Mark (1988). *Attribute Value Logic and The Theory of Grammar*. Number 16 in CSLI Lecture Notes Series. Chicago University Press.

Johnson, Mark (1999). A resource sensitive interpretation of Lexical Functional Grammar. *The Journal of Logic, Language and Information, 8*(1).

Johnson, Mark, Stuart Geman, Stephen Canon, Zhiyi Chi, and Stefan Riezler (1999). Estimators for stochastic "unification-based" grammars. In *The Proceedings of the 37th Annual Conference of the Association for Computational Linguistics* (pp. 535–541), San Francisco. Morgan Kaufmann.

Kaplan, Ronald M. (1995). The formal architecture of LFG. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen (eds), *Formal Issues in Lexical-Functional Grammar*, number 47 in CSLI Lecture Notes Series (chapter 1, pages 7–28). CSLI Publications.

Kaplan, Ronald M. and Joan Bresnan (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations* (chapter 4, pp. 173–281). The MIT Press.

Karttunen, Lauri (1998). The proper treatment of optimality in computational phonology. In Lauri Karttunen and Kemal Oflazer, editors, *Proceedings of the International Workshop on Finite-State Methods in Natural Language Processing* (pp. 1–12). Bilkent University, Ankara, Turkey.

Kuhn, Jonas (2000a). Generation and parsing in Optimality Theoretic syntax-issues in the formalization of OT-LFG. In Peter Sells (ed.), *Formal and Empirical Issues in Optimality-theoretic Syntax*. CSLI Publications, Stanford, California.

Kuhn, Jonas (2000b). Faithfulness violations and bidirectional optimization. In Miriam Butt (ed.), *Proceedings of LFG 2000*, Stanford, California. CSLI Publications.

Legendre, Geraldine. to appear. Clitics, optimality, and modularity in bulgarian. In F. van der Leeuw, J. Dekkers, and J. van de Weijer (eds), *The Pointing Finger: Conceptual Studies in Optimality Theory*. Oxford University Press.

Pereira, Fernando C.N. and David H.D. Warren (1983). Parsing as deduction. In *The Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics* (pp. 137–144). M.I.T., Cambridge, Massachusetts.

Prince, Alan and Paul Smolensky (1998). *Optimality Theory*. The MIT Press, Cambridge, Massachusetts.

Shieber, Stuart M. (1992). *Constraint-based Grammar Formalisms*. The MIT Press, Cambridge, Massachusetts.

Smolensky, Paul (1997). On the comprehension/production dilemma in child language. *Linguistic Inquiry, 27*(4), 720–731.

Stevenson, Suzanne and Paul Smolensky (1997). Optimal sentence processing. Paper presented at the Hopkins Optimality Theory Fest.

Tesar, Bruce (1995). *Computational Optimality Theory*. Ph.D. thesis, University of Colorado.

Tesar, Bruce and Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry, 29*(2), 229–268.

The lexicon and the laundromat

Jerry Fodor
Center for Cognitive Science, Rutgers University

This paper offers a modest proposal about what can be put in the lexicon and what can't. Here is the proposal: nothing belongs to a lexical entry for a lexical item except what that item contributes to the grammatical representation of its hosts. This is because languages are compositional – if you understand the constituent expression, then you understand the hosts –, but also reverse compositional – if you understand the host you understand the constituents. If you accept this, then the lexicon is extremely exiguous, containing only definitional information. How does such lexical exiguity converge with the notion of lexical mastery that psycholinguistic theorizing requires to account for parsing and learning? My guess is that language acquisition delivers shallow lexical entries and parsing delivers shallow structural descriptions, and that everything else is performance theory. It might be then that the recent consensus on the lexicon, and its centrality, is just a label, and might not have much to do with what lexical entries actually contain.

Way back when I was a boy just out of Graduate School, the unspoken rule in linguistics was this: if you have a thing that you don't know what to do with and that you don't want to have to think about, put it in the semantics. But then along came truth definitions and model theories, and the like, and it began to seem that semantics might actually turn into a respectable kind of intellectual activity. So they changed the rule. The new rule was: If you have a thing that you don't know what to do with and that you don't want to have to think about, put it in the pragmatics. But then along came Griceian implications and relevance theories, and discourse theories, and the like, and it began to seem that pragmatics too might actually turn into a respectable kind of intellectual activity. So they changed the rule again. The new rule is: If you have a thing that you don't know what to do with and that you don't want to have to think about, put it in the lexicon.

Now, I would be the last person in the world who would wish to suggest that theorizing about the lexicon might also turn into a respectable kind of intellectual activity. Still, I think that we should, if we can, try to arrive at at least a *partial* consensus about what can be allowed to go into the lexicon and what cannot. A respect for tidiness commends that, and besides, I don't see how we can seriously raise psycholinguistic questions about how the lexicon is learned, or about the role of the lexicon in parsing, except against the background of such a consensus.

So this talk offers a modest proposal about what can be put in the lexicon and what can't; indeed, about what must be put in the lexicon and what mustn't. The story comes in four parts. First, I'll say what the proposal is; then I'll give what justification I can for endorsing it; then I'll give examples of one or two kinds of theories of the lexicon that the proposal appears to preclude; then I'll say just a word about the implications of all this for psycholinguistics.

Some terminology to begin with. I will think of constraints on the lexicon as, in particular, constraints on the *lexical entries* for *lexical items*. A 'lexical item' is anything that gets a lexical entry. A lexical entry is whatever a grammar says about a lexical item. I suppose, for example, that 'cat' is a lexical item in English. I suppose it's lexical entry in the grammar of English says something about how 'cat' is pronounced (presumably that it's pronounced /cat/); and something about its syntactic form class (presumably that it's a noun); and something about what it means (presumably, at a minimum, that it means *cat*). Most of the discussion to follow will concern the third, 'semantic' parameter of lexical entries, so questions about phonology or about syntax won't matter much for our purposes. It also won't matter just what kinds of linguistic expressions are to count as lexical items. I'll assume that it's something like morphemes. In which case, 'cat' and 'plural' are lexical items, but 'cats' isn't.

I need just one more bit of terminology. Let the 'hosts' H of any expression E be the expressions of which E is a constituent. So, among the hosts of the lexical item 'cat' are the expressions ' ___ s', 'the ___ ', and 'the ___ is on the mat'. And among the hosts of the expression 'the cat is on the mat' are 'the cat is on the mat and the frost is on the pumpkin.' I take it as not seriously a matter of dispute that every expression of any natural language has infinitely many hosts in that language.

Ok, so here is my constraint: *Nothing belongs to the lexical entry for a lexical item except what that item contributes to the grammatical representation of its hosts.*

This is to put the case very informally of course, since I haven't said what either 'contributes to' or 'determines' means. Nor would it be in the least a trivial matter to do so. But that's ok. A loose formulation will suffice for my polemical purposes; and the kind of thing I have in mind is actually quite familiar. So, for example, it is compatible with my constraint that *is pronounced with a /k/* belongs to the lexical entry for 'cat,' since being pronounced with a /k/ is part of what 'cat' contributes to determining the pronunciation of host expressions like 'the cat' and 'the gray cat', etc. Likewise, it is compatible with my constraint that *is a Noun* belongs to the lexical entry for 'cat' since that is part of what 'cat' contributes to the grammatical analysis of such host expressions as 'the cat,' 'the gray cat,' and 'the cat is on the mat.' Likewise, it is compatible with my constraint that *applies to cats* or, for that matter *applies to domestic felines*, or, for that matter, *applies to certain mammals*, should belong to the lexical entry for 'cat', since 'cat' contributes all of these to all of its hosts. Thus, 'gray cat' applies to cats, and to domestic felines, and to certain mammals; and it's part of the story about why it does so that 'gray cat' numbers 'cat' among its constituents. (I'm aware that there are, (ahem!) certain difficulties about 'toy cat,' and 'decoy cat', and 'political fat cat', and so on. But I shall assume, rather grandly, that none of these are mortal for the project that I have in mind.)

So much for telling you what the constraint is that I'm proposing. Now a little about what I take to be its justification.

I begin with a caution: I expect it's clear to everyone that, given the grammar of a language (and putting idioms to one side) the linguistic structural description of a host expression must be entirely determined by the linguistic structural description of its constituents. That's pretty generally assumed to be part of the explanation of how natural languages can be finitely represented and finitely assimilated. Indeed, it's just a way of expressing the idea that natural languages are 'compositional'. All there is that contributes to the analysis of a linguistic expression is the analysis of its constituents together with its principles of construction. Notice that this sort of picture of the relation between hosts, constituents and their respective structural descriptions recurs at every level of a grammar, not just at the semantic level. That the pronunciation of 'the dog' is exhaustively determined by the pronunciations of 'the' and 'dog' (together with its syntax) is part of the explanation of how English contrives to contain, and how English speakers contrive to master, infinitely many pronounceable sequences.

But please notice that the principle I'm commending is different from, and stronger than, the familiar one about compositionality. For example, according to my story, not only is the grammar of a host expression exhaustively determined by the grammar of its constituents, *but the grammar of the constituent expressions is exhausted by what they contribute to their hosts*. It's hardly in dispute that there's nothing more to 'the cat' than what it gets from 'the' and 'cat', and how what it gets from them is put together. But also, according to me, there is nothing more to lexical entries for 'the' and 'cat' than what they contribute to 'the cat' and their other hosts. Notice that, unlike compositionality, this second constraint does *not* follow just from the usual considerations about the finite representability and finite learnability of productive languages. English could be both learnable and finitely representable *and productive* even though the lexical entry for 'cat' includes stuff that 'cat' does *not* contribute to its hosts (so long as it contains no more than finitely much such stuff). To put it slightly differently, the usual considerations establish a *floor* on what a lexical item can have in its lexical entry: the lexical entry for a lexical item must be rich enough to determine the linguisitically salient properties of its hosts. But they don't, in and of themselves, establish a *ceiling* on what can appear in a lexical entry. Whereas, by contrast, the constraint that I'm urging does so. A lexical item must contain all and *only* what determining the linguistics of its hosts requires it to contain.

So, then, what justifies adding this ceiling constraint? It's this: Not only are natural languages compositional, but they are also (what I'll call) *reverse compositional*. Roughly, compositionality says that if you understand the constituent expressions, then you understand the host. Whereas, reverse compositionality says that *if you understand the hosts, then you understand their constituents. Compositionality* is required to explain why everybody who has the entries for 'gray' and 'cat' in his lexicon has 'gray cat' in his ideolect. *Reverse compositionality* is required to explain why everybody who has 'gray cat' in his ideolect has the entries for 'gray' and 'cat' in his lexicon. Notice that it is *not* a truism that everybody who understands 'gray cat' understands 'gray' and 'cat'; nor is it a mere byproduct of the compositionality of English. You can imagine, for example, a perfectly compositional language in which 'gray cat' means just what it means in English (i.e. *gray and a cat*), but in which 'gray' means (*gray, viz the color of Granny's hair*). Understanding 'gray cat' would *not* guaranty understanding 'gray' for the speaker of such a language; indeed, it might reasonably be denied that 'gray' would count as a *semantic* constituent of its hosts in this language (since it doesn't, as it were, contribute *all* of its meaning to them.) Still, 'gray' might count as a *phonological*, and or a *syntactic*, constituent of its hosts; and there is nothing, so far, to stop this language from being finitely learned or, a fortiori, finitely represented.

But, of course, natural languages don't work that way. The way natural languages work is that, idioms excepted, if you understand a host, then, a fortiori, you understand its constituents. Natural languages are thus not just compositional, but also reverse compositional. And, to repeat, just as compositionality puts a floor on lexical entries, so reverse compositionality puts a ceiling on them. In fact, between the two, they damned near lick the platter clean.

Consider, for example, a hypothesis about the lexicon that I gather many psychologists actually endorse; viz that typical lexical entries are specifications of the stereotypes or prototypes of the corresponding lexical items. (Linguists into 'conceptual semantics' or 'cognitive semantics' have been know to endorse this thesis too, which goes to show how bad for linguists hanging around with psychologists can be.) Well, it's a standard objection to this proposal that stereotypes don't, in the general case, compose. So, to invoke the classical example, if a lexical entry provided only the stereotypes of 'fish' and 'pet' in specifyng the semantics of these words, then it would be possible for a speaker who understands 'pet' and 'fish' not to understand 'pet fish'. This is because pet fish are neither stereotypic pets nor stereotypic fish. The long and short is that, on the assumption that lexical meanings are stereotypes, the lexicon fails to play the required role in explaining linguistic productivity. So there must be something wrong with the assumption that lexical meanings are stereotypes. So the story goes.

Amen, say I. However, I'm proposing to add a verse to this litany. For, although the present line of thought rules it out that the entry for 'pet' or 'fish' is *just* a stereotype, it's thus far left open that the lexical entries could be stereotypes *plus some other stuff*. For example, the lexical entry for 'fish' might say that it means *fish and* has such and such a stereotype. That would be consistent with the idea that 'pet fish' gets its meaning from, and only from, its constituents *and it would also be compatible* with the idea that the fish stereotype is part of the lexical entry of *fish*; viz a part that 'fish' does *not* transmit to its hosts. It would, ipso facto, be an assumption of this revised stereotype story that a lexical entry can contain *more* than the corresponding lexical item contributes to its hosts; hence that there can be more to understanding a lexical item than knowing what its hosts inherit from it.

To repeat: I think that that such a view would be internally perfectly coherent and that it would be perfectly compatible with explaining the productivity, systematicity, etc. of the hosts of 'pet' and 'fish'. However, reverse compositionality rules it out. For, on the revised stereotype account, it would be possible to know the meaning of 'pet fish' but *not* to know the meaning of 'pet' or of 'fish'. This is because, by assumption, their meanings include their stereotypes; and, patently, you could know what 'pet fish' means and *not* know what *stereotypical* pets and fish are like. Indeed, you could even *know the pet fish stereotype* and not know what stereotypical pets and fish are like. Stereotypical pet fish are goldish in color, and live in bowls, and are not good to eat. None of this is true of stereotypic fish. Stereotypic pets are warm and cuddly and say things like woof and meow. None of this is true of stereotypic pet fish either.

In fact, the reverse compositionality argument that stereotypes aren't parts of lexical entries generalizes to preclude *any* merely epistemic properties of lexical item from being part of a lexical entry. Suppose you're *absolutely certain* that all brown cows live in New Jersey. You'd go to the wall for it. Indeed, you're *more* certain that all brown cows live in New Jersey than you are that 7 + 5 = 12. So, just as you might think that it's part of what '5' means that if you add it to '7' you get *12*, so you might think that it's part of what 'brown' means that if you add it to 'cow' you get *lives in New Jersey*. I've actually heard this kind of suggestion made by a ravening semantic holist who was hell bent to identify the content of a lexical item with its entire inferential role. But reverse compositionality (to say nothing of minimal common sense) prevents. Reverse compositionality says that *only* what it transmits to its hosts can be part of the semantic analysis of a constitutent. That the cow ones live in New Jersey can't be part of the meaning of 'brown' since, presumably, 'brown' doesn't contribute that fact to such of its hosts as aren't about cows; for example, to 'brown is my Granny's favorite color.'

Between them, compositionality and reverse compositionality rule out practically all the candidates for meanings that are currently in favor in cognitive science; so, at least, I'm inclined to believe. For just one more example: the relative frequency of the constituents of a lexical item doesn't predict the relative frequency of its hosts or vice versa. So if frequency information were part of the entry for a lexical item, compositionality and reverse compositionality would both be violated. So lexical entries cannot contain such information. A fortiori, learning such information can't be a condition for learning the lexicon.

As I say, I think this pattern of argument ramifies widely. But I won't argue for that here. Instead I propose to say a little about what candidates for lexical information are left; i.e. about what account of lexical entries the compositionality constraint and the reverse compositionality constraint do tolerate. Then a word about what all this implies for psycholinguistic issues about the role of the lexicon in mental processes like sentence processing and language acquisition.

The first point to make is that, if you accept the argument so far, then the lexicon that you're left with is extremely exiguous. In fact, as far as I can see, only two theories of lexical entries are possibile: A lexical entry for an item might specify *logically necessary and sufficient* conditions for the item to apply; or it might be merely 'disquotational' (that is, it might say that 'cat' means *cat*;

that 'dog' means *dog*, and the like and not say anything else). In fact, I favor the second of these alternatives, and I think the 'merely' in 'merely disquotational' is misleading and invidious. But never mind. Suffice it that, if you accept the kind of arguments I've been giving and you *don't* think that lexical entries are merely disquotational, then about all that's left is that they are something like *definitions*. Nothing except definitional information *could* be available in lexical entries if, in particular, reverse compositionality is to be enforced. That's because anything that's not part of the definition of 'cow' is ispso facto something you might not know about cows compatible with understanding 'brown cow' or 'that cow'.

Of course, this impacts the connection between linguistics and psycholinguistics. Consider, in particular, the role of the lexicon in parsing. You might well have thought that any realistic theory of how we understand sentences would have to assume a pretty rich lexicon. 'Semantic' effects are practically ubiquitous in what are generally supposed to be experiments on how people parse; and these are paradigmatically the effects of plausibility and context. Well, if the effects of plausibility and context really *are* semantic, then the information they exploit must be in the lexicon, contrary to the very minimalist account of the lexical entries that I've been suggesting. And if they *aren't* literally semantic, and the information they exploit *isn't* in the lexicon, then there's no hope for the idea that the information parsing a language exploits is the very information that *learning the grammar provides*.

The situation is no happier when one thinks about the acquisition problem itself. If only defining inferences are contributed by lexical items to hosts, and if only what it contributes to hosts can appear in a lexical item's entry, then the child has to insure that *nothing but* defining inferences are allowed to get into his lexicon. The problem is that lexical items are learned (not from their occurrences in isolation, but) precisely from their occurrences in host expressions. So, just as *parsing with* the lexicon requires drawing inferences that run from the previously given semantic properties of constituents to the properties of hosts consonant with the constraints compositionality imposes, likewise *learning* the lexicon requires drawing inferences that run from the *previously given* semantic properties of hosts to the properties of their constituents consonant with the constraints that reverse compositionality imposes. And reverse compositionality says that only the properties of hosts *that they have in virtue of their compositional structure*s are to count in inferences of the latter kind. Well, if that's right, then it's hard to see how you can learn a lexical item unless you already know quite a lot about what the compostional semantics of its host expressions *is*. It appears that, in order for the child to learn the lexicon, he needs information about the semantical properties of host expressions of the exactly the sort that knowledge of the lexicon is supposed to provide. This is all too reminiscent of a paradox that's familiar from learnability theory in syntax: Presumably you can't parse without a grammar. But presumably the child can't get at the data that are supposed to constrain his selection of a grammar unless he can *already* parse the sentences he hears.

The psycholinguistic moral of all this might be: So much the worse for the linguistic lexicon; if it's defined by the principles of compositionality and reverse compositionality, then it hasn't much to do with what people parse with when they understand language, or with what they learn when they learn words. And these are, I suppose, what psycholinguists care about most. Perhaps you're prepared to live with this; undifferentiated, interactionistic theories of learning and parsing are everywhere these days, and it's part and parcel of such theories that questions like 'what do you know (/learn) when you know (/learn) a language (/the meaning of a word)?' don't have *principled* answers. Still, it ought to give you pause if grammar (which is, after all, primarily the theory of the compositional structure of a language) comes largely unstuck from theories of learning the language and parsing it. Does it really seem plausible that learning a language does *not* provide one with the information that using the language routinely employs? If so, then what are we to say *is* the relation between learning a language and being able to understand it?

Actually, I'm a little less pessimistic than these last couple of paragraphs may sound. Since I think the lexicon is disquotational, the lexical entry for 'cow' is required to specify only that it means *cow*. And, since the meaning of 'brown cow' is exhausted by what its constituents contribute, all you have to know to know what 'brown cow' means is that it means *brown and cow*. If the bad news is that an exiguous lexicon leaves you with very little information about lexical items to parse with, the good news is that it leaves the information about an utterance that successful parsing requires you to recover pretty exiguous too. A shallow lexicon implies a correspondingly shallow parser. So shallow, indeed, that it could well turn out that *all* that parsing does is to assign linguistic tokens to their linguistic types. That would be to say that, if you understand *a language*, all there is to understanding an *utterance* in that language is figuring out *which* linguistic type it's a token of. (Mark Steedman reminds me, rightly, that this can't be true where the utterance contains indexicals and the like. Point taken.)

On the other hand, on this sort of account, the experimental problem of isolating bona fide parsing processes from the background of inferential elaboration in which they are normally embedded might prove to be formidable. For example, showing that parsing isn't, as one says, 'autonomous' or 'modular' would require showing that the parser routinely exploits information which, on the one hand, is *not* implicated in the compositional structure of the language and, on the other hand, *is* implicated in computing token-to-type relations. I don't, to put it mildly, know of any anti-modularity experiment that even *purports* to show this. What they do instead, according to the present construal, is demonstrate that the available experimental measures respond to a lot of stuff that isn't, in fact, part of sentence parsing.

I do think there eventually will be – indeed, has to be – some sort of significant convergence between, on the one hand, the notion of a lexical entry that grammar requires in order to explicate the compositionality and reverse compositionality of linguistic structures; and, on the other hand, the notion of lexical mastery that the theory of the speaker/hearer requires to account for parsing and learning. My guess is that language acquisition delivers shallow lexical entries consonant with reverse compositionality, and that parsing delivers correspondingly shallow structural descriptions consonant with assigning tokens to their types, and that just about everything else will turn out to be 'performance theory' in the invidious sense of that term. But for the psychology and the linguistics to fit together in that way would be for them to converge at a level of considerable abstractness; light years away from the current squabbles about how to account for the odd millisecond of 'semantic' priming.

Whether I'm right about all this only time will tell. But the methodological moral surely ought not to be in dispute. It's no use us all agreeing that what's in the lexicon is where the action is unless 'being in the lexicon' is *independently* defined. The only way I know to do that is to tie the constraints on lexical entries very closely to the conditions that compositionality and reverse compositionality impose, since that's the only way I know of to tie the understanding of hosts to the understanding of parts *biconditionally*. But if you do constrain the lexicon that way, it seems almost certain that most of what cognitve science blithely refers to as lexical effects in parsing and language learning aren't, in fact, mediated by information of the kind that lexical entries contain. Likewise, the appearance of an emerging consensus between psychologists and linguists as to the centrality of the lexicon in their respective disciplines begins to look a little spurious. That's quite a lot like what happened to pragmatics and semantics too, of course, when people started to take *them* seriously. So maybe it's a sign of progress.

Anyhow, let me conclude with a bona fide New York anecdote. My wife was one day in a hurry to get some dry cleaning done, so she brought it to a shop that had a large sign over the door saying '24 hour service'. 'I'll pick it up tomorrow,' she told the man at the desk. 'Not so,' he assured her, 'it won't be ready till next week.' 'But,' Janet protested, 'your sign says 24 hour service'. '*That's just the name*,' the man New Yorkishly replied. Well, likewise with the lexicon; I'm afraid it's not the solution of our psycholinguistic problems; it's just what we've recently gotten into the habit of calling them.

Semantics in the spin cycle

Competence and performance criteria for the creation of lexical entries

Amy Weinberg
Linguistics Department/UMIACS
University of Maryland

My brother is a practical man, with little patience for things philosophical.[1] Coward that I am, I was surreptitiously working on this commentary while visiting him in California, but to no avail. One morning, after glancing at my copy of Jerry Fodor's submission to this volume he came down with a big grin on his face prepared to ask me what dispute I had with a theory that claimed that 'cat' means cat and that one learned the meaning of 'gray' and 'cat' from seeing these words in expressions like 'gray cat'. This time I had my arguments (cribbed from Jerry's paper) ready. I told him that the nature of lexical representation had profound influence on the conduct and import of recent psycholinguistic experimentation. In mid flight (even in mid sentence) I realized though that I was not at all sure that this was true, or that Fodor's criteria would have many foes in the psycholinguistic community, or that work would progress any differently if his criteria were taken into account. In the rest of this commentary I would like to reflect on the import that adopting Fodor's criterion of *reverse compositionality* has for the field of lexical and sentence processing.

The point of these remarks is **not** to argue that these theories provide the correct account of sentence processing, but to look at whether these theories are ruled out by the adoption of Fodor's criterion.

What is 'Reverse Compositionality'?

Fodor gives the following definition of 'Reverse Compositionality' (RC):

> "Nothing belongs to the lexical entry for a lexical item except what that item contributes to the grammatical representation of its hosts" where host is defined as "any expression E… of which E is a constituent."[2]

RC places an upper bound on what can appear in a lexical entry. For example, it limits the semantic component of a word to semantic pieces that can be gleaned from the meaning of the host. Fodor's example of 'gray cat' limits the meaning of 'gray' and 'cat' to those elements that can be obtained by inspection of this and other phrases of which these words are component parts. For Fodor, this is an extremely stringent requirement. Lexical entries must be either "logically necessary properties of the thing an item applies to or disquotational."[3] So 'cat' can mean either 'feline four legged mammal' or it can simply denote a cat. In this outing, Fodor refrains from attacking definitions.[4] Instead, he tries to exclude many of psychology's favorite candidate structures for lexical entries. In this article, Fodor specifically excludes stereotypes, frequency information, and epistemic properties from lexical entries. Stereotypes are excluded because they are not always invoked in understanding the meaning of phrases. The stereotype of 'pet' might be 'dog' or 'cat', the stereotype of 'fish' might be 'trout' but 'pet fish' has a goldfish like stereotype that is not formed from the component stereotypes and since these don't contribute to the meaning of this phrase, they cannot be in the lexical entries for the component parts even if they contribute to the meaning of some other phrases of which they are components (like 'big pet' or 'big fish'). If a word contains a subcomponent, this piece must be part of the meanings of all the hosts that contain this word. No mixed theories of the type where certain subcomponents are part of some hosts, while not surfacing as components of the meanings of other hosts are allowed.

Fodor's justification comes from learnability considerations. "… Reverse compositionality is required to explain why everyone who has 'gray cat' in his idiolect has the entries for 'gray' and 'cat' in his lexicon."[5] We learn the meaning, syntactic or morphological features, or other characteristics of words or other linguistic components from seeing these words in phrasal or sentential contexts. Therefore components of meaning have to be derivable from inspection of these contexts. Simple as that.

RC and psycholinguistics

Except that Fodor would have us believe that it is not so simple for much philosophical and psycholinguistic work on lexical and sentence processing. Fodor claims that

> "... most of what cognitive science blithely refers to as lexical effects in parsing and language learning aren't in fact mediated by information of the kind that lexical entries contain..." and "... that language acquisition delivers shallow lexical entries consonant with reverse compositionality, and that parsing delivers correspondingly shallow lexical entries consonant with assigning tokens to their types, and that everything else will turn out to be 'performance theory'..."[6]

This distinction between acquisition and parsing crosses with another well known dichotomy between competence and performance.

In the standard generative tradition, a criterion like 'reverse compositionality' is naturally placed in the competence theory as it plays a role in explaining how an idealized speaker/hearer could come to know the meaning of lexical items from their occurrences in their hosts. Theories that enforce a competence/performance distinction typically do not enforce a one to one mapping between representations derived from competence and those used by the performance theory. All features of the representation must clearly be learnable but, additional features can be added to representations based on the actual details of acquisition, or the extra demands that processing places on these representations. In this case, given Fodor's assumptions about learning and the fact that grammatical features become part of static lexical representations, RC follows with respect to grammatical features. It is clear though, that the learning model doesn't prohibit other features from being part of the performance lexicon, provided they are also learnable. Recoding from the representation that is explanatory at the competence level can occur when viewing the same principle as an algorithm.[7]

Fodor does not cite specific examples of theories that handle psycholinguistic phenomena as "lexical effects" but it will be handy to have some cases available in order to see how enforcement of a reverse compositional view of the lexicon bears on the interpretation of experimental results.

As an example, consider Ford, Bresnan, and Kaplan's (1981) early use of frequency information to explain various PP attachment decisions. FBK assumed a theory where the lexical entries for verbs contained subcategorization information at the level of competence. Verbs had representations as in (1) at the competence level.

(1) a. put___NP PP
b. see___NP
c. know___NP
d. bring___NP (PP)

These representations were necessary to explain the fact that there is a non arbitrary association between verb choice and argument array. The PP in (2a) is an obligatory piece of the verb's syntactic environment. In (2b), it is an optional part of the argument structure.

(2) a. I put the dogs on the beach.
*I put the dogs.
b. I brought a toy from France.
I brought a toy.

In addition there was an algorithmic component to the theory that manipulated these representations during processing. This algorithm required that the initial choice for PP attachment was governed by the following algorithm (3).

(3) "If a set of alternatives has been reached in the expansion of a phrase structure rule, give priority to the alternatives that are coherent with the strongest lexical form of the predicate."[8]

(3) implicitly requires attachment as an argument if that argument is specified in the competence theory as part of the lexical representation. If no PP is specified as part of the argument structure, it requires attachment as an adjunct inside a complement NP. This explains the preferred attachment of the PP in (2a) as (4a), and the attachment of the same PP as an adjunct to the verb in (4b). Attachment as an adjunct allows the syntactic structure to conform to the argument array for 'see', which simply has a complement NP.

(4) a. I [$_{\text{VP}}$ put [$_{\text{NP}}$ the dogs] [$_{\text{PP}}$ on the beach]]
b. I [$_{\text{VP}}$ saw [$_{\text{NP}}$ the dogs [$_{\text{PP}}$ on the beach]]]

In order to handle cases like (5), FBK invoke frequency as part of a performance theory to be overlain on the representations provided by the competence theory. In addition to learning that different verbs have different subcategorization frames as required by the competence theory, native speakers of a language have the ability to count the number of occurrences in which an item occurs in a particular frame. This detail is abstracted away from in the competence theory, as it is not required to explain how the native speaker comes to associate a verb with its argument structures. It is represented at the level of performance by augmenting the lexical templates for verbs with the relative frequencies of occurrence for each subcategorization frame. This information is used by the parser at run time in order to pick its first choice for PP attachment in ambiguous cases where there is also a preference for adjunct attachment as part of the complement NP even though the PP is part of the argument structure of 'bring' bearing the source argument role.

(5) I brought the toy from France yesterday.

The PP will be attached so that the resulting structure is compatible with the most frequent lexically specified frame. This theory uses representations provided by the grammar in the form of subcategorization frames augmented by a theory of linguistic memory that claims that humans can count the number of occurrences of a particular frame. Fodor seems to want to bar frequency information even in the performance theory because "the relative frequency of a lexical item doesn't predict the relative frequency of its hosts or vice versa".[9] This is surely correct, but it is just as surely true that hosts can be decomposed into their component parts and the frequency with which a category occurs with a particular set of features overall can be tracked. This tracking has to be justified on extragrammatical grounds, but it is clearly learnable and no one doubts that frequency effects exist. Their inclusion in the lexicon implicates a perfectly sensible learning mechanism that is simply different from the one used to supply semantic features. The fact that language learners know the subcategorization frames associated with particular verbs (a fact about competence) is at right angles to whether or not they can count the number of times these frames occur (a fact about performance). Features like frequency of occurrence may be added in the performance model as they further specify how learning actually occurs (in this case by the intersection of the competence theory with a theory of linguistic memory). Similar remarks apply to the criterion of reverse compositionality. Reverse compositionality is placed in the lexicon by considerations of competence. Other features like frequency (and plausibility (see below)) can also be added as the performance theory's contribution to the lexical representation viewed from the level of the actual learning or parsing algorithm. It may be true that you can't learn that a particular semantic feature is part of a word, except by seeing it in context, justifying RC with respect to features. It is not true though that one cannot break hosts into their component parts and then separately track these frequencies.

Similar remarks apply to a more recent example of a "lexicalist theory" that invokes frequency. Trueswell et al (1993) have pointed out that initial interpretations of potential relative clause constructions can be mediated by particular lexical choices. For example, the underlined portion of (6) is strongly interpreted as a main clause leading to difficulty in interpreting the full sentence. By contrast, the underlined portion of (7) admits the relative clause interpretation, which is appropriate, making this sentence much easier to interpret.

(6) The witness examined by the lawyer was useless.

(7) The evidence examined by the lawyer was useless.

This difference is explained as a "lexical effect." The lexicon contains an entry for the verb 'examine' that specifies that it selects an agent theta role. The competence entry for this verb looks something like (8)

(8) examine: [________verb]
(agent, patient)

Somewhere there is also a statement that agents are generally animate. The standard representation for this in the competence grammar is some kind of redundancy rule that links the feature of agenthood to the feature of animacy. This expresses the generalization that crosslinguistically, there is a link between these two features. Turning this into a performance theory requires several decisions. One must decide whether this general redundancy rule is accessed online each time a verb that selects an agent is computed, or whether this link is precomputed and stored in the lexical entry yielding the entry in (9)

(9) examine: [________verb]
(agent, patient)
+animate

This decision will be made based on algorithmic considerations such as the speed with which this linking is made at runtime and whether one can show that storing the effect of the redundancy rule makes it quicker to access than actively computing it on each occasion, as is standardly assumed.

Another factor that may influence this decision is whether one can show that this linking is probabilistic, and that different verbs can enforce this linking to varying degrees. If this is true, then one may want to indicate the probability (or frequency) of this association as part of each lexical entry. We idealize at the competence level, but allow more nuanced algorithmic representations using primitives that are irrelevant to competence. This recoding reflects the contribution of the experience of individual speaker/hearers at the algorithmic level, and is standard in theories that assume both of these levels. Assuming that (8) is correct, the improved acceptability of (6) comes from the fact that inspection of the lexical entry for 'examine' now provides evidence against taking the inanimate 'evidence' as the subject of this verb, thus correctly favoring a relative clause interpretation.

This theory conforms to RC in several ways. First, at the competence level, its lexical entries only contain definition or other information that plays a role in the compositional structure of its grammatical host. Second, the non reverse compositional information (frequency with which two competence features are linked) is added as part of the theory of performance, as required by Fodor. This kind of frequency based lexical explanation proceeds as before, with or without the "reverse compositionality" criterion.

The previous two cases that we have discussed make use of features that could be stored as part of the definition of a lexical item as a semantic feature. What about the storage of so called "pragmatic or inferential" information as part of a lexical entry. One might think that this would be disallowed as part of processing, since surely pragmatic information does not apply across the board to all hosts for a particular lexical category. As Fodor says "semantic effects are practically ubiquitous in what are generally supposed to be experiments on how people parse; and these are paradigmatically the effects of plausibility and context..." Plausibility effects being neither definitional nor denotational cannot be in the lexicon according to Fodor because, if they were there would be "no hope for the idea that information parsing a language exploits is the very information that learning the grammar provides." Again though, we will see that this is a very weak criterion that excludes the use of pragmatic information only if one assumes that this information is stored as a feature on lexical entries, which is not generally assumed. To see this consider one theory allowing pragmatic factors to influence parsing as a "lexical effect" as discussed by MacDonald et al. (1994). These authors claim that context effects can influence various parsing decisions. We should be a bit more precise here because their claim is more radical, consisting of three parts which are:

a. words are composed of component features
b. lexical choice and disambiguation both involve dynamic selection of features
c. frequency plays a role both at the semantic and pragmatic level

(b) implies that there is really no such thing as a "lexical entry." Lexical entries are dynamically constructed by putting features together that are strongly associated. The features 'feline' and 'with tail' are strongly associated by being co-activated whenever one discusses cats. Context is seen as being decomposable into features as well. These features have a different status because they depend on particular situations though, and so are not strong enough to determine lexical selection, but may help to disambiguate. For example, the fact that a context indicates that one is talking about animals does not allow you to choose the word 'crow' in the following context:

I saw a ___ on the telephone pole.

This is because this ___ could be filled in by many animals. A definition however such as: A black bird smaller than a raven a ___, might allow you to fill in this blank, because these features are all strongly associated together by being activated whenever we hear the word 'crow'. Since this theory does not include static entries, it cannot distinguish between features that are or are not part of lexical entries.

Different feature pairings have different levels of constraint though, and in a theory that made a lexical/nonlexical distinction, definitions would be implemented separately from contextual cues. This interpretation is justified by the role that these authors give to these two sets of features during processing. The first set of definitional (now interpreted as high frequency constraints) is used in lexical choice, but context effects interact with these baseline frequencies for purposes of disambiguation. Context effects provide "an effective basis for deciding between a small number of alternatives... Moreover, the effects of contextual information are limited by lexical factors, specifically the (prior (ASW)) frequencies of the alternatives."[10] There are large unanswered questions of exactly which mechanisms are used by context features to influence lexical disambiguation, but it is clear that limiting a lexicon to features justified by RC is consistent with the use of context that MacDonald et al. want to make. Lexical effects govern initial choice and context can (perhaps in a nonmodular way) influence disambiguation.

Conclusion

Three things should be clear from the discussion above. Learning semantic features may require reverse compositionality, but that does not preclude other learning models for other types of features such as frequency. A competence performance distinction allows a distinction between features justified solely by considerations of learning, and features like frequency that are learnable but required only to explain facts about performance. Finally, there is no reason that lexical entries should not contain both types of information. Given this, the **Reverse Compositionality** criterion, while perhaps justified for semantic learning does not bar use of frequency or pragmatic information as it used by most current lexicalist theories. It is of course a much more ambitious enterprise to show whether these theories provide a correct account of processing phenomena,[11] but it should be clear that they are not ruled out by imposing RC.

Notes

[1] James Weinberg (personal communication).

[2] Fodor, J. (this volume), p. 76.

[3] Ibid, p. 80.

[4] See Fodor and Lepore (1997), for example.

[5] Fodor, J. (this volume), p. 78.

[6] Ibid, p. 83.

[7] See Berwick and Weinberg (1984).

[8] Ford, M., Bresnan, J. and Kaplan, R. (1981).

[9] Fodor, op. cit., p. 80.

[10] MacDonald et al. (1994) pg. 697.

[11] See Weinberg (2000), for example for criticism of a purely lexicalist approach.

References

Berwick, R.C. and Weinberg, A. (1984). *The Grammatical Basis of Linguistic Performance*. MIT Press.

Fodor, J. “The lexicon and the laundromat.” (this volume).

Ford, M., Bresnan, J. and Kaplan, R. “A Competence based Theory of Syntactic Closure.” In J. Bresnan, *The Mental Representation of Grammatical Relations*. MIT Press.

MacDonald, M.E., Pearlmutter, N. and Seidenberg, M. “Lexical Nature of Syntactic Ambiguity Resolution.” *Psychological Review*, pp. 676–703.

Trueswell, J.C., 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 Cognition*, pp. 528–553.

Weinberg, A.S. (2000). “Minimalist Parsing.” In Epstein, S. and Hornstein, N. (eds.) *Working Minimalism*. MIT Press.

Connectionist and symbolist sentence processing

Mark Steedman
University of Edinburgh

This chapter argues that claims for recurrent networks as plausible models of human sentence processing from which generalizations that have seemed to require the mediation of symbolically represented grammars are "emergent" are misplaced. Rather, it is argued that the linguistic relevance of connectionist networks lies in their application to lexicalist grammar induction.

1. What do SRNs compute?

The Simple Recurrent Network (SRN, Elman 1990, 1995) has recently received quite a bit of attention as a mechanism for modelling human natural language processing. The device approximates the more costly but exact "backpropagation through time" algorithm of Rumelhart, Hinton, & McClelland (1986) for learning sequential dependencies. It does so by using a single set of context units which store the activations of the hidden units at time $t-1$ as an input to the hidden units at time t, along with the activations of the normal input units corresponding to the current item, $item_t$, as in Figure 1.

Since the activations of the hidden units at time $t-1$ were themselves partly determined by the activations on the hidden units at time $t-2$, which were in turn determined by those at time $t-3$, and so on, the context units carry ever-diminishing echoes of the items in the preceding sequence.

Because there is no clear bound to the extent of the preceding sequence about which information can be captured in the context units, it is not entirely clear what the precise automata-theoretic power of such "graded state machines" is (see Cleeremans 1993; Cleeremans, Servan-Schreiber, & McClelland 1995; Casey 1996). However, the signal-to-noise ratio for information concerning distant items falls off very rapidly with this mechanism, and it is fairly clear that in practice SRNs of the kind that can actually be built and

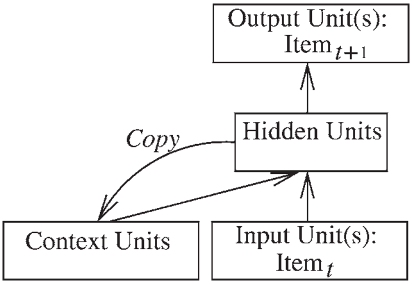


Figure 1. Simply Recurrent Network (SRN).

trained end up approximating the class of Finite State Markov Machines that can be learned using the exact technique of Rumelhart et al. (1986) to a degree of accuracy that depends on the maximum number of timesteps required.

Finite state machines are interesting devices, and it is often surprising to see the extent to which they can approximate the output of devices that are intrinsically higher on the automata-theoretic hierarchy. For example, there have recently been some startling demonstrations that it is possible to automatically induce large finite state machines that sufficiently capture the redundancies characteristic of instructional text and that their measure of the similarity of student essays to the original text are highly correlated with the grades assigned to the essays by human graders – as in the Latent Semantic Analysis (LSA) approach of Landauer, Laham, Rehder, & Schreiner (1997). It is interesting to ask whether similar mechanisms play any part in natural language processing.

Such results are possible because some neural network algorithms are capable of inducing extremely efficient – and correspondingly opaque – representations, when compared with standard Hidden Markov Models (HMMs – see Williams & Hinton 1990 on this point). However, as SRNs are actually used by psychologists and linguists they appear to approximate something much closer to a familiar standard symbolist finite-state device, namely the *n*-gram part-of-speech (POS) tagger. (This also seems to be their role in "hybrid architecture" connectionist parsers such as those proposed by Mikkulainen 1993, which combine them with a push-down stack and structure-building modules.)

2. Finite-state part-of-speech tagging

N-gram POS tagging – that is, the determination of the form-class of ambiguous lexical items like *bear* on the basis of sequential probabilities at the word level can be remarkably successful. Accuracy over 97% is quoted (Merialdo 1994).

Such results need to be put in context. 91% accuracy can be achieved on the basis of unigram frequency alone, and both in theory and in practice, accuracy of 97% implies that only half the sentences in texts such as the Wall Street Journal will be without error if only the most likely candidate is chosen (see Church & Mercer 1993; Ratnaparkhi 1998). If more than one candidate is allowed when the top candidates are close in probability, then the likelhood that the correct category will be among them goes up to 99.9% for an average set size of around 1.3 categories per word (De Marcken 1990; Elworthy 1994; Carroll & Briscoe 1996). In either case, more work needs to be done. But such techniques can massively reduce the degree of nondeterminism that practical parsing algorithms must cope with.

Moreover, there is growing evidence that if the standard Brown Corpus POS categories like VB are replaced with more informative lexical categories like Tree Adjoining Grammar (TAG, Srinivas and Joshi 1999) elementary trees or Combinatory Categorial Grammar (CCG Steedman 2000) categories, and if different senses of the same syntactic type are also distinguished as different grammatical categories, this effect may do a great deal of the work of parsing itself, leaving only structural or "attachment" ambiguities to be resolved by parsing proper (see B. Srinivas 1997; Kim, B. Srinivas, & Trueswell, this volume). This is not to imply that the resolution of attachment ambiguities is trivial: it may in fact be exponentially complex in the worst case.

2.1 Is part-of-speech tagging "psychologically real"?

Despite the real practical success of finite-state POS tagging, none of the computational linguists who have built such devices seems ever to have claimed that predictive POS tagging corresponds to a component of human processing, any more than the LSA discussed in Landauer et al. (1997) corresponds to a component of human essay writing and (one hopes) grading. The idea that human sentence processors deal with lexical ambiguity by anything like the predictive POS tagging implicit in SRNs based on global statistical characteristics of large volumes of text seems quite unlikely on a number of counts, despite the claims of Corley & Crocker (1996) and Kim et al. (this volume).

First, the sequential properties that are typically discovered by HMM taggers and by implication SRNs vary widely across different types of text, so that reliability degrades rapidly with genre changes. Rule-based POS disambiguators more closely linked to syntax proper, of the type discussed by Brill (1992), Cussens (1997), and Voutilainen (1995), or parsers that integrate probabilistic categorial disambiguation more closely with the grammar, may be more resistant to this effect.

Second, experiments like those of Swinney (1979) and Tanenhaus, Leiman, & Seidenberg (1979) showing early transient activation of irrelevant lexical alternatives, including irrelevant syntactic types, which are only subsequently eliminated as an effect of prior biasing extra-sentential context seem hard to reconcile with n-gram POS taggers or any other mechanism based on local string context. While related low-level probabilistic mechanisms using a larger or even an unbounded string context could in principle define a probability model that could be used to adjust the activation of the category set in keeping with Swinney's results, such models are rather different from the SRN, and have problems of their own (such as sparseness of data). Mechanisms that adjudicate between alternatives on the basis of more dynamic and transient properties of the text and the context seem to be needed.

In particular, experiments by Marslen-Wilson et al. (1978), Marslen-Wilson & Tyler (1980), Crain (1980), Altmann (1985), van Berkum et al. (1999) and the present author, showing immediate effects of inferential plausibility and referential context on parsing decisions, are hard to reconcile with any mechanism based on *a priori* preferences based on global sequential probabilities of text.

In fact, it seems likely that the real interest of reentrant networks may lie elsewhere. To see why this might be, we should ask ourselves why finite state devices work as well as they do.

2.2 Why do SRNs and part-of-speech taggers work?

Finite-state POS taggers, and by assumption SRNs, work reasonably well on tasks like category- and sense- disambiguation and prediction of succeeding category because the implicit Markov processes encode a lot of the redundancy (in the information-theoretic sense of the term) that is implicit in grammar, interpretation, and world-knowledge. For example, the SVO word-order of English and the way the world actually is between them determines the fact that the transitive category for the word *arrested* is more likely to follow the noun *cop* than the past participial category, while these preferences are reversed following the word *crook*.

(1) a. The cop arrested by the detective was guilty.
    b. The crook arrested by the detective was guilty.

This means that, like standard Markov processes, SRNs can be made the basis of quite good predictors of processing difficulty – as measured by increased reading times at the word *by* in (a) as compared with (b), for example.

Because the context defined by the context units is in a limited sense unbounded, SRNs can at least in theory be used to model long distance agreement dependencies (Elman 1990; Christiansen & Chater 1994), although because of the properties of the context unit representation, reliability falls off with distance, and these dependencies cannot be regarded as unbounded in the technical grammatical sense.

2.3 Are grammars emergent from SRNs?

While claims exist in the literature to the effect that recognisers for string sets of kinds that in general require grammars of higher than finite-state power have been acquired by SRNs (Christiansen 1994), none of these results suggests that the grammars in question are therefore "emergent" properties of mechanisms like SRN, any more than they are of n-gram POS taggers. Even error-free sequences of grammatical categories fall short of semantic interpretability, as can be seen from the fact that the following sequence has *two* interpretations based on identical Brown corpus categories:

(2) Put the block in the box on the table.

Although SRNs can be regarded as disambiguating lexical items, this other kind of ambiguity – structural or attachment ambiguity – remains, just as it does for the symbolist TAG-categorial and CCG categorial disambiguators of B. Srinivas (1997) and B. Srinivas & Joshi (1999) (although such extended category sets can have the effect of trading attachment ambiguity for categorial ambiguity, further simplifying but not eliminating the task of the parser). The same is true for the "shallow" parser hypothesised by Fodor in this volume, which still has to conduct search to identify the linguistic type of which the utterance to hand is a token.

For the same reason, it does not seem sound to regard "trajectories" through the high-dimensional space defined by the hidden units by sequences of words as the equivalent of parses or interpretation (Elman, 1995; Tabor et al. 1997).

Many other defining properties of interpretations – such as the ability to do the kind of structure-dependent transformations characteristic of inference – seem to be lacking in trajectories or category sequences of this kind.

3. Interpretable structure and associative memory

Much connectionist work has explicitly or implicitly taken on board the need for explicitly representing the equivalent of trees or pointer structures to represent syntactic or semantic analyses (see papers in Hinton 1990a), using associative memories of various kinds.

Such devices are of interest for (at least) two reasons: First, they inherit some psychologically desirable properties of distributed representations, such as content-addressability and graceful degradation under noise and damage. Second, they offer a way to think about the interface between neurally embedded map-like sensory-motor inputs and outputs, and symbolic knowledge representation, in the following sense.

3.1 The Associative Net

One very simple early associative memory model for pointers in structures is the Associative (a.k.a Willshaw) Net (Willshaw et al. 1969; Willshaw 1981), Figure 2.

Pointers can be represented as associations between addresses represented as binary vectors. To store an association between the input vector on the left

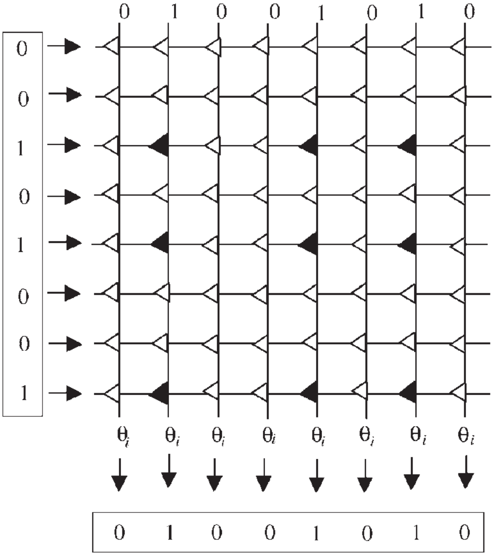

0 1 0 0 1 0 1 0
0
0
1
0
1
0
0
1
θ_i θ_i θ_i θ_i θ_i θ_i θ_i θ_i
0 1 0 0 1 0 1 0

Figure 2. The Associative Net storing one pointer.

and an associate vector, the associate is input along the top, and switches are turned on (black triangles) at the intersection of lines which correspond to a 1 in both patterns. To retrieve the associate from the input, a signal is sent along each horizontal line corresponding to a 1 in the input. When such an input signal encounters an "on" switch, it increments the signal on the corresponding output line by one unit. These lines are then thresholded at a level corresponding to the number of on-bits in the input, to yield the associated vector as the output at the bottom.

With such thresholding, an associative memory can store a number of associations in a distributed fashion, with interesting properties of graceful degradation in the face of noise and ablation. If two (or more) devices of this kind share input lines, binary (or n-ary) trees and other graphs can be represented – see figure 3.

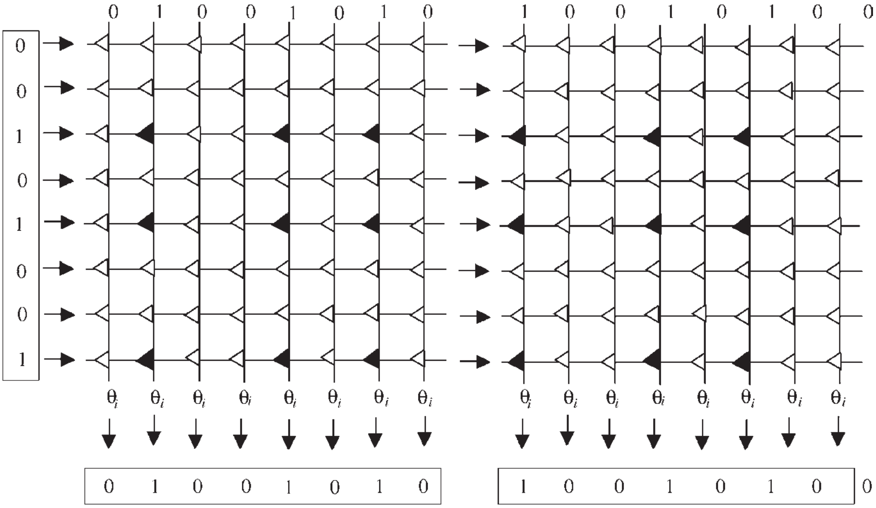


Figure 3. The Associative Net storing one binary node.

3.2 Recursive Auto-Associative Memory (RAAM)

Recursive Auto-Associative Memory (RAAM, Pollack 1990) is a related device that uses hidden unit activation patterns in place of the Willshaw net's sparse matrix (figure 2). It is called "auto-associative" because it uses the same patterns as input and output

An n-ary recursive structure can be stored bottom-up in the RAAM starting with the leaf elements by recursively auto-associating vectors comprising up to *n* hidden-unit activation patterns that resulted from encoding their daughters. The activation pattern that results from each auto-association of the daughters can then be treated as the address of the parent.

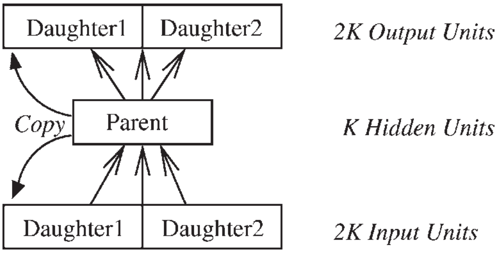


Figure 4. Recursive Auto-Associative Memory (RAAM).

Since by including finitely many further units on the input and output layers we can associate node-label or content information with the nodes, a modification sometimes referred to as Labelled RAAM (LRAAM), this device can store recursive parse structures, thematic representations, or other varieties of logical form of sentences.

The device should not be confused with a parser: it is trained with fully articulated structures which it merely efficiently stores. The hidden units can be regarded as encoding to some approximation the context-free productions that defined those structures, in a fashion similar to the way Hinton (1990b) encoded part-whole relations. In that sense the device has been claimed to be capable of inducing the corresponding grammar from the trees (Pollack 1990).

This is probably a more appropriate use for RAAM than building parse trees, since RAAM is slow to train, and inherits poor scaling properties from its use of backpropagation. Devices more closely related to Willshaw nets, such as the Holographic Reduced Representations (HRR) proposed by Plate (1994) are an interesting alternative. Their properties for the storage and holistic transformation of such structures has been investigated by Neuman (2000a, b).

4. Associative memory and the lexicon

4.1 Using classifiers to learn categories

One use of associative memory might be to learn the bounded structures that are associated with lexical items, particularly verbs.

We might assume that a subset of such structures are available prelinguistically, and result relatively directly from the evolved or learned structure of connections to the sensorium, short term memory, and the like. At higher levels, such structure may arise from non-linguistic network concept-learning along lines set out by Hinton (1990b), without mediating symbolic forms.

One use might be for learning lexical grammars in the form of CCG lexical categories or the elementary trees of a lexicalized TAG grammar. Part of the interest of this proposal lies in the possibility that the interaction of such structured sensory-motor manifolds and this novel form of concept learning might give rise to the kind of shallowly grounded categories that Fodor argues for in the present volume within a standard symbolist approach. (Fodor's argument centers exclusively on the conceptually primitive nature of most nouns, and many linguists have noted that verbs seem more susceptible to analysis in terms of underlying structure involving causative elements and the like. While Fodor (1975) has argued against the idea that the *sentences* "He killed the dog" and "He caused the dog to die" are structurally equivalent, these arguments seem entirely consistent with decomposition of concepts like *kill* within the lexicon.)

4.2 Lexicalized grammar

The advantage of such theories lies in a closer integration of the lexicon, syntax, semantics, and phonology including intonation In CCG, each word and constituent is associated with a directional syntactic type, a logical form, and a phonological type. "Combinatory" syntactic rules combine such entities to produce not only standard constituents associated with the same three components, such as predicates or VPs, but also non-standard constituents corresponding to substrings such as *Anna married*.

The latter are involved in phenomena such as coordination and intonational phrasing, as in (3) and (4) (in which % marks an intonational boundary marked by a rise and/or lengthening, and capitals indicate pitch accent).

(3) Mary loved, and Anna married, Manny.

(4) Q: I know that Mary married DEXTER, but who did ANNA marry?
A: ANNA married% MANNY

Such non-standard constituents may also be involved in the fine-grain incremental interpretation by the processor and its use (referred to earlier in the discussion of experiments by Crain and Altmann), under a strict version of the competence hypothesis, according to which the processor is only allowed access to the categories and interpretations that are defined by the competence grammar.

4.3 Learning lexicalized grammars with networks

Within such frameworks, grammar acquisition mainly reduces to decisions such as whether the syntactic type corresponding to the *walking* concept looks for its subject to the left or to the right in the particular language that the child is faced with – in CCG terms, whether it is $S\backslash NP$ or S/NP – and to how the multiple arguments of transitives, ditransitives, and the like map onto the underlying universal logical form, as reflected for example in the possibilities for reflexivization. Since directionality can be represented as a value on an input unit, and since the categories themselves can be defined as a finite state machine, and their relation to universal logical form as a finite state transduction, such categories are good candidates for learning with neurally computational devices.

A similar tendency towards lexical involvement is evident in current statistical computational linguistic research. Much recent work in probabilistic parsing including proposals by Jelinek et al. (1994), Magerman (1995), Collins (1997) and Charniak (1997) moves away from autonomous Markovian POS tagging and prefiltering, and towards a greater integration of probabilities with grammar – see Manning & Schütze (1999) for a review.

Part of the interest of this proposal lies in the possibility that such learning might capture word-order generalizations over the lexical categories, a point that has been made by Christiansen & Devlin (1997). Constraints such as that semantically related categories (such as tensed transitive verbs) have the same directionality (such as SVO order) are "soft" constraints, which can have exceptions (such as English auxiliary verbs), have been discussed within Optimality Theory (Prince & Smolensky 1997). Since most Optimality-Theoretic constraint systems appear to be equivalent to Finite State Transducers (Eisner 1997), it seems likely that the associative memory-based lexical acquisition device sketched above might also be suited to acquiring such soft-constraint-based lexicons. If so, then the claim that the form of possible human lexicons was "emergent" from the neural mechanism would have some force.

Acknowledgements

This paper was first presented to the 1999 CUNY Natural Language Processing Conference. Thanks for comments to the participants and to the editors and referees for this volume, The research was supported in part by ESRC Award Number M/423/28/4002, EPSRC grant GR/M96889 and EU IST grant MAGICSTER.

References

Altmann, G. (1985). Reference and the resolution of local syntactic ambiguity. Ph.D. thesis, University of Edinburgh.

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191–238.

van Berkum, J., Brown, C., & Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language, 41*, 147–182.

Brill, E. (1992). A simple rule-based part-of-speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, Trento*, pp. 152–155, San Francisco, CA. Morgan Kaufmann.

Carroll, J. & Briscoe, E. (1996). Apportioning development effort in a probabilistic lr parsing system through evaluation. In *Proceedings of the 2nd ACL SIGDAT Conference on Empirical Methods in Natural Language Processing, Philadelphia PA*, pp. 92–100.

Casey, M. (1996). The dynamics of discrete-time computation, with application to recurrent neural networks and finite state machine extraction. *Neural Computation, 8*, 1135–1178.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the 14th National Conference of the American Association for Artificial Intelligence, Providence, RI., July*, pp. 598–603.

Christiansen, M. (1994). Infinite languages, finite minds: Connectionism, learning and linguistic structure. Ph.D. thesis, University of Edinburgh.

Christiansen, M. & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language, 9*, 273–287.

Christiansen, M. & Devlin, J. (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th Annual Cognitive Science Society Conference*, pp. 113–118, Mahwah, NJ. Lawrence Erlbaum Associates.

Church, K. & Mercer, R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics, 19*, 1–24.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning*. Cambridge MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. (1995). Graded state machines: The representation of temporal contingencies in feedback. In Y. Chauvin & D.E. Rumelhart (eds.), *Backpropagation: Theory, Architectures, and Applications*, Developments in connectionist theory, pp. 274–312. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, M. (1997). Three generative lexicalized models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid*, pp. 16–23, San Francisco, CA. Morgan Kaufmann.

Corley, S. & Crocker, M. (1996). Evidence for a tagging model of human lexical disambiguation. In *Proceedings of the 18th Annual Meeting of the Cognitive Science Society, San Diego CA*.

Crain, S. (1980). Pragmatic constraints on sentence comprehension. Ph.D. thesis, University of California, Irvine.

Crain, S. & Steedman, M. (1985). On not being led up the garden path: the use of context by the psychological parser. In L.K. David Dowty & A. Zwicky (eds.), *Natural Language Parsing: Psychological, Computational and Theoretical Perspectives*, pp. 320–358. Cambridge: Cambridge University Press.

Cussens, J. (1997). Part-of-speech tagging using progol. In N. Lavrac & S. Dzeroski (eds.), *Inductive Logic Programming: Proceedings of the 7th International Workshop (ILP-97)*, vol. 1297 of *Lecture Notes in Artificial Intelligence*, pp. 93–108. Berlin, Springer.

Eisner, J. (1997). Efficient generation in primitive optimality theory. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Association for Computational Linguistics, Madrid*, pp. 313–320, San Francisco, CA. Morgan Kaufmann.

Elman, J. (1990). Representation and structure in connectionist models. In G. Altmann (ed.), *Cognitive Models of Speech Processing*, pp. 345–382. Cambridge, MA: MIT Press.

Elman, J. (1995). Language as a dynamical system. In R. Port & T. van Gelder (eds.), *Mind as Motion*, pp. 195–225. Cambridge, MA: MIT Press.

Elworthy, D. (1994). Automatic error detection in part of speech tagging. In *Conference on New Methods in Language Processing, Manchester*.

Fodor, J. (1975). *The Language of Thought*. Cambridge, MA: Harvard.

Hinton, G. (ed.) (1990a). *Connectionist Symbol Processing*. Cambridge, MA: MIT Press/ Elsevier. Reprint of *Artificial Intelligence, 46*, 1–2.

Hinton, G. (1990b). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence, 46*, 47–75. Reprinted in Hinton (1990a).

Jelinek, F., Lafferty, J., Magerman, D., Mercer, R., Ratnaparkhi, A., & Roukos, S. (1994). Decision tree parsing using a hidden derivation model. In *Proceedings of the 1994 DARPA Human Language Technology Workshop*, pp. 272–277, San Francisco, CA. Morgan Kaufmann.

Kim, A., Srinivas, B. & Trueswell, J. A computational model of the grammatical aspects of word recognition as supertagging. In P. Merlo & S. Stevenson (eds.), *The Lexical basis of sentence Processing: Formal, computational and experimental issues*. Amsterdam: John Benjamins.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? In *Proceedings of the Nineteenth Conference of the Cognitive Science Society*, pp. 412–417, Mahwah, NJ. Lawrence Erlbaum Associates.

Magerman, D. (1995). Statistical decision tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge MA*, pp. 276–283, San Francisco, CA. Morgan Kaufmann.

Manning, C. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

de Marcken, C. (1990). Parsing the lob corpus. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics, Pittsburgh*, pp. 243–251, San Francisco, CA. Morgan Kaufmann.

Marslen-Wilson, W. & Tyler, L. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1–74.

Marslen-Wilson, W., Tyler, L.K., & Seidenberg, M. (1978). Sentence processing and the clause boundary. In W.J.M. Levelt & G.F. d'Arcais (eds.), *Studies in the Perception of Language*, pp. 219–246. New York: Wiley.

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics, 20*, 155–171.

Mikkulainen, R. (1993). *Subsymbolic Natural Language Processing*. Cambridge, MA: MIT Press.

Neuman, J. (2001a). Holistic processing of hierarchical structures in connectionist networks. Ph.D. thesis, University of Edinburgh.

Neuman, J. (2001b). Learning holistic transformations of HRR from examples. *Cognitive Systems Research*, to appear.

Plate, T. (1994). Distributed representations and nested compositional structure. Ph.D. thesis, University of Toronto.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence, 46*, 77–105. Reprinted in Hinton (1990a).

Prince, A. & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*, 1604–1610.

Ratnaparkhi, A. (1998). Maximum entropy models for natural language ambiguity resolution. Ph.D. thesis, University of Pennsylvania.

Rumelhart, D., Hinton, G., & McClelland, J. (1986). A general framework for parallel distributed processing. In D. Rumelhart, J. McClelland, & the PDP Research Group (eds.), *Parallel Distributed Processing,* Vol. 1, *Foundations*, pp. 45–76. Cambridge, MA: MIT Press.

Srinivas, B. (1997). Complexity of lexical descriptions and its relevance to partial parsing. Ph.D. thesis, University of Pennsylvania. IRCS Report 97–10.

Srinivas, B. & Joshi, A. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics, 25*(2), 237–265.

Steedman, M. (2000). *The Syntactic Process*. Cambridge, MA: MIT Press.

Swinney, D. (1979). Lexical access during sentence comprehension: (re)considerations of context effects. *Journal of Verbal Learning and Behaviour, 18*, 645–659.

Tabor, W., Juliano, C., & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12*, 211–271.

Tanenhaus, M., Leiman, J., & Seidenberg, M. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic context. *Journal of Verbal Learning and Behaviour, 18*, 427–441.

Voutilainen, A. (1995). A syntax-based part-of-speech analyser. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, Dublin*, pp. 157–164. San Francisco, CA. Morgan Kaufmann.

Williams, C. & Hinton, G. (1990). Mean field networks that learn to discriminate temporally distorted strings. In D. Touretsky, J. Elman, T. Sejnowski, & G. Hinton (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, pp. 18–22.

Willshaw, D. (1981). Holography, association and induction. In G. Hinton & J. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 83–104. Hillsdale, NJ: Erlbaum.

Willshaw, D., Buneman, P., & Longuet-Higgins, C. (1969). Non-holographic associative memory. *Nature, 222*, 960–962.

A computational model of the grammatical aspects of word recognition as supertagging

Albert E. Kim, Bangalore Srinivas and John C. Trueswell
University of Washington / AT&T Labs-Research / University of Pennsylvania

We describe a Constraint-Based Lexicalist model of human sentence processing. Highlighting a convergence of developments in multiple fields toward lexicalist and statistical processing perspectives, we argue that much of the syntactic ambiguity of language can be understood as lexical ambiguity, which is resolved during word recognition. The model is a connectionist system, which acquires wide coverage grammatical knowledge from supervised training on highly variable, naturally occurring text. The model learns to map each of the words in a sentence to an elementary tree from Lexicalized Tree Adjoining Grammar (Joshi & Schabes, 1996). These elementary trees are rich in grammatical information, encoding, among other things, the number and type of complements taken by a verb. The syntactic richness of these lexical representations results in substantial lexico-syntactic ambiguity. At the same time, statistical mechanisms for lexical ambiguity resolution are shown to effectively resolve this ambiguity. Simulations show that the model accounts for previously reported patterns in human sentence processing, including frequency-shaped processing of verb subcategory (e.g., Juliano & Tanenhaus, 1994) and effects of subtle contextual cues in lexical category ambiguity resolution (e.g., MacDonald, 1993).

In the last fifteen years, there has been a striking convergence of perspectives in the fields of linguistics, computational linguistics, and psycholinguistics regarding the representation and processing of grammatical information. First, the lexicon has played an increasingly important role in the representation of the syntactic aspects of language. This is exemplified by the rise of grammatical formalisms that assign a central role to the lexicon for characterizing syntactic forms, e.g., LFG (Bresnan & Kaplan, 1982), HPSG (Pollard & Sag, 1994), CCG (Steedman, 1996), Lexicon-Grammars (Gross, 1984), LTAG (Joshi &

Schabes, 1996), Link Grammars (Sleator & Temperley, 1991) and the Minimalist Program within GB (Chomsky, 1995). Second, theories of language processing have seen a shift away from "rule-governed" approaches for grammatical decision-making toward statistical and constraint-based approaches. In psycholinguistics, this has been characterized by a strong interest in connectionist and activation-based models (e.g., Lewis, 1993; McRae, Spivey-Knowlton & Tanenhaus, 1998; Stevenson, 1994; Tabor, Juliano & Tanenhaus, 1996). In computational linguistics, this is found in the explosion of work with stochastic approaches to structural processing (cf. Church & Mercer, 1993). In linguistics, this interest is most apparent in the development of Optimality Theory (Prince & Smolensky, 1997).

In this chapter, we highlight how the shift to lexical and statistical approaches has affected theories of sentence parsing in both psycholinguistics and computational linguistics. We present an integration of ideas developed across these two disciplines, which builds upon a specific proposal from each. Within psycholinguistics, we discuss the development of the Constraint-Based Lexicalist (CBL) theory of sentence processing (MacDonald, Pearlmutter & Seidenberg, 1994; Trueswell & Tanenhaus, 1994). Within computational linguistics, we discuss the development of statistical approaches to processing Lexicalized Tree-Adjoining Grammar (LTAG, Joshi & Schabes, 1996). Finally, we provide a description of the CBL theory, which is based on LTAG.

A constraint-based theory of sentence processing

Psycholinguistic thinking about the syntactic aspects of language comprehension has been deeply influenced by theories that assign a privileged role to supra-lexical syntactic representations and processes. This view has been most extensively developed in the theory of Frazier (1979, 1989), which proposed that syntactic processing is controlled by a two-stage system. In the first stage, a single syntactic representation of the input is computed using a limited set of phrase structure rules and basic grammatical category information about words. When syntactic knowledge ambiguously allows multiple analyses of the input, a single analysis is selected using a small set of structure-based processing strategies. In a second stage of processing, the output of this structure-building stage is integrated with and checked against lexically specific knowledge and contextual information, and initial analyses are revised if necessary. The basic proposal of this theory – that syntactic processing is, at least in the earliest stages, independent from lexically specific and contextual influences –

has been one of the dominant ideas of sentence processing theory (e.g., Ferreira & Clifton, 1986; Perfetti, 1990; Mitchell, 1987, 1989; Rayner, Carlson & Frazier, 1983).

A diverse group of recent theories has challenged this two-stage structure-building paradigm by implicating some combination of lexical and contextual constraints and probabilistic processing mechanisms in the earliest stages of syntactic processing (Crocker, 1994; Corley & Crocker, 1996; Ford, Bresnan & Kaplan, 1982; Gibson, 1998; Jurafsky, 1996; MacDonald et al., 1994; Pritchett, 1992; Stevenson, 1994; Trueswell & Tanenhaus, 1994). We focus in this chapter on the body of work known as the Constraint-Based Lexicalist theory (MacDonald et al., 1994; Trueswell & Tanenhaus, 1994), which proposes that all aspects of language comprehension, including the syntactic aspects, are better described as the result of pattern recognition processes than the application of structure building rules. Word recognition is proposed to include the activation of rich grammatical structures (e.g., verb argument structures), which play a critical role in supporting the semantic interpretation of the sentence. These structures are activated in a pattern shaped by frequency, with grammatically ambiguous words causing the temporary activation of multiple structures. The selection of the appropriate structure for each word, given the context, accomplishes much of the work of syntactic analysis. That is, much of the syntactic ambiguity in language is proposed to stem directly from lexical ambiguity and to be resolved during word recognition.[1] The theory predicts that initial parsing preferences are guided by these grammatical aspects of word recognition.

The CBL framework can be illustrated by considering the role of verb argument structure in the processing of syntactic ambiguities like the Noun Phrase/Sentence Complement (NP/S) ambiguity in sentences like (1a) and (1b).

(1) a. *The chef forgot the recipe was in the back of the book.*
b. *The chef claimed the recipe was in the back of the book.*

In (1a), a temporary ambiguity arises in the relationship between the noun phrase *the recipe* and the verb *forgot*. Due to the argument structure possibilities for *forgot*, the noun phrase could be a direct object or the subject of a sentence complement. In sentences like this, readers show an initial preference for the direct object interpretation of the ambiguous noun phrase, resulting in increased reading times at the disambiguating region *was in...* (e.g., Holmes, Stowe & Cupples, 1989; Ferreira & Henderson, 1990; Rayner & Frazier, 1987). On the CBL theory, the direct object preference in 1a is due to the lexical representation of the verb *forgot*, which has a strong tendency to take a direct object rather than a sentence complement. The CBL theory proposes that word recognition includes the activation of not only semantic and phonological representations of a word, but also detailed syntactic representations. These lexico-syntactic representations, and the processes by which they are activated, are proposed to play critical roles in the combinatory commitments of language comprehension. The direct object preference should therefore be eliminated when the verb *forgot* is replaced with a verb like *claimed*, which has a strong tendency to take a sentence complement rather than a direct object. These predictions have been confirmed experimentally (Trueswell, Tanenhaus & Kello, 1993; Garnsey, Pearlmutter, Myers & Lotocky, 1997), and connectionist models have captured these preferences (Juliano & Tanenhaus, 1994; Tabor et al., 1996).

Experimental work has also indicated that the pattern of processing commitments is not determined solely by individual lexical preferences, but involves an interaction between argument structure preference and lexical frequency. NP-biased verbs result in strong direct object commitments regardless of the lexical frequency of the verb. S-bias verbs, on the other hand, show an effect of frequency, with high frequency items resulting in strong S-complement commitments and low frequency items resulting in much weaker S-complement commitments (Juliano & Tanenhaus, 1993; though see Garnsey et al., 1997). This interaction between frequency and structural preference is explained by Juliano & Tanenhaus (1993) as occurring because the argument structure preferences of S-bias verbs must compete for activation with the regular pattern of the language – that an NP after a verb is a direct object. The ability of the S-bias verbs to overcome this competing cue depends upon frequency. Juliano & Tanenhaus (1994) present a connectionist model that shows that such interactions emerge naturally from constraint-based lexicalist models, since the models learn to represent more accurately the preferences of high frequency items. In later sections, we return to the issue of interactions between lexical frequency and "regularity" and discuss its implications for the architecture of computational models of language processing.

The CBL theory has provided an account for experimental results involving a wide range of syntactic ambiguities (e.g., Boland, Tanenhaus, Garnsey & Carlson, 1995; Garnsey et al., 1997; Juliano & Tanenhaus, 1993; Trueswell & Kim, 1998; MacDonald, 1993, 1994; Spivey-Knowlton & Sedivy, 1995; Trueswell et al., 1993; Trueswell, Tanenhaus & Garnsey, 1994; cf. MacDonald et al., 1994). As this body of experimental results has grown, there has been a need to expand the grammatical coverage of computational modeling work to match that of the most comprehensive descriptions of the CBL theory, which have been wide in scope, but have not been computationally explicit (MacDonald et al., 1994; Trueswell & Tanenhaus, 1994). Existing computational models have focused on providing detailed constraint-based accounts of the pattern of processing preferences for particular sets of experimental results (McRae et al., 1998; Tabor et al., 1996; Spivey-Knowlton, 1996; Juliano & Tanenhaus, 1994). These models have tended to be limited syntactic processors, with each model addressing the data surrounding a small range of syntactic ambiguities (e.g., the NP/S ambiguity). This targeted approach has left open some questions about how CBL-based models "scale up" to more complicated grammatical tasks and more comprehensive samples of the language. For instance, the Juliano & Tanenhaus model learns to assign seven different verb complement types based on co-occurrence information about a set of less than 200 words. The full language involves a much greater number of syntactic possibilities and more complicated co-occurrence relationships. It is possible that the complexities of computing the fine-grained statistical relationships of the full language may be qualitatively greater than in these simple domains, or even intractable (Mitchell, Cuetos, Corley & Brysbaert, 1995). It is also possible that these targeted models are so tightly focused on specific sets of experimental data that they have acquired parameter settings that are inconsistent with other data (see Frazier, 1995). Thus, there is a need to examine whether the principles of the theory support a model that provides comprehensive syntactic coverage of the language but which still predicts fine-grained patterns of argument structure availability.

Lexicalized grammars and supertagging

In developing a broader and more formal account of psycholinguistic findings, we have drawn insights from work on statistical techniques for processing over LTAG (Srinivas & Joshi, 1999). This section introduces LTAG and representational and processing issues within it.

The idea behind LTAG is to localize the computation of linguistic structure by associating lexical items with rich descriptions that impose complex combinatory constraints in a local context. Each lexical item is associated with at least one "elementary tree" structure, which encodes the "minimal syntactic environment" of a lexical item. This includes such information as head-complement requirements, filler-gap information, tense, and voice. Figure 1 shows some of the elementary trees associated with the words of the sentence *The police officer believed the victim was lying*.[2] The trees involved in the correct parse of the sentence are highlighted by boxes. Note that the highlighted tree for *believed* specifies each of the word's arguments, a sentential complement and a noun phrase subject.

Encoding combinatory information in the lexicon rather than in supralexical rules has interesting effects on the nature of structural analysis. One effect is that the number of different descriptions for each lexical item becomes much larger than when the descriptions are less complex. For instance, the average elementary tree ambiguity for a word in Wall Street Journal text is about 47 trees (Srinivas & Joshi, 1999). In contrast, part-of-speech tags, which provide a much less complex description of words, have an ambiguity of about 1.2 tags per word in Wall Street Journal text. Thus, lexicalization increases the local ambiguity for the parser, complicating the problem of lexical ambiguity resolution. The increased lexical ambiguity is partially illustrated in Figure 1, where six out of eight words have multiple elementary tree possibilities. The flip-side to this increased lexical ambiguity, however, is that resolution of lexical ambiguity yields a representation that is effectively a parse, drastically reducing the amount of work to be done after lexical ambiguity is resolved (Srinivas & Joshi, 1999). This is because the elementary trees impose such complex combinatory constraints in their own local contexts that there are very few ways for the trees to combine once they have been correctly chosen. The elementary trees can be understood as having "compiled out" what would be rule applications in a context-free grammar system, so that once they have been correctly assigned, most syntactic ambiguity has been resolved. Thus, the lexicalization of grammar causes much of the computational work of structural analysis to shift from grammatical rule application to lexical ambiguity resolution. We refer to the elementary trees of the grammar as "supertags", treating them as complex analogs to part-of-speech tags. We refer to the process of resolving supertag ambiguity as "supertagging". One indication that the work of structural analysis has indeed been shifted into lexical ambiguity resolution is that the run-time of the parser is reduced by a factor of thirty when the correct supertags for a sentence are selected in advance of parsing.[3]

Importantly for the current work, this change in the nature of parsing has been complemented by the recent development of statistical techniques for lexical ambiguity resolution. Simple statistical methods for resolving part-of-speech ambiguity have been one of the major successes in recent work on statistical natural language processing (cf. Church & Mercer, 1993). Several algorithms tag part-of-speech with accuracy between 95% and 97% (cf. Charniak, 1993). Applying such techniques to the words in a sentence before parsing can substantially reduce the work of the parser by preventing the construction

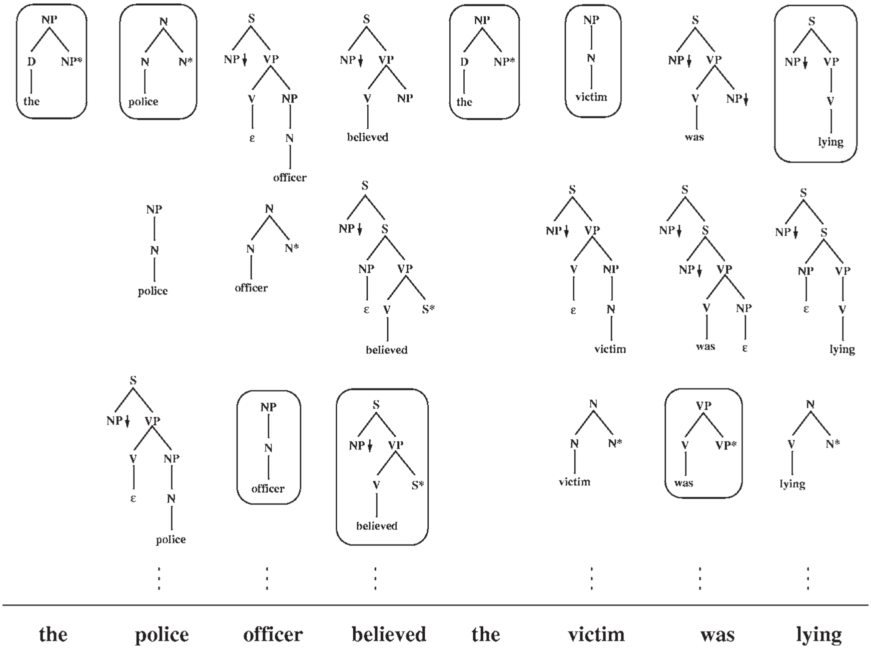


Figure 1. A partial illustration of the elementary tree possibilities for the sentence *the police officer believed the victim was lying*. Trees involved in the correct parse of the sentence are highlighted in boxes

of spurious syntactic analyses. Recently, Srinivas and Joshi (1999) have demonstrated that the same techniques can be effective in resolving the greater ambiguity of supertags. They implemented a tri-gram Hidden Markov Model of supertag disambiguation. When trained on 200,000 words of parsed Wall Street Journal text, this model produced the correct supertag for 90.9% of lexical items in a set of held out testing data.

Thus, simple statistical techniques for lexical ambiguity resolution can be applied to supertags just as they can to part-of-speech ambiguity. Due to the highly constraining nature of supertags, these techniques have an even greater impact on structural analysis when applied to supertags than when applied to part-of-speech tags. These results demonstrate that much of the computation of linguistic analysis, which has traditionally been understood as the result of structure building operations, might instead be seen as lexical disambiguation. This has important implications for how psycholinguists are to conceptualize structural analysis. It expands the potential role in syntactic analysis of simple pattern recognition mechanisms for word recognition, which have played a very limited role in classical models of human syntactic processing.

Note that the claim here is not that supertagging accomplishes the entire task of structural analysis. After elementary trees have been selected for the words in a sentence, there remains the job of connecting the trees via the LTAG combinatory operations of adjunction and substitution. The principal claim here is that in designing a system for syntactic analysis there are sound linguistic and engineering reasons for storing large amounts of grammatical information in the lexicon and for performing much of the work of syntactic analysis with something like supertagging. If such a system is also to be used as a psycholinguistic model, it is natural to predict that many of the initial processing commitments of syntactic analysis are made by a level of processing analogous to supertagging. In the following section, we discuss how an LTAG-based supertagging system resolves at the lexical level many of the same syntactic ambiguities that have concerned researchers in human sentence processing, suggesting that a supertagging system might provide a good psycholinguistic model of syntactic processing. Thus, although the question of how such a system fits into a complete language processing system is an important one, it may be useful to begin exploring the psychological implications of supertagging in advance of a complete understanding of how to design the rest of the system.[4]

A model of the grammatical aspects of word recognition using LTAG

In the remaining sections of this paper, we describe an ongoing project which attempts to use LTAG to develop a more fully-specified description of the CBL theory of human sentence processing. We argue that the notion of supertagging can become the basis of a model of the grammatical aspects of word recognition, provided that certain key adjustments are made to bring it in line with the assumptions of psycholinguistic theory (Kim, Srinivas & Trueswell, in preparation). Before introducing this model, we outline how LTAG can be used to advance the formal specification of the CBL theory.[5] We then turn to some of the findings of the model, which capture some of the major phenomena reported in the human parsing literature.

LTAG lexicalizes syntactic information in a way that is highly consistent with descriptions of the CBL theory, including the lexicalization of head-complement relations, filler-gap information, tense, and voice. The value of LTAG as a formal framework for a CBL account can be illustrated by the LTAG treatment of several psycholinguistically interesting syntactic ambiguities, e.g., prepositional phrase attachment ambiguity, the NP/S ambiguity, the reduced relative/main clause ambiguity, and the compound noun ambiguity. In all but one of these cases, the syntactic ambiguity is characterized as stemming from a lexical ambiguity.

Figure 2 presents the LTAG treatment of these ambiguities. Each of the sentence fragments in the figure ends with a syntactically ambiguous word and is accompanied by possible supertags for that word. First, the prepositional phrase attachment ambiguity is illustrated in Figure 2a. The ambiguity lies in the ability of the prepositional phrase *with the …* to modify either the noun phrase *the cop* (e.g., *with the red hair*) or the verb phrase headed by *saw* (e.g., *with the binoculars*). Within LTAG, prepositions like *with* indicate lexically whether they modify a preceding noun phrase or verb phrase. This causes prepositional phrase attachment ambiguities to hinge on the lexical ambiguity of the preposition. Similarly, the NP/S ambiguity discussed in the Introduction arises directly from the ambiguity between the elementary trees shown in Figure 2b. In this case, these trees encode the different complement-taking properties of the verb *forgot* (e.g., *the recipe* vs. *the recipe was …*). Figure 2c shows a string that could be parsed as a Noun-Noun compound (e.g., *the warehouse fires were extinguished.*) or a Subject-Verb sequence (e.g., *the warehouse fires older employees.*). In non-lexicalist grammars, this ambiguity is treated as arising from the major category ambiguity of *fires*. In LTAG, this ambiguity involves not only the category ambiguity but also a more fine-grained ambiguity regarding the previous noun *warehouse*. Due to the nature of combinatory operations of LTAG, nouns that appear as phrasal heads or phrasal modifiers are assigned different types of elementary trees (i.e., the *Alpha-/Beta-* distinction in LTAG, see Doran, Egedy, Hockey, Srinivas & Zaidel, 1994). Figure 2d illustrates the reduced relative/main clause ambiguity (e.g., *the defendant examined by the lawyer was …* vs. *the defendant examined the pistol.*). Here again, the critical features of the phrase structure ambiguity are lexicalized. For instance, the position of the gap in an object-extraction relative clause is encoded at the verb (right-hand tree in Figure 2d). This is because LTAG trees encode the number, type, and position of all verb complements, including those that have been extracted. Finally, Figure 2e illustrates a structural ambiguity that is not treated lexically in LTAG. As in Figure 2a, the preposition *with* is associated with two elementary trees, specifying verb phrase or noun phrase modification. However, in this example, both attachment possibilities involve the same tree (NP-attachment), which can modify either *general* or *secretary*. The syntactic information that distinguishes between local and non-local attachment is not specified lexically. So, within LTAG, this final example is a case of what we

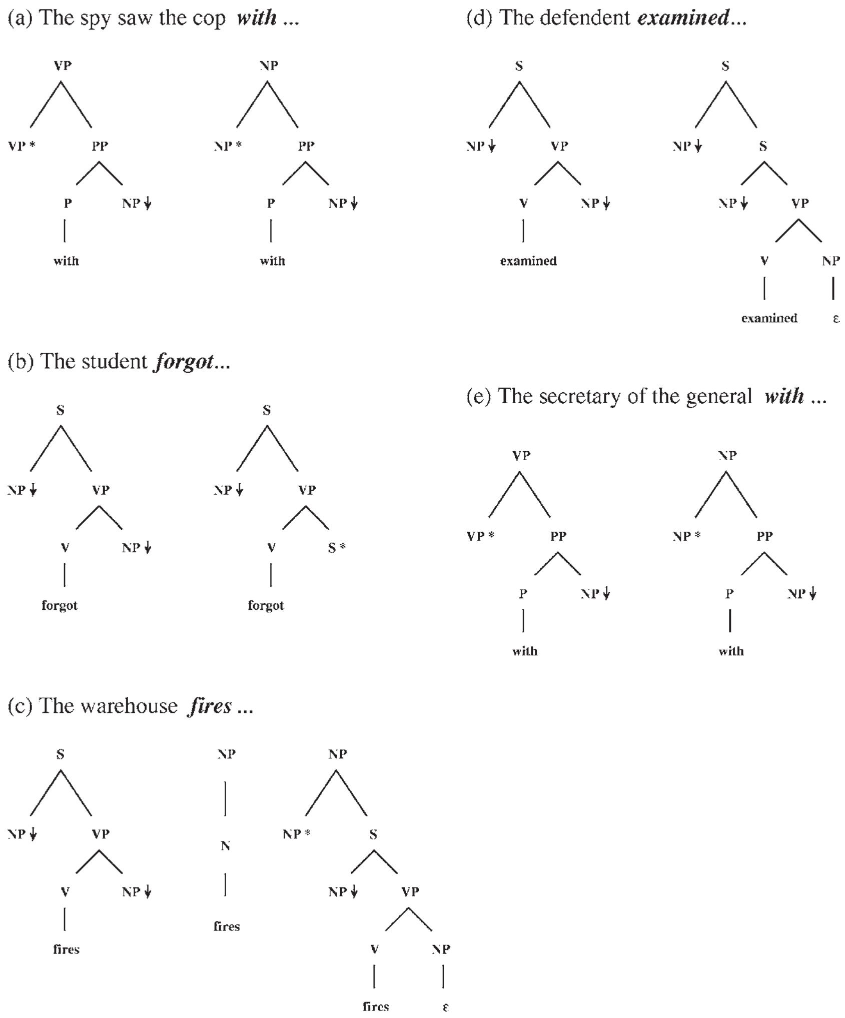


Figure 2. LTAG treatment of several psycholinguistically interesting syntactic ambiguities: (a) PP-attachment ambiguity; (b) NP/S ambiguity; (c) N/V category ambiguity; (d) reduced relative/main clause ambiguity; (e) PP-attachment ambiguity with both attachment sites being nominal.

might call true attachment ambiguity. This example illustrates the point made earlier that even when a lexical tree is selected, syntactic processing is not complete, since lexical trees need to be combined together through the operations of substitution and adjunction. In the first four examples, the selection of lexical trees leaves only a single way to combine these items. In the final example, however, multiple combinatory possibilities remain even after lexical selection.

The examples in Figure 2 illustrate the compatibility of LTAG with the CBL theory. The two frameworks lexicalize structural ambiguities in similar ways, with LTAG providing considerably more linguistic detail. This suggests that LTAG can be used to provide a more formal statement of the representational claims of the CBL theory. For instance, one can characterize the grammatical aspects of word recognition as the parallel activation of possible elementary trees. The extent to which a lexical item activates a particular elementary tree is determined by the frequency with which it has required that tree during an individual's linguistic experience. The selection of a single tree is accomplished through the satisfaction of multiple probabilistic constraints, including semantic and syntactic contextual cues. The CBL theory has traditionally focused on the activation of verb argument structure. The introduction of a wide-coverage grammar into this theory generates clear predictions about the grammatical representations of other classes of words. The same ambiguity resolution processes occur for all lexical items for which LTAG specifies more than one elementary tree.

The grammatical predictions of LTAG are worked out in an English grammar, which is the product of an ongoing grammar development project at the University of Pennsylvania (Doran et al., 1994). The grammar provides lexical descriptions for 37,000 words and handles a wide range of syntactic phenomena, making it a highly robust system. The supertagging work described in this chapter makes critical use of this grammar. The comprehensiveness of the grammar makes it a valuable tool for psycholinguistic work, by allowing formal statements about the structural properties of a large fragment of the language. In our case, it plays a critical role in our attempt to "scale up" CBL models in order to investigate the viability of such models on more complex grammatical situations than they have been tested on before.

Implementation

In this section, we describe preliminary results of a computational modeling project exploring the ability of the CBL theory to integrate the representations of LTAG. We have been developing a connectionist model of the grammatical aspects of word recognition (Kim et al., in preparation), which attempts to account for various psycholinguistic findings pertaining to syntactic ambiguity resolution. Unlike previous connectionist models within the CBL approach (McRae et al., 1998; Tabor et al., 1997; Spivey-Knowlton, 1996; Juliano

& Tanenhaus, 1994), this model has wide coverage in that it has an input vocabulary of 20,000 words and is designed to assign 304 different LTAG elementary trees to input words. The design of the model was not guided by the need to match a specific set of psycholinguistic data. Rather, we applied simple learning principles to the acquisition of a wide coverage grammar, using as input a corpus of highly-variable, naturally occurring text. Certain patterns of structural preferences and frequency effects, which are characteristic of human data, fall directly out of the model's system of distributed representation and frequency-based learning.

The model resembles the statistical supertagging model of Srinivas & Joshi (1999), which we briefly described above. We have, however, made key changes to bring it more in line with the assumptions behind the CBL framework. The critical assumptions are that human language comprehension is characterized by distributed, similarity-based representations (cf. Seidenberg, 1992) and by incremental processing of a sentence. The Srinivas and Joshi model permits the use of information from both left and right context in the syntactic analysis of a lexical item (through the use of Viterbi decoding). Furthermore, their model has a "perfect" memory, which stores the structural events involving each lexical item separately and without error. In contrast, our model processes a sentence incrementally, and its input and internal representations are encoded in a distributed fashion. Distributed representations cause each representational unit to play a role in the representation of many lexical items, and the degree of similarity among lexical items to be reflected in the overlap of their representations.

These ideas were implemented in a connectionist network, which provided a natural framework for implementing a distributed processing system.[6] The model takes as input information about the orthographic and semantic properties of a word and attempts to assign the appropriate supertag for the word given the local left context. The architecture of the model consists of three layers with feed-forward projections, as illustrated in Figure 3.

The model's output layer is a 95 unit array of syntactic features which is capable of uniquely specifying the properties of 304 different supertags. These features completely specify the components of an LTAG elementary tree: 1) part-of-speech, 2) type of "extraction," 3) number of complements, 4) category of each complement, and 5) position of complements. Each of these components is encoded with a bank of localist units. For instance, there is a separate unit for each of 14 possible parts of speech, and the correct activation pattern for a given supertag activates only one of these units (e.g., "Noun").

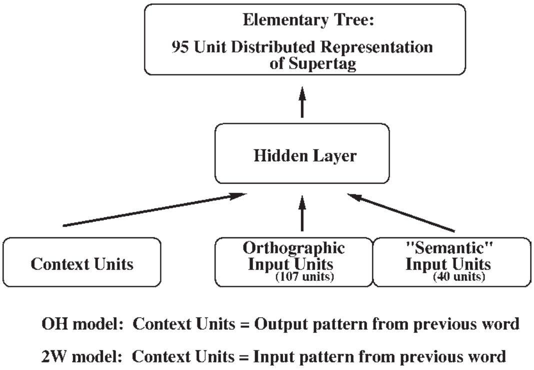

Elementary Tree:
95 Unit Distributed Representation of Supertag
Hidden Layer
Context Units
Orthographic Input Units (107 units)
"Semantic" Input Units (40 units)
OH model: Context Units = Output pattern from previous word
2W model: Context Units = Input pattern from previous word

Figure 3. Architecture of the model

The model was given as input rudimentary orthographic information and fine-grained distributional information about a word. 107 of the units encoded orthographic features, namely the 50 most common three-letter word-initial segments (e.g., *ins*), the 50 most common two-letter word-final segments (e.g., *ed*), and seven properties such as capitalization, hyphenation, etc. The remaining 40 input units provide a "distributional profile" of each word, which was derived from a co-occurrence analysis.

The orthographic encoding scheme served as a surrogate for the output of morphological processing, which is not explicitly modeled here but is assumed to be providing interactive input to lexico-syntactic processes that are modeled. The scheme was chosen primarily for its simplicity – it was automatically derived and easily applied to the training and testing corpus, without requiring the use of a morphological analyzer. It was expected to correlate with the presence of common English morphological features.

Similarly, the distributional profiles were used as a surrogate for the activation of detailed semantic information during word recognition. Although space prevents a detailed discussion, we note that several researchers have found that co-occurrence-based distributional profiles provide detailed information about the semantic similarity between words (cf. Burgess & Lund, 1997; Landauer & Dumais, 1997; Schuetze, 1993). The forty-dimensional profiles used here were created by first collecting co-occurrence statistics for a set of 20,000 words in a large corpus of newspaper text.[7] The co-occurrence matrix was compressed by extracting the 40 principal components of a Singular Value Decomposition (see Kim et al., in preparation, for details). An informal inspection of the space reveals that it captures certain grammatical and semantic information. Table 1 shows the nearest neighbors in the space for some selected words. These are some of the better examples, but in general the information in the space consistently encodes semantic similarities between words.

Table 1. Nearest neighbors of sample words based on distributional profiles

Word	Nearest neighbors by distributional profile
scientist	researcher, scholar, psychologist, chemist
london	tokyo, chicago, atlanta, paris
literature	poetry, architecture, drama, ballet
believed	feared, suspected, convinced, admitted
bought	purchased, loaned, borrowed, deposited
smashed	punched, cracked, flipped, slammed
confident	hopeful, optimistic, doubtful, skeptical
certainly	definitely, obviously, hardly, usually
from	with, by, at, on

We implemented two variations on the basic architecture described above, which gave the model an ability to maintain information over time so that its decisions would be context sensitive. The first variation expanded the input pattern to provide on each trial a copy of the input pattern from the previous time step along with the current input. This allowed the network's decisions about the current input to be guided by information about the preceding input. We will call this architecture the "two-word input" model (2W). The second variation provided simple recurrent feedback from the output layer to the hidden layer so that on a given trial the hidden layer would receive the previous state of the output layer. This again allowed the model's decision on a given trial to be contingent on activity during the previous trial. We call this architecture the "output-to-hidden" architecture (OH). For purposes of brevity, we discuss only the results of the 2W architecture. In all statistical analyses reported here, the OH architecture produced the same effects as the 2W architecture.

The model was trained on a 195,000 word corpus of Wall Street Journal text, which had been annotated with supertags. The annotation was done by translating the annotations of a segment of the Penn Treebank (Marcus, Santorini & Marcinkiewicz, 1993) into LTAG equivalents (Srinivas, 1997). During training, for each word in the training corpus, the appropriate orthographic units and distributional profile pattern were activated in the input layer. The input activation pattern was propagated forward through the hidden layer to the output layer. Learning was driven by back propagation of the error between the model's output pattern and the correct supertag pattern for the current word (Rumelhart, Hinton & Williams, 1986).

We tested the overall performance of the model by examining its supertagging accuracy on a 12,000 word subset of the training corpus that was held out of training. The network's syntactic analysis on a given word was considered to be the supertag whose desired activation pattern produced the lowest error with respect to the model's actual output (using least squares error). On this metric, the model guessed correctly on 72% of these items. Using a slightly relaxed metric, the correct supertag was among the model's top three choices (the three supertags with the lowest error) 80% of the time. This relaxed metric was used primarily to assess the model's potential for increased overall accuracy in future work; if the correct analysis was highly activated even when it was not the most highly activated analysis, then future changes might be expected to increase the model's overall accuracy (e.g., improvements to the quality of the input representation). Accuracy for basic part of speech on the relaxed metric was 91%. The performance of the network can be compared to 79% accuracy for a "greedy" version of the tri-gram model of Srinivas & Joshi (1999), which was trained on the same corpus. The greedy version eliminated the previously mentioned ability of the original model to be influenced by information from right context in its decisions about a given word.

Although these results indicate that the model acquired a substantial amount of grammatical knowledge, the main goal of this work is to examine the relationship between the model's operation and human behavioral patterns, including the patterns of misanalysis characteristic of human processing. In pursuing this goal, we measure the model's degree of commitment to a given syntactic analysis by the size of its error to that analysis relative to its error to other analyses. We make the linking hypothesis that reading time elevations due to misanalysis and revision in situations of local syntactic ambiguity should be predicted by the model's degree of commitment to the erroneous syntactic analysis at the point of ambiguity. For example, in the NP/S ambiguity of Example 1, the model's degree of commitment to the NP-complement analysis over the S-complement analysis should predict the amount of reading time elevation at the disambiguating region *was in…* . Examination of the model's processing of syntactic ambiguities revealed patterns characteristic of human processing.

Modeling the NP/S ambiguity

One pattern of behavioral data that our model aims to account for is the pattern of processing difficulty around the NP/S ambiguity, illustrated by *The chef forgot the recipe was in the back of the book* (discussed in the Introduction as

(1)). In (1a), comprehenders can initially treat the noun phrase *the recipe* as either a NP-complement of *forgot* or the subject of a sentential complement. Numerous experiments have found that readers of locally ambiguous sentences like 1a often erroneously commit to a NP-complement interpretation (Holmes et al., 1989; Ferreira & Henderson, 1990; Trueswell et al., 1993; Garnsey et al., 1997).

Several experiments have found that the general processing bias toward the NP-complement is modulated by the structural bias of the main verb (Trueswell et al., 1993; Garnsey et al., 1997). Erroneous commitments to the NP-complement interpretation are weakened or eliminated when the main verb has a strong S-bias (e.g., *claimed*). Similar effects have also been found when verb bias information is introduced to processing through a lexical priming technique (Trueswell & Kim, 1998). Thus, the language processing system appears to be characterized simultaneously by an overall bias toward the NP-complement analysis and by the influence of the lexical preferences of S-bias verbs.

The coexistence of these two conflicting sources of guidance may be explained in terms of "neighborhoods of regularity" in the representation of verb argument structure (Seidenberg, 1992; Juliano & Tanenhaus, 1994). NP-complement and S-complement verbs occupy neighborhoods of representation, in which the NP-complement neighborhood dominates the "irregular" S-complement neighborhood, due to greater membership. The ability of S-complement items to be represented accurately is dependent on frequency. High frequency S-complement items are accurately represented, but low frequency S-complement items are overwhelmed by their dominant NP-complement neighbors. Juliano & Tanenhaus (1993) found evidence in support of this hypothesis in a study in which the ability of verb bias information to guide processing was characterized by an interaction between the frequency and the subcategory of the main verb. The ability of S-complement verbs to guide processing commitments was correlated with the verb's lexical frequency. Low frequency S-complement verbs allowed erroneous commitments to the NP-complement analysis in spite of the verb's bias, while high frequency S-complement items caused rapid commitments to the correct S-complement analysis.

Our model provides such a neighborhood-based explanation of the human processing data for NP/S ambiguities. We presented the model with NP/S ambiguous fragments, such as *The economist decided* …, which contained either a verb that strongly tended to take S-complements in the training corpus or strongly tended to take NP-complements. The model assigned either a

NP- or S-complement analysis to 96% of such verbs, indicating that it clearly recognized NP/S verbs. In resolving the NP/S ambiguity, the model showed a general bias toward the NP-complement structure, which can be overcome by lexical information from high frequency S-complement verbs. All NP-biased verbs were correctly analyzed, but S-biased verbs were misanalyzed on 9 of 14 items, with 8 of 9 misanalyses being to the NP-complement. The dominance of the NP-complement analysis, however, is modulated by the frequency of exposure to S-complement items. The model accurately subcategorized S-biased verbs when they were high in frequency (5 of 7) but was highly inaccurate on low frequency items (none were correctly classified; 6 of 7 were mis-analyzed as NP-complement verbs).

The model's frequency-by-subcategory interaction arises from its system of distributed representation and frequency sensitive learning. S-complement verbs and NP-complement verbs have a substantial overlap in input representation, due to distributional and orthographic similarities (-ed, -ng, etc.) between the two types of verbs and the fact that S-complement verbs are often NP/S ambiguous. NP-complement tokens dominate S-complement tokens in frequency by a ratio of 4 to 1, causing overlapping input features to be more frequently associated with the NP-complement output than the S-complement output during training. The result is that a portion of the input representation of S-complement verbs becomes strongly associated with the NP-complement output, causing a tendency for the model to misanalyze S-complement items as NP-complement items. The model is able to identify non-overlapping input features that distinguish S-complement verbs from their dominant neighbors, but its ability to do so is affected by frequency. When S-complement verbs are seen in high frequencies, their distinguishing features are able to influence connection weights enough to allow accurate representation; however, when S-complement verbs are seen in low frequencies, their NP-complement-like input features dominate their processing. The explanation here is similar to the explanation given by Seidenberg & McClelland (1989) for frequency-by-regularity interactions in word naming (e.g., the high frequency irregularity of *ha*ve vs. the regularity of *gave*, *wave*, *save*) and past tense production.

The theoretical significance of this interaction lies partly in its emergence in a comprehensive model, which is designed to resolve a wide range of syntactic ambiguities over a diverse sample of the language. These results provide a verification of conclusions drawn by Juliano & Tanenhaus (1994) from a much simpler model, which acquired a similar pattern of knowledge about NP-complement and S-complement verbs from co-occurrence information about verbs and the words that follow them. It is important to provide such follow-up work for Juliano & Tanenhaus (1994), because their simplifications of the domain were extreme enough to allow uncertainty about the scalability of their results. Although their training materials were drawn from naturally occurring text (Wall Street Journal and Brown corpus), they sampled only a subset of the verbs in that text and the words occurring after those verbs. S-complement tokens were more common in their corpus than in the full language, and only past-tense tokens were sampled. This constitutes a substantial simplification of the co-occurrence information available in the full language. In our sample of the Wall Street Journal corpus, non-auxiliary verbs account for only 10.8% of all tokens, suggesting that the full language may contain many co-occurrence events that are "noise" with respect to the pattern detected by the Juliano & Tanenhaus (1994) model. For instance, as Juliano & Tanenhaus observe, their domain restricts the range of contexts in which the determiner *the* occurs, obscuring the fact that in the full language, *the* often introduces a subject noun phrase rather than an object noun phrase. It is conceivable that the complexity of the full language would obscure the pattern of co-occurrences around the NP/S ambiguity sufficiently to prevent a comprehensive constraint-based model from acquiring the pattern of knowledge acquired by the Juliano & Tanenhaus (1994) model. Our results demonstrate that the processing and representational assumptions that allow constraint based models to naturally express frequency-by-regularity interactions are scalable – they continue to emerge when the domain is made very complex.

Modeling the noun/verb lexical category ambiguity

Another set of behavioral data that our model addresses is the pattern of reading times around lexical category ambiguities like that of *fires* in (4).

(4) a. *the warehouse* **fires** *burned for days.*
b. *the warehouse* **fires** *many workers every spring*.

The string *warehouse fires* can be interpreted as a subject-verb sequence (4a) or a compound noun phrase (4b). This syntactic ambiguity is anchored by the lexical ambiguity of *fires*, which can occur as either a noun or a verb.

Several experiments have shown that readers of sentences like (4a) often commit erroneously to a subject-verb interpretation, as indicated by processing difficulty at the next word (*burned*), which is inconsistent with the erroneous interpretation and resolves the temporary ambiguity. Corley (1998) has shown that information about the category bias of the ambiguous word is rapidly employed in the resolution of this ambiguity. When the ambiguous word is one that tends statistically to be a verb, readers tend to commit erroneously to the subject-verb interpretation, but when the word tends to occur as a noun, readers show no evidence of misanalysis. MacDonald (1993) has found evidence of more subtle factors, including the relative frequency with which the preceding noun occupies certain phrase-structural positions, the frequency of co-occurrence between the preceding noun and ambiguous word, and semantic fit information. Most important for the current work, MacDonald found that when the ambiguous word was preceded by a noun that tended to occur as a phrasal head, readers tended to commit to the subject-verb interpretation. However, when the preceding noun tended to occur as a noun modifier, readers tended to commit immediately to the correct noun-noun compound analysis. The overall pattern of data suggests a complex interplay of constraints in the resolution of lexical category ambiguity. Lexically specific information appears to be employed very rapidly and processing commitments appear to be affected by multiple sources of information, including subtle cues like the modifier/head likelihood of a preceding noun.

Like human readers, our model shows sensitivity to both lexical category bias and fine-grained contextual cues when processing locally ambiguous fragments like *the warehouse fires*. We presented the model with fragments ending in noun/verb ambiguous verbs (e.g., *the emergency plans*). The ambiguous words were either noun biased (e.g., *plans*), verb-biased (e.g., *pay*), or equi-biased (e.g., *bid*). The preceding noun was either one that tended to occur as a phrasal head in the training corpus (e.g., *division*) or one that tended to occur as a noun modifier in the corpus (e.g., *emergency*). Lexical bias was determined by frequency properties in the training corpus.

The model clearly recognized the target words as nouns and verbs, as indicated by the fact that 97% of the test items were assigned either a noun supertag or a verb supertag. More subtle aspects of the model's operation were revealed by an examination of the activation values of the noun and verb part-of-speech units separately from the rest of the output layer. The model showed strong commitments to the contextually supported category when that category was either the dominant sense of a biased word or when the word was equi-biased – the contextually supported unit had superior activation in 90% of such cases. In contrast, the model had difficulty activating the contextually supported category when it was the subordinate category of a biased word – showing superior activation for the contextually supported category in only 35% of such cases. Thus, context and lexical bias interacted such that the model showed a strong tendency to activate a contextually-supported pattern when it was either the dominant pattern or had an equally frequent alternative, but when context supported the subordinate pattern, the model was unable to activate this pattern.

Interestingly, this interaction resembles the "subordinate bias" effect observed in the semantic aspects of word recognition (Duffy, Morris & Rayner, 1988). When semantically ambiguous words are encountered in biasing contexts, the effects of context depend on the nature of the word's bias. When preceding discourse context supports the subordinate sense of a biased ambiguous word, processing difficulty occurs. When context supports the dominant sense or when it supports either sense of an equi-biased word, no processing difficulty occurs. Our model shows a qualitatively identical effect with respect to category ambiguity. We take this as further support for the idea, central to lexicalist theories, that lexical and syntactic processing obey many of the same processing principles. On the basis of this kind of effect in the model, we predict that human comprehenders should show subordinate bias effects in materials similar to the ones used here. Furthermore, because the subordinate bias effects found here are quite natural given the model's system of representation and processing, we would expect similar effects to arise in the model and in humans with respect to other syntactic ambiguities that are affected by local left context (see Trueswell, 1996, for similar predictions about subordinate bias effects involving the main clause/relative clause ambiguity).

General discussion

We have attempted to advance the grammatical coverage and formal specification of Constraint-based Lexicalist models of language comprehension. A convergence of perspectives between CBL theory in psycholinguistics and work in theoretical and computational linguistics has supported and guided our proposals. We have attempted to give a concrete description of the syntactic aspects of the CBL theory by attributing to human lexical knowledge the grammatical properties of a wide coverage Lexicalized Tree Adjoining Grammar (Doran et al., 1994). In developing a processing model, we have drawn insight from work on processing with LTAG which suggests that statistical mechanisms for lexical ambiguity resolution may accomplish much of the computation of parsing when applied to rich lexical descriptions like those of LTAG (Srinivas & Joshi, 1999). We have incorporated these ideas about grammar and processing into a psychologically motivated model of the grammatical aspects of word recognition, which is wide in grammatical coverage.

The model we describe is general in purpose; it acquires mappings between a large sample of the lexical items of the language and a large number of rich grammatical representations. Its design does not target any particular set of syntactic ambiguities. Nevertheless, it qualitatively captures subtle patterns of human processing data, such as the frequency-by-regularity interaction in the NP/S ambiguity (Juliano & Tanenhaus, 1993) and the use of fine-grained contextual cues in resolving lexical category ambiguities (MacDonald, 1993).

The wide range of grammatical constructions faced by the model and the diversity of its sample of language include much of the complexity of the full language and support the idea that constraint-based models of sentence processing are viable, even on a large grammatical scale. The model provides an alternative to the positions of Mitchell et al. (1995) and Corley & Crocker (1996), which propose statistical processing models with only coarse-grained parameters such as part-of-speech tags. Their argument is that the sparsity of some statistical data causes the fine-grained parameters of constraint-based models to be "difficult to reliably estimate" (Corley & Crocker, 1996) and that the large number of constraints in constraint-based models causes the management of all these constraints to be computationally intensive. Such arguments assume that a coarse-grained statistical model is more viable and more "compact" than a fine-grained model.

The issue of whether fine-grained statistical processing is viable may hinge on some basic computational assumptions. The observation that sparsity of statistical data affects the performance of statistical processing systems is certainly valid. But there are a number of reasons why this does not support arguments against fine-grained statistical processing models. First, there is a large class of statistical processing models, including connectionist systems like the one used here, that are well suited to the use of imperfect cues. For instance, a common strategy employed by statistical NLP systems to deal with sparse data is to "back off" to statistics of a coarser grain. This is often done explicitly, as in verb subcategorization methods, where decisions are conditionalized on lexical information (individual verbs) when the lexical item is common, but are conditionalized on (backed off to) basic category information (all verbs), when the lexical item is rare (Collins, 1995). In connectionist systems like ours, statistical back-off is the flip-side of the network's natural tendency to generalize but also to be guided by fine-grained cues when those cues are encountered frequently. Fine grained features of a given input pattern are able to influence behavior when they are encountered frequently, because they are given repeated opportunities to influence connection weights. When such fine-grained features are not encountered often enough, they are overshadowed by coarser-grained input features, which are by their very nature more frequent. Systems like our model can be seen as discovering back-off points. We argue that systems that do such backing off are the appropriate class of system for modeling much of sentence processing. As a back-propagation learning system with multiple grammatical tasks competing for a limited pool of processing resources, our model is essentially built to learn to ignore unreliable cues.

Thus, the interaction between frequency and subcategory that we have discussed emerges naturally in the operation of statistical processing devices like the model described here. Fine-grained information about S-complement verbs is able to guide processing when it is encountered often enough during training to influence connection weights in spite of the dominance of NP-complement signals. The ability of Head/Modifier likelihood cues about nouns to influence connection weights is similarly explained.

In general, we view the sparsity of data as an inescapable aspect of the task of statistical language processing rather than as a difficulty that a system might avoid by retreating to more easily estimable parameters. Even part-of-speech tagging models like Corley & Crocker's (1996) include a lexical component, which computes the likelihood of a lexical item given a candidate part-of-speech for that word, and their model is therefore affected by sparsity of data for individual words – this is true for any tagger based on the dominant Hidden Markov Model framework. Furthermore, as mentioned earlier, work in statistical NLP has increasingly indicated that lexical information is too valuable to ignore in spite of the difficulties it may pose. Techniques that count lexically specific events have generally out-performed techniques that do not, such as statistical context-free grammar parsing systems (see Marcus, this volume). It seems to us that, given a commitment to statistical processing models in general, there is no empirical or principled reason to restrict the granularity of statistical parameters to a particular level, such as the part-of-speech tags of a given corpus. Within the engineering work on part-of-speech tagging, there are a number of different tag-sets, which vary in the granularity of their tags for reasons unconnected to psychological research, so that research does not motivate a psychological commitment to any particular level of granularity. Furthermore, the idea that the language processing system should be capable of counting statistical events at only a single level of granularity seems to be an assumption that is inconsistent with much that is known about cognition, such as the ability of the visual processing system to combine probabilistic cues from many levels of granularity in the recognition of objects. The solution to the data sparsity problem, as manifested in humans and in successful engineering systems, is to adopt the appropriate learning and processing mechanisms for backing off to more reliable statistics when necessary.

We have argued that the complexities of statistical processing over fine grained lexical information do not warrant the proposal of lexically-blind processing mechanisms in human language comprehension. Although the complexities may be unfamiliar, they are tractable, and there are large payoffs to dealing with them. An increasingly well understood class of constraint-satisfaction mechanisms is well suited to recognizing fine-grained lexical patterns and also to backing off to coarser-grained cues when fine-grained data is sparse. The modeling work described here and research in computational linguistics suggests that such mechanisms, when applied to the rich lexical representations of lexicalized grammars, can accomplish a substantial amount of syntactic analysis. Furthermore, the kind of mechanism we describe shows a pattern of processing that strongly resembles human processing data, suggesting that such mechanisms are good models of human language processing.

Acknowledgements

This work was partially supported by National Science Foundation Grant SBR-96-16833; the University of Pennsylvania Research Foundation; and the Institute for Research in Cognitive Science at the University of Pennsylvania (NSF-STC Cooperative Agreement number SBR-89-20230). The authors thank Marian Logrip for assistance in the preparation of this paper and thank Paola Merlo, Suzanne Stevenson, and two anonymous reviewers for helpful comments on the paper.

Notes

[1] The amount of syntactic structure that is lexically generated goes beyond the classical notion of argument structure. In lexicalized grammar formalisms such as LTAG, the entire grammar is in the lexicon. For instance, the attachment site of a preposition can be treated as a lexically specific feature. Noun attaching prepositions and verb attaching prepositions have different senses. We will discuss this in further detail in the following sections.

[2] The down-arrows and asterisks in the trees mark nodes at which trees make contact with each other during the two kinds of combinatory operations of Tree Adjoining Grammar, substitution and adjunction. Down-arrows mark nodes at which the substitution operation occurs, and asterisks mark footnodes, which participate in the adjunction operation. The

details of the combinatory operations of TAG are beyond the scope of this chapter. See Joshi and Schabes (1996) for a discussion.

[3] This is based on run-times for a sample of 1300 sentences of Wall Street Journal text, reported by Srinivas and Joshi (1999). Running the parser without supertagging took 120 seconds, while running it with correct supertags pre-assigned took 4 seconds.

[4] Srinivas (1997) has suggested that this can be done by a process that is simpler than full parsing. He calls this process "stapling".

[5] Of course, formal specification of this theory can be achieved by using other lexicalized grammatical frameworks, e.g., LFG (Bresnan & Kaplan, 1982), HPSG (Pollard & Sag, 1994), CCG (Steedman, 1996).

[6] This is not to say that left-to-right processing and overlapping representations cannot be incorporated into a symbolic statistical system. However, most attempts within psycholinguistics to incorporate these assumptions into a computationally explicit model have been made within the connectionist framework (e.g., Elman, 1990; Juliano & Tanenhaus, 1994; Seidenberg & McClelland, 1989). By using a connectionist architecture for the current model, we are following this precedent and planning comparisons with existing modeling results.

[7] For each of the 20,000 target words, we counted co-occurrences with a set of 600 high frequency "context" words in 14 million words of Associated Press newswire. Co-occurrences were collected in a six-word window around each target word (three words to either side of the word).

References

Boland, J.E., Tanenhaus, M.K., Garnsey, S.M. & Carlson, G.N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory & Language, 34*, 774–806.

Burgess, C. & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language & Cognitive Processes, 12*, 177–210.

Charniak, E. (1993). *Statistical language learning*. Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Church, K. & Mercer, R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics, 19*, 1–24.

Corley, S. (1998). A Statistical Model of Human Lexical Category Disambiguation. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK.

Corley, S. & Crocker, M.W. (1996). Evidence for a tagging model of human lexical category disambiguation. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*.

Crocker, M.W. (1994). On the nature of the principle-based sentence processor. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*, pp. 245–266. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Doran, C., Egedi, D., Hockey, B.A., Srinivas, B. & Zaidel, M. (1994). XTAG system – a wide coverage grammar for English. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING '94)*. Kyoto, Japan.

Duffy, S.A., Morris, R.K. & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory* & *Language, 27*, 429–446.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Ferreira, F. & Henderson, J.M. (1990). The use of verb information in syntactic parsing: A comparison of evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 555–568.

Ford, M., Bresnan, J. & Kaplan, R.M. (1982). A competence-based theory of syntactic closure. In J. Bresnan (Ed.), *The mental representation of grammatical relations*, pp. 727–796. Cambridge, MA: MIT Press.

Frazier, L. (1995). Constraint satisfaction as a theory of sentence processing. *Journal of Psycholinguistic Research, 24*, 437–468.

Frazier, L. (1989). Against lexical generation of syntax. In W.D. Marslen-Wilson (Ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press.

Frazier, L. (1979). *On comprehending sentences: Syntactic parsing strategies*. Bloomington, IN: Indiana University Linguistics Club.

Garnsey, S.M., Pearlmutter, N.J., Myers, E. & Lotocky, M.A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*, 1–76.

Gross, M. (1984). Lexicon-grammar and the syntactic analysis of French. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING '84)*, Stanford, CA.

Holmes, V.M., Stowe, L. & Cupples, L. (1989). Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language, 28*, 668–689.

Joshi, A. & Schabes, Y. (1997). *Tree-adjoining grammars*. In Rozenberg, G. and Salomaa, A. (Eds.). Handbook of Formal Languages, Volume 3, pp. 63–124. Berlin, New York: Springer-Verlag.

Juliano, C. & Tanenhaus, M.K. (1994). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research, 23*, 459–471.

Juliano, C. & Tanenhaus, M.K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Kaplan, R. & Bresnan, J. (1982). Lexical functional grammar: A formal system of grammatical representation. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*, pp. 173–281. Cambridge, MA: The MIT Press.

Kim, A., Srinivas, B. & Trueswell, J. (in preparation). A computational model of lexico-syntactic processing during language comprehension: syntactic analysis through word recognition.

Landauer, T.K. & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction & representation of knowledge. *Psychological Review, 104*, 211–240.

Lewis, R.L. (1993). *An architecturally-based theory of human sentence comprehension*. Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, PA.

Lund, K., Burgess, C. & Atchley, R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*.

MacDonald, M. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*, 157–201.

MacDonald, M.C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory* & *Language, 32*, 692–715.

MacDonald, M.C., Pearlmutter, N.J. & Seidenberg, M.S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313–330.

McRae, K., Spivey-Knowlton, M.J. & Tanenhaus, M.K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283–312.

Mitchell, D.C. (1989). Verb-guidance and other lexical effects in parsing. *Language and Cognitive Processes, 4*, 123–154.

Mitchell, D.C. (1987). Lexical guidance in human parsing: Locus and processing characteristics. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mitchell, D.C., Cuetos, F., Corley, M.M.B. & Brysbaert, M. (1995). Exposure-based models of human parsing. *Journal of Psycholinguistic Research, 24*, 469–488.

Perfetti, C.A. (1990). The cooperative language processors: Semantic influences in an autonomous syntax. In G.B. Flores d'Arcais, D.A. Balota & K. Rayner (Eds.), *Comprehension processes in reading*. Hillsdale, NJ: Erlbaum.

Pollard, C. & Sag, I. (1994). *Head-driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Prince, A. & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science, 275*, 1604–1610.

Pritchett, B.L. (1992). *Grammatical competence and parsing performance*. Chicago, IL: The University of Chicago Press.

Rayner, K., Carlson, M. & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing. *Journal of Verbal Learning and Verbal Behavior, 22*, 358–374.

Rayner, K. & Frazier, L. (1987). Parsing temporarily ambiguous complements. *Quarterly Journal of Experimental Psychology, 39A*, 657–673.

Rumelhart, D., Hinton, G. & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Schütze, H. (1993). Word space. In S. Hanson, J. Cowan, & C. Giles (Eds.), *Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kaufmann.

Seidenberg, M.S. (1992). Connectionism without tears. In S. Davis (Ed.), *Connectionism: Theory & Practice*. New York, NY: Oxford University Press.

Seidenberg, M.S. & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Sleator, D. & Temperley, D. (1991). Parsing English with a link grammar. *Technical report CMU-CS-91-196, Department of Computer Science, Carnegie Mellon University*.

Spivey-Knowlton, M.J. (1996). *Integration of visual and linguistic information: Human data and model simulations*. Unpublished doctoral dissertation, University of Rochester, Rochester, NY.

Spivey-Knowlton, M.J. & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55*, 227–267.

Srinivas, B. (1997). *Complexity of lexical descriptions and its relevance to partial parsing*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Srinivas, B. & Joshi, A.K. (1999). Supertagging: An approach to almost parsing. *Computational Linguistics, 25*(2), 237–265.

Steedman, M. (1996). *Surface Structure and Interpretation*. Cambridge, MA: MIT Press.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research, 23*, 295–322.

Tabor, W., Juliano, C. & Tanenhaus, M. (1996). A dynamical system for language processing. In *Proceedings of the 18th Annual Conference of the Cognitive Science Society*.

Trueswell, J. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35*, 566–585.

Trueswell, J.C. & Kim, A.E. (1998). How to prune a garden-path by nipping it in the bud: Fast-priming of verb argument structures. *Journal of Memory and Language, 39*, 102–123.

Trueswell, J. & Tanenhaus, M. (1994). Toward a lexicalist framework for constraint-based syntactic ambiguity resolution. In C. Clifton, K. Rayner & L. Frazier (Eds.), *Perspectives on sentence processing*, pp. 155–180. Hillsdale, NJ: Lawrence Erlbaum Associates.

Trueswell, J., Tanenhaus, M., & 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., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 528–553.

Incrementality and lexicalism

A treebank study

Vincenzo Lombardo and Patrick Sturt

DSTA, Università del Piemonte Orientale and Dipartimento di Informatica, Università di Torino, Italy, and HCRC, Department of Psychology, University of Glasgow, Scotland

Current evidence suggests that human parsing is highly incremental, but the consequences of incrementality have not been fully explored. In this paper, we consider one of the consequences of incrementality which poses important questions for lexicalist models of sentence processing; the problem of non-lexical structure building. This problem occurs when a new input word can only be connected to the current partial phrase marker via one or more "headless" projections, whose heads have not yet been read in the input. The necessity to hypothesise such headless projections in the absence of direct lexical evidence raises the potential for serious computational problems. We describe a parsing simulation algorithm which takes a parsed corpus (treebank) and determines the parsing steps which would be required to attach each word in the corpus, assuming an incremental parser. We record the amount of non-lexical structure building produced during the simulation. The results show that the non-lexical structure building required to process realistic input is actually very limited; 80% of words can be attached without any headless projections at all, and very few words require more than one headless projection. Furthermore, there are systematic patterns which suggest that lexical information associated with the current word and left context can aid in the construction of non-lexical structure.

1. Introduction

Evidence from psycholinguistics suggests that human parsing is largely incremental, in the sense that structural commitments are made and interpretations become available on a word-by-word basis (Marslen-Wilson, 1973). A widely held assumption is that the human parser is "strongly incremental"; that is, the parser maintains a fully connected structure as each word is received in the input, without allowing partially structured input to be stored in a disconnected state (for examples of theories which assume strong incrementality, see Frazier and Rayner, 1987; Gibson, 1991, 1998; Gorrell, 1995; Inoue and Fodor, 1995; Stabler, 1994). Strong incrementality contrasts with other proposals such as the head-driven strategy advocated by Abney (1989) and Pritchett (1991), according to which the attachment of a phrase is delayed until its head is reached; such proposals require a mechanism for storing unattached phrases, and are thus not strongly incremental. Recent years have seen increasing empirical evidence against the head-driven strategy. For example, on-line studies of head-final languages (Bader & Lasser, 1994; Yamashita, 1994) show evidence that attachment decisions are not postponed until the head of a phrase is reached, and eye-movement studies of visual recognition show that the semantic interpretation of pre-modified noun phrases can occur well before the head of the phrase is reached (Eberhard, Spivey-Knowlton, Sedivy & Tanenhaus, 1995). However, strong incrementality is not a universally held assumption (see e.g. Merlo, 1996), and it is possible that the parser has some means of interpreting disconnected pieces of structure in its memory (Stabler, 1991; Shieber & Johnson, 1993). Nevertheless, as pointed out by Stabler (1994), strong incrementality yields "a simpler, more natural and more restrictive theory." Our aim in this paper will not be to argue for or against strong incrementality, but to explore some of its consequences.

In tandem with the increasing evidence for incrementality, psycholinguistic models have been strongly influenced by the trend towards lexicalism in linguistic theory. In sentence processing research, this trend has resulted in models which strongly emphasise the role played by words in the parsing process. For example, MacDonald, Pearlmutter, & Seidenberg (1994) suggest that "syntactic structure is built through links between individual lexical items." However, the computational consequences of incrementality for the lexicalist view have not been fully explored. In order to maintain connectedness, the portion of structure connecting each input word to the current representation sometimes has to include links which are not licensed by direct lexical knowledge. These links include nodes that are part of projections whose heads have not yet been read in the input. We will call these projections *headless projections*. As an example, consider the following sentence fragment from the Penn Treebank II (Marcus, Santorini, & Marcinkiewicz, 1993):

(1) He thinks [$_{\text{SBAR}}$ [$_{\text{S}}$ [$_{\text{NP}}$ [$_{\text{ADJ}}$ steeper] prices] have come about because…]]

Here, a strongly incremental parser incorporating the word *steeper* into the representation needs two headless projections, namely, the NP which *steeper* modifies, and the S projection. In general, an incremental parser needs to insert each new word into the representation by means of a subtree, which we will call a *connection path*. The connection path for this example is illustrated in Figure 1.

Structures such as these are not motivated by traditional lexical projection, as assumed in recent linguistic theories (Chomsky, 1995; Pollard & Sag, 1994). However, they are an essential component of models of incremental processing, whether they are stored explicitly with lexical entries, as in MacDonald et al. (1994), or built dynamically by exploiting general syntactic knowledge (Crocker, 1995; Sturt, 1997; Lombardo, Lesmo, Ferraris, & Seidenari, 1998). Our aim is not to argue about where connection paths come from or how they are built, but to address two potential problems concerning headless projections in connection paths. First, we need to know what syntactic knowledge is needed beyond traditional lexical projection, in order to build connection paths. In traditional (non-lexicalised) context-free grammars, for example, this knowledge is available in the form of production rules that do not involve terminal symbols. Second, we need to know the extent of headless projections in the connection paths. If the extent of headless projections is high,

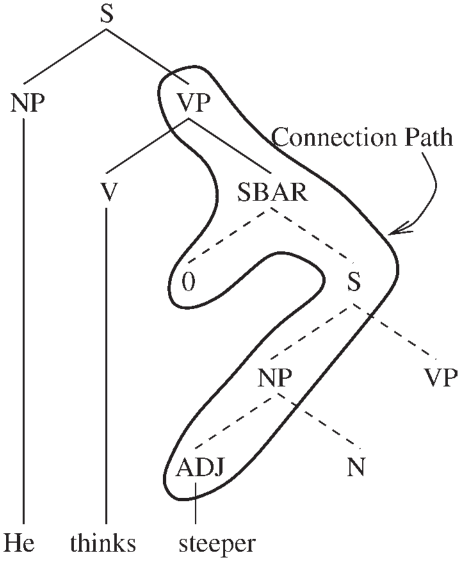


Figure 1. A connection path.

then the compatibility between strong incrementality and lexicalism would be low, since inserting any given word into any given left context would involve a great deal of blind guessing for the parser. Since such guessing would involve extra-lexical knowledge, it would contrast with the tenets of lexicalism, which views sentence processing as being driven mainly by the lexical items.

In this paper, we use a parsed corpus (treebank) to address these problems. We simulate a strongly incremental parser that runs on the trees in the treebank, and collects the connection paths that are needed to incorporate each word from left to right. Note that, because our goal is to study the portions of the connection paths which are not lexically licensed, the connection paths do not include specific lexical items (refer again to Figure 1). The form of the connection paths provides an empirical answer to the first problem: they could be further analysed in terms of some formalism that expresses the extra-lexical knowledge. The results of the analysis of the number of headless projections in the collected connection paths provide an answer to the second problem. The results will provide a quantitative estimation of the extent of the blind guessing required to connect words to the syntactic structure, and will therefore be informative about the compatibility between incrementality and lexicalism.

2. Methodology

Our methodology contrasts with previous computational modelling of incremental processing, which has focussed on the construction of incremental parsing algorithms (Crocker, 1995; Gibson, 1991, 1998; Konieczny, 1996; Sturt & Crocker, 1996), sometimes with incremental semantic interpretation (Lombardo et al., 1998; Steedman, 1989). These have tended to be relatively small scale systems, and it may be difficult to extend them to deal with realistic input. Our approach is not to build a parser, which takes a word string as input and returns a tree, but rather to build a *parsing simulation algorithm*, which takes a full tree as input, and identifies the word-by-word steps by which this tree would have been built by a strongly incremental parser. In order to have a realistic idea of the form of the connection paths and the statistical distribution of the headless projections within them, we have run our algorithm on a *treebank* which covers a realistic sample of natural language (the Penn Treebank (Marcus et al., 1993)).

We believe that our approach is more practical than that of building a traditional parsing model, which requires a grammar, a disambiguation mechanism, and some form of control system for changing from one analysis to another, such as a reanalysis module. We do not claim that our approach is informative about the time-course of processing of the human parser. We make the simplifying assumption that the parser works on the basis of perfect knowledge, and never misanalyses while it is processing a sentence, even though it is well known that people often make temporary misanalyses, and have to revise them later. For example, consider again the sentence fragment (1) *He thinks steeper prices have come about because....* Our parsing simulation algorithm assumes that *steeper* is initially attached in its eventual position as an adjective modifying the subject of the embedded clause. This may well accurately reflect the action of the human parser for this example, since the verb *thinks* very frequently takes a clausal complement. However, as the algorithm contains almost no grammatical or lexical knowledge, it would behave in exactly the same way if the sentence included the verb which frequently takes a noun phrase direct object, as in *He accepts steeper prices have come about because...* In this case, there is evidence that *steeper prices* is initially attached as a direct object of *accepts* (Trueswell, Tanenhaus, & Kello, 1993; Garnsey, Pearlmutter, Myers, & Lotocky, 1997). Hence, in this case, our simulation algorithm would predict a longer connection path than is actually built by the human parser (see also the discussion section).

We do not believe that this problem will cause a serious distortion of our results. Above, we have discussed an example in which our algorithm would probably overestimate the number of headless projections. However, the opposite problem, in which the algorithm *under*estimates the number of headless projections, is likely to be much rarer. In the principle of *minimal attachment*, Frazier (1978) proposed that the human parser minimizes the amount of structure that it has to build on the input of each node. While there have been many claims that this preference can be removed (see the discussion in MacDonald et al. (1994) and references therein), there is virtually no evidence that the preference can actually be *reversed*; that is, it is very unlikely that humans ever systematically follow a principle of *maximal attachment*. If these observations can be generalized to the wider domain of the treebank corpus, then we can assume that when our algorithm makes a wrong estimation of the amount of non-lexical structure building, it is consistently over estimating. We will return to this point later in the discussion, where we consider some ways of introducing misanalyses into the parsing simulation.

At this point, we should point out a hypothesis that we are assuming here; namely that humans can build any of the connection paths in the corpus, in one stage (i.e. on the input of one word, before moving to the next). Note that this hypothesis does not imply that humans would actually build all the connection paths if they were processing the corresponding sentences of the corpus. It is merely a claim about their ability to construct connection paths.

3. The parsing simulation algorithm

Before describing the simulation algorithm, we need to clarify the notion of connection path with respect to the total parse tree.

Given a sentence $s = w_0 w_1 \ldots w_i \ldots w_{n-1}$ and a tree T for it, we define recursively the *incremental trees* $\mathrm{T}_i (i = 0, 1, \ldots, n-1)$ spanning $w_0 \ldots w_i$ as follows:

- T_0 consists of the chain of nodes and edges from w_0 to its maximal projection;
- T_i consists of all the nodes and edges in T_{i-1}, and, for each leaf L between w_{i-1} (excluded) and w_i (included) in the linear order of tree nodes, T_i includes the chain of nodes and edges from *L* to *N*, where *N* is
  - — either a node of T_{i-1},
  - — or the lowest node of T dominating both the root of T_{i-1} and w_i.

For example, given the sentence "Investors welcomed the move," the incremental trees are depicted in Figure 2.

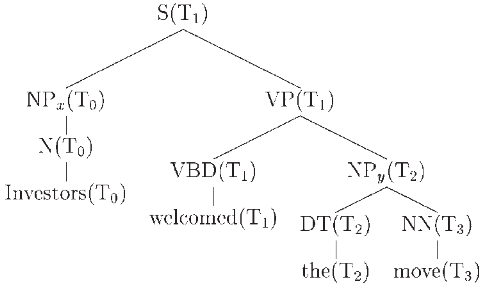


Figure 2. Incremental trees for the sentence "Investors welcomed the move." Each node is labelled with the first incremental tree that includes it. Also notice that T_3 coincides with the total tree.

Note that the definition is general enough to include the cases which include (phonologically) empty nodes (see Figure 3).

Given two incremental trees T_1 and T_2, we define the *difference* between T_1 and T_2 as the tree formed by all the edges which are in T_1 and not in T_2, and all the nodes touched by such edges.

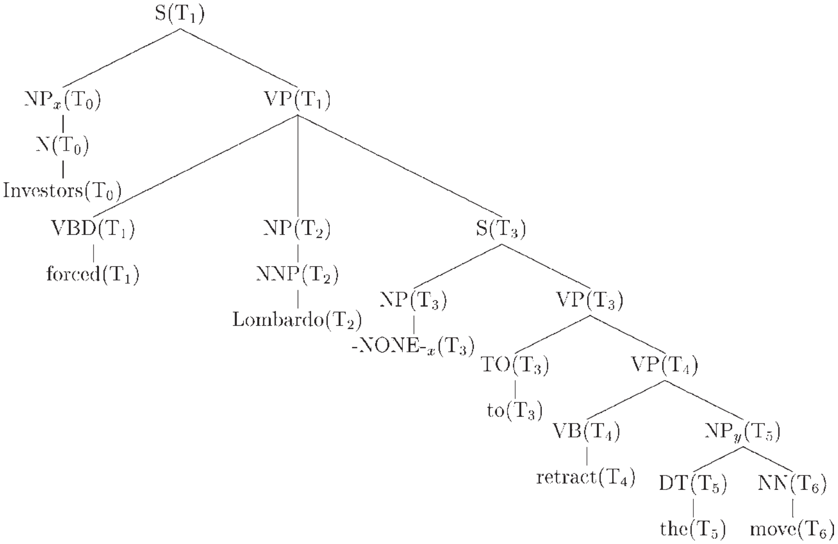


Figure 3. A tree for the sentence "Investors forced Lombardo to retract the move," that shows the representation and processing of empty-category nodes. In this example, the incremental tree T_6 coincides with the total tree.

Now, given a sentence $s = w_0 w_1 \ldots w_{n-1}$ and a tree T for it, the *connection path* for w_i is the difference between the incremental trees T_i and T_{i-1}, except the path of edges between w_i and its maximal projection. Moreover,

- A node both in T_i and in T_{i-1} is called an *anchor* (that is, a node where the connection path anchors to T_{i-1}).
- A node which is the maximal projection of w_i is called a *foot.*
- All the other nodes are called *path* nodes. The path nodes for which the head daughter is in T_i are said to be *headed.* The path nodes for which the head daughter is not in T_i are said to be *headless.*

Given a tree from the treebank as input, the algorithm scans the sentence covered by this tree in a word-by-word fashion from left to right, and for each word w_i, finds the subset of branches in the tree (the connection path) which would be built in order to attach w_i to the current left context. The algorithm simulates the building of structure by *marking*, during the search of the tree, the branches which would be built by an incremental parser.

The parsing simulation algorithm has two stages.

1. The **first stage** of the algorithm projects up from the new word w_i to find its maximal projection WP. To do this, we maintain a set of rules for de-

termining the maximal projection of a lexical category (these rules were adapted from Magerman and Collins (M&C's rules), who used them in their respective statistical parsers (Magerman, 1995; Collins, 1997)). The algorithm moves up the tree, marking each branch until the maximal projection WP is reached.

2. If there is no left context (i.e. w_i is the first word w_0), the algorithm moves on to the next word, and begins again at the first stage. Otherwise, the algorithm moves into its **second stage**, in which WP will be connected with its current left context. This can happen in one of two ways (see Figure 4 for a schematic illustration):

    a. In the first case, the left context of WP contains one or more incomplete nodes on its right frontier.[1] In this case, the algorithm takes the lowest of these incomplete nodes, XP, which will be an anchor node for WP. If XP is the same node as WP, then the second stage is complete. If XP immediately dominates WP, then the branch between the two nodes is marked, and the second stage is complete. Otherwise, the algorithm descends the leftmost unmarked branch from XP, marking each branch traversed, and saving the path until a node N is reached which *does* immediately dominate WP. Then, the branch is marked, and the second stage is complete.

    b. In the second case, the left context of WP is an orphan node, YP,[2] which has no incomplete nodes on its right frontier. YP will be the anchor node for WP. If WP immediately dominates YP in the input tree, then the branch between WP and YP is marked, and the second stage is complete. If WP does not immediately dominate YP, then the algorithm climbs the tree from YP, marking each branch traversed, and saving the path until a node N is reached which is immediately dominated by WP. Then, the branch between N and WP is marked, and the second stage is complete.

    Because of the presence of phonologically empty nodes, the second stage is sometimes a mixture of cases (a) and (b). After WP is connected to its left context, the connection path is stored in a database. Then the algorithm moves to the next word, and starts again at stage one.

We count the number of headless projections using M&C's rules to determine which is the head daughter of a node in the list of its daughters. These rules are applied to the connection paths in the database after excluding foot and anchor nodes. Such nodes can be excluded from the count because anchor nodes do

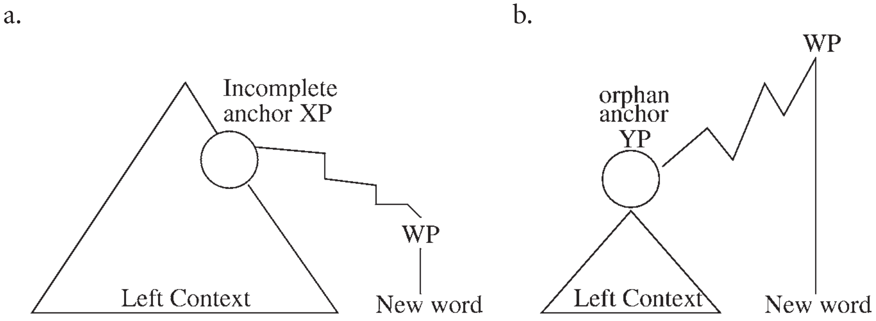


Figure 4. The two cases of connecting the projection of a new word to an anchor in the left context.

not form part of the new structure which has to be built on the input of a new word, while foot nodes, being the projections of words, are headed by definition. Return to Figure 2 to see an example of how the algorithm marks the branches (note that we are assuming an extended lexical projection S for the verb *welcomed* (cf. Grimshaw, 1997)). Figure 3 illustrates an example of labelling a tree containing empty nodes. For the algorithmic details, we refer the reader to the report in (Lombardo & Sturt, 1999).

The algorithm described assumes perfect knowledge. It is accurate enough for our task of estimating the number of headless projections in strongly incremental parsing of language. As noted above, the simulation does not consider the misanalyses that humans make when parsing natural language. In Section 5, we will discuss extensions of the algorithm to deal with certain cases of misanalysis.

4. Experiments and results: The Treebank study

In order to run the simulation algorithm on the treebank, we randomly selected samples of approximately 100,000 words from each of the Wall Street Journal section of the Penn II Treebank, and the Brown corpus section of the previous Penn Treebank release (Marcus et al., 1993). We used two corpora to allow us to assess the cross-corpus reliability of our results. It should be noted here that the Brown Corpus, being from an earlier version of the treebank, used a slightly different notation, and included more markup errors. However, we included this corpus because it gives a more representative sample of English, with text from many different genres. The simulation algorithm was run on these samples, and the full set of connection paths collected. We counted the number of headless projections produced for the input of each word. We did not count empty nodes or their immediately dominating non-terminals as headless projections, since these cannot be seen as projections which are waiting for the input of a head. Finally, we removed punctuation symbols from the corpora before analysis.

4.1 The universe of connection paths

The analysis looks at the full set of connection paths (or *universe* of paths) from the sample, which we partitioned according to the number of headless projections which appear in each path (so, for example, we have the class of paths containing zero headless projections, the class containing one, the class containing two, and so on). Part of the analysis looks at the distributional frequency of each of these classes. This distribution will give us an idea of how many headless projections are typically necessary on the input of a word. Then, we consider the make-up of the paths in terms of the patterns of categories that appear in them. This allows us to see the extent to which headless structures are predictable, and whether systematic patterns can be found.

The number of path tokens in the sample is identical to the number of word tokens (WSJ: 104,989; BROWN: 113,258), because each word requires precisely one connection path. The numbers of distinct path types were 1,896 (WSJ), and 2,307 (Brown).

The most frequent connection path in both corpora corresponds to projecting an NP in the absence of a left context, for example, a sentence initial subject. It can also be noted that of the ten most frequent connection paths in the Brown corpus, there are only two which involve any headless projections, and both of these have only one headless projection. These two connection paths correspond to the attachment of a headless NP (to a PP and a VP respectively), on the input of a determiner. In the WSJ corpus, these two connection paths reduce to only one, namely the PP attachment case (the VP attachment case is not in the top ten).

4.2 Headless projections

Below we give an analysis of the path tokens in terms of the number of headless projections included in each.

Number of headless projections	0	1	2	3	4
Number of path tokens (WSJ Corpus)	86439	16022	2395	133	0
Number of path tokens (Brown Corpus)	88587	22100	2437	133	1

From these results it can be seen that around 80% of path tokens involve no headless projections (WSJ: 82%; BROWN: 78%). This means that in the processing of realistic English, around 80% of the words can be attached to the current representation without involving headless projections. Another striking aspect of the data is that very few of the path tokens involve more than two headless projections (WSJ: 0.13%; BROWN: 0.11%), and four appears to be the absolute limit on the number of headless projections that can appear in a connection path (although, we have found only one case of this, in the Brown Corpus).

We have identified two notable patterns which account for substantial proportions of our headless projections. In the percentages which follow, the number outside the parentheses refers to the WSJ corpus, and the number inside the parentheses refers to the Brown corpus (in fact, it can be seen that the figures are very similar).

The first pattern concerns headless NP projections. Of the headless projections which appear in the path tokens, 41% (40%) consist of a headless NP projection immediately dominating a determiner foot node. In parsing, this corresponds to a situation in which the current word is a determiner, and this word can be connected with the current left context by hypothesising an immediately dominating NP. Note that on the DP hypothesis (cf. Abney, 1987), in which what are traditionally referred to as noun phrases are analysed as determiner phrases, these would not count as headless projections, as the determiner would be classified as the head of the projection. Whether or not the DP hypothesis is assumed, note that a determiner is a good indicator of the existence of a noun phrase, and this could act as lexical guidance for the construction of such connection paths by the human parser. A further 11% (6%) of the headless projections consist of a headless noun phrase which immediately dominates an adjective foot node, which can also be taken as good lexical evidence for the existence of a noun phrase. Taken together, 52% (46%) of the headless projections consist of NP nodes which can be predicted from a determiner or adjective in the current input.

The second common pattern involves clausal complements. We found that 11% (10%) of all headless projections were headless tensed S nodes which could be predicted from lexical material in the left context. These cases involve headless S projections which appear on a path whose anchor is either an immediately dominating SBAR node, or a VP node which dominates an immediately dominating SBAR node with an empty complementizer position. The first case corresponds to a situation in which the headless S node can be predicted by the presence of an overt complementizer in the left context, and the second case corresponds to a case in which the S node can be predicted by the presence of a verb in the left context, which subcategorizes for a clause. In both cases, it is clear that there is good lexical evidence for the presence of an S node.

In parsing terms, the two common patterns which we have outlined above can be thought of in terms of bottom-up and top-down lexical information respectively. In the first case, a headless NP is predicted from bottom-up information associated with the current input word. In the second case, a headless S node is predicted from top-down information associated with the left context. We have found that, of all the path tokens which involve two headless projections, 23% (24%) can be accounted for by a combination of these two common patterns; that is, the two headless projections consist of an NP node which can be predicted from the current input word (a determiner or an adjective), together with an S node which can be predicted from a previously read word (a complementizer or a verb).

These results demonstrate that large scale non-lexical structure building is not necessary in incremental parsing of realistic input, and lexicalism is viable under the strong incrementality assumption. Note also the striking similarity in the data between the two corpora. This suggests that the frequency distributions are stable across corpora, and do not simply reflect idiosyncrasies of a particular genre, for example, that of Wall Street Journal. It is interesting to consider the reasons underlying the distributions which we have found. One possibility is that the lack of connection paths with large numbers of headless projections serves to ease the processing burden on comprehenders. This would be consistent, for example, with minimal attachment (Frazier, 1978), or with theories such as (Gibson, 1991, 1998), which place a memory cost on each headless projection during parsing. Alternatively, and more speculatively, the distribution could reflect the concerns of language production mechanisms; structures with headless projections could impose large memory burdens in planning utterances.

5. Discussion

In this section we will consider some extensions to the algorithm and analyses presented above. First, we discuss the problem of simulating misanalysis. We will then discuss how to enhance the results by considering further linguistic knowledge that can help in predicting headless projections.

As we have mentioned, our results have been based on the simplifying assumption that the parser works with perfect knowledge, and does not simulate the misanalyses that are known to accompany human language processing. As a test case, we will describe an extension to the algorithm which simulates misanalysis and reanalysis in the processing of left-recursive structures. The motivations for including such a treatment of left recursion can be found in Lombardo and Sturt (1997) (see also Stabler, 1994; Thompson, Dixon, & Lamping, 1991 for similar ideas). In left recursive structures, the parser cannot know how deeply embedded the foot node WP is in the connection path, when, given the local context, the legal site can be at more than one level. Consider, for example, the sentence "Urban is the company's first telephone subsidiary in Wisconsin" and the corresponding tree in Fig. 5.[3] To solve the problem of connecting *the company* to its left context in an incremental parse would require guessing how deeply embedded this NP is. However, this embedding can be arbitrarily deep, and there is no way of predicting the depth in advance. Our solution to this problem (Lombardo & Sturt, 1997) relies on the fact that a left-recursive structure, by definition, includes a repeated pattern of symbols, which we call the *Minimal Recursive Structure* (MRS). We assume that the human parser never hypothesises more than one occurrence of each MRS on a connection path, until further MRS's are confirmed by incoming lexical input. A left recursive structure is built through repeated insertions of MRS's onto a connection path. So, in the above example (Fig. 5), the NP immediately dominating *the company's* is initially structured as a daughter of VP (NP_x); when *first* is read, a new NP (NP_y) is built and inserted into the structure above NP_x; when the preposition *in* is read, a third NP (NP_z) is added into the structure above NP_y. Note, therefore, that a left recursive structure is built through an alternating sequence of misanalysis and reanalysis.

To simulate this behaviour on the parsing simulation algorithm, we introduce the notion of temporary links, which simulate the notion of misanalysis, exemplified above by the provisional attachment of NP_x as a daughter of VP. A *temporary link* is a branch between two nodes which are not directly linked in the input tree, but which would be temporarily linked in the incremental tree construction, given the assumptions above on misanalysis. With the introduction of temporary links, we must update the definition of connection path to include that some edge *e* on the right frontier of T_{i-1} is replaced in T_i by a chain of edges having *e*'s nodes as extreme nodes. In Figure 5, it happens that the same T labels appear on nodes that are far apart in the tree. In fact, they are linked by a temporary link at the appropriate point in the simulation. Incidentally, the notion of temporary links has some interesting connections with D-Theory (Description Theory (Marcus, Hindle, & Fleck, 1983). In D-Theory, the parser does not build a fully specified tree, but rather a *description* of a tree, defined in terms of dominance relations (in contrast with traditional immediate dominance relations). At each stage in the parse, new dominance relations are added to the description monotonically, and no existing dominance relation may be deleted. In our simulation algorithm, links in the incremental tree always represent immediate dominance relations. In addition, temporary links have the property of being removable (i.e. non-monotonic). However, a temporary link always corresponds to a non-immediate dominance relation in the global tree from the treebank (as opposed to the incremental tree). If the temporary links were maintained (monotonically), we would have a closer correspondence to D-theory (see also Sturt and Crocker, 1996). The reason for not maintaining the temporary links is purely implementational, since the semantically interpretable incremental tree is immediately recoverable.

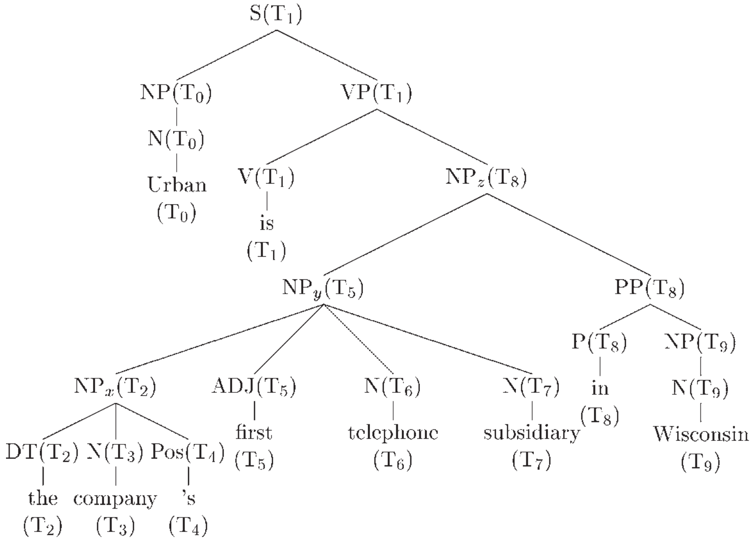


Figure 5. A tree from the treebank for the sentence "Urban is the company's first telephone subsidiary in Wisconsin."

In this table we present the results for the simulation using the treatment of misanalysis in left recursion.

Number of headless projections (LR)	0	1	2	3	4
Number of path tokens (WSJ Corpus)	86330	16555	2048	56	0
Number of path tokens (Brown Corpus)	88534	22524	2136	64	0

It can be seen that the use of the temporary links reduces the count of word tokens whose input requires two or more headless projections (WSJ: 2.40% vs. 2.00%; Brown: 2.27% vs. 1.94%), and those tokens whose input requires three or more headless projections (WSJ: 0.13% vs. 0.05; Brown: 0.11% vs. 0.06%). To see why this is so, consider the insertion of the word *the* into the incremental tree in Figure 5. With the temporary link mechanism, the three NP nodes, NP_x, NP_y and NP_z, on the connection path between VP and *the* are introduced in three different incremental trees, namely T_2, T_5 and T_8, while without the temporary links mechanism, they would all be introduced in the same step (T_2). Thus, in this case, the mechanism reduces the number of headless projections from three to one in attaching *the*.

Another common case of misanalysis that shares some commonalities with left recursion is general left embedding ambiguity, of which the NP/S ambiguity is a typical example, e.g., *John understood the theory was complicated.*, where *the theory* is typically interpreted initially as the direct object of *understood* and subsequently reanalysed as the subject of *was* (see Mitchell, 1994 and references therein). Here, given some knowledge about specific verb subcategorization, we could adopt a similar mechanism for simulating misanalysis and reanalysis of such NP/S sentences.[4] Furthermore, incorporating verb subcategorization into the simulation algorithm would make a number of projections lexically licensed, even if they were headless. Referring again to Figure 1, the SBAR node could be considered licensed by the subcategorization requirements of the verb *thinks*, and would be the anchor of the connection path according to a revised definition.

We are aware that our experiments have a strong dependency on the treebank sentences and syntactic representations. These representations have some shortcomings (Johnson, 1997; Manning & Carpenter, 1997), and may not accurately reflect the representation employed by people in parsing sentences. One of the most evident cases is the flat treatment of VP and NP modification in the Penn treebank. In this case, the results would not be different if the representation were closer to the standard Chomsky-adjunction analysis

(Johnson, 1997), since the further projections introduced by adjunction would all share the same head. In the case of post-modification, these further projections would be on the left corner (so, they would not be headless projections), while in the case of pre-modification, these further projections would be on the right corner, but would all form part of the same head projection, and would therefore not add to the number of headless projections. A different case is the standard structure of noun compounds and genitives, which are often left-recursive, but the head is not in the left corner. In this case, the results would be quite different with the basic algorithm. However, the extension introduced above for the general treatment of left recursion would bring us back to the same results as in the previous table.

The question of whether or not the incidence of non-lexical structure building varies between languages is a very interesting one. A point of particular interest is that connection paths which include headless projections are usually left-branching, and therefore it may be thought that the rarity of headless projections can be attributed to the fact that left-branching structures are, in any case, rare in English (Sampson, 1997; Yngve, 1960). In order to test this claim, we believe it would be informative to conduct investigations of the kind described in this paper on languages in which left-branching structures are very common, such as Japanese and Turkish. We suspect that a substantial number of the left branching structures in these languages involve left recursion, and are built up step-by-step by the human parser, rather than forming single connection paths.

6. Conclusions

This paper has presented the results of a treebank study devoted to assess a quantitative analysis of the amount of non-lexical structure building needed by a strongly incremental parser. The study suggests that the amount of syntactic knowledge required over and above the maximal projections of lexical items is limited. Furthermore, headless projections can in many cases be predicted from the combinatorial properties of lexical items.

Acknowledgements

We acknowledge Michael Collins, who kindly provided us with his rules for computing the head-daughter relations. We would also like to thank Ted Gibson and an anonymous reviewer for their helpful comments. This research was supported by bilateral collaboration grants awarded by the British Council and the Consiglio Nazionale delle Ricerche in 1997 and 1998. The order of authors is alphabetical.

Notes

[1] An *incomplete* node refers to node which has not yet been connected to all its daughter nodes (i.e. in our terms, a node with at least one unmarked daughter branch).

[2] An *orphan node* refers to a node which currently has no mother (i.e. in our terms, a node whose mother branch has not yet been marked).

[3] The flat structure for the possessive NP follows the conventions of the Penn Treebank bracketing style. However, a more conventional analysis would pose the same problem. We will return to the issue of representation later in this section.

[4] D-theory (see above) raises the possibility that incremental tree descriptions are interpreted in an underspecified manner. However, there exists psychological evidence against the underspecification approach to semantic interpretation in such cases of misanalysis (Pickering & Traxler, 1998). For a general discussion of such issues, see Sturt (1997).

References

Abney, S.P. (1987). *The English noun phrase in its sentential aspect*. Ph.D. thesis, MIT, Cambridge, MA.

Abney, S.P. (1989). A computational model of human parsing. *Journal of Psycholinguistic Research, 18*(1), 129–144.

Bader, M. & Lasser, I. (1994). German verb-final clauses and sentence processing. In Clifton, C., Frazier, L., & Rayner, K. (Eds.), *Perspectives on Sentence Processing* (p. 225–242). Lawrence Erlbaum Associates, New Jersey.

Chomsky, N. (1995). *The minimalist program*. MIT Press, Cambridge, MA.

Collins, M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th annual meeting of the Association for Computational Linguistics*, pp. 16–23.

Crocker, M.W. (1996). *Computational Psycholinguistics: An Interdisciplinary Approach to the Study of Language*. Studies in Theoretical Psycholinguistics 20. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Eberhard, K.M., Spivey-Knowlton, M.J., Sedivy, J.C., & Tanenhaus, M.K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research, 24*(6), 409–436.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies*. Ph.D. thesis, University of Connecticut.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language, 26*, 505–526.

Garnsey, S.M., Pearlmutter, N.J., Myers, E., & Lotocky, M.A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.

Gibson, E.A.F. (1991). *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing breakdown*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Gibson, E.A.F. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76.

Gorrell, P. (1995). *Syntax and Parsing*. Cambridge University Press, Cambridge, England.

Grimshaw, J. (1997). Projection, heads, and optimality. *Linguistic Inquiry, 28*(3).

Inoue, A., & Fodor, J.D. (1995). Information-paced parsing of Japanese. In Mazuka, R., & Nagai, N. (Eds.), *Japanese Sentence Processing*, chap. 2, pp. 9–63. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Johnson, M. (1997). The effect of alternative tree representations on treebank grammars. In *Proceedings of NeMLAP 3 (New Empirical Methods in Language Processing)*, Sydney, pp. 39–48.

Konieczny, L. (1996). *Human Sentence Processing: A Semantics-Oriented Approach*. Ph.D. thesis, University of Freiburg.

Lombardo, V., Lesmo, L., Ferraris, L., & Seidenari, C. (1998). Incremental processing and lexicalized grammars. In *Proceedings of the XXI Annual Meeting of the Cognitive Science Society*.

Lombardo, V. & Sturt, P. (1997). Incremental processing and infinite local ambiguity. In *Proceedings of the 19th Annual Conference of the Cognitive Science Society, Stanford CA*, pp. 448–453.

Lombardo, V. & Sturt, P. (1999). An algorithm for simulating incremental parsing of a treebank. Technical report, Università di Torino. URL ftp://ftp.cogsci.ed.ac.uk/pub/sturt/papers/tr99.ps.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). Lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.

Magerman, D. (1995). Statistical decision-tree models for parsing. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp. 276–283.

Manning, C.D., & Carpenter, B. (1997). Probabilistic parsing using left-corner language models. In *Proceedings of the 5th International Workshop on Parsing Technologies* Boston, MA, pp. 147–158.

Marcus, M., Hindle, D., & Fleck, M. (1983). D-theory: Talking about talking about trees. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pp. 129–136. Cambridge, MA.

Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics, 19*, 313–330.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–533.

Merlo, P. (1996). *Parsing with principles and classes of information*. Studies in Linguistics and Philosophy. Kluwer, Dordrecht.

Mitchell, D.C. (1994). Sentence parsing. In Gernsbacher, M.A. (Ed.), *Handbook of Psycholinguistics*, pp. 375–410. Academic Press, San Diego, CA.

Pickering, M.J., & Traxler, M.J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(4), 940–961.

Pollard, C., & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. CSLI and University of Chicago Press, Stanford, Ca. and Chicago, Ill.

Pritchett, B.L. (1991). Head position and parsing ambiguity. *Journal of Psycholinguistic Research, 20*(3), pp. 251–270.

Sampson, G. (1997). Depth in English grammar. *Journal of Linguistics, 33*, 131–151.

Shieber, S., & Johnson, M. (1993). Variations on incremental interpretation. *Journal of Psycholinguistic Research, 22*(2), 287–318.

Stabler, E.P. (1991). Avoid the pedestrian's paradox. In Berwick, R.C., Abney, S.P., & Tenny, C. (Eds.), *Principle based parsing: computation and psycholinguistics*. Kluwer, Boston, MA.

Stabler, E.P. (1994). The finite connectivity of linguistic structure. In Clifton, C., Frazier, L., & Rayner, K. (Eds.), *Perspectives on Sentence Processing*, chap. 13, pp. 303–336. Lawrence Erlbaum.

Steedman, M.J. (1989). Grammar, interpretation and processing from the lexicon. In Wilson, W.M. (Ed.), *Lexical Representation and Process*, chap. 16, pp. 463–504. MIT Press.

Sturt, P. (1997). *Syntactic Reanalysis in Human Language Processing*. Ph.D. thesis, Centre for Cognitive Science, University of Edinburgh, Edinburgh, Scotland.

Sturt, P., & Crocker, M.W. (1996). Monotonic syntactic processing: a cross-linguistic study of attachment and reanalysis. *Language and Cognitive Processes, 11*(5), 449–494.

Thompson, H., Dixon, M., & Lamping, J. (1991). Compose-reduce parsing. In *Proceedings of the 29th Meeting of the Association for Computational Linguistics*, pp. 87–97. Berkeley, California.

Trueswell, J.C., Tanenhaus, M.K., & Kello, C. (1993). Verb-specific constraints on sentence processing: Separating effects of lexical preference from garden paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 528–553.

Yamashita, K. (1994). *Processing of Japanese and Korean*. Ph.D. thesis, Ohio State University, Columbus, Ohio.

Yngve, V.H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*, 444–466.

Modular architectures and statistical mechanisms

The case from lexical category disambiguation[1]

Matthew W. Crocker and Steffan Corley

Computational Linguistics, Saarland University, Saarbrücken, Germany
and Sharp Laboratories of Europe, Oxford Science Park, Oxford, U.K.

This paper reviews the modular, statistical model of human lexical category disambiguation (SLCM) proposed by Corley and Crocker (2000). The SLCM is distinct lexical category disambiguation mechanism within the human sentence processor, which uses word-category frequencies and category bigram frequencies for the initial resolution of category (part-of-speech) ambiguities. The model has been shown to account for a range of existing experimental findings in relatively diverse constructions. This paper presents the results of two new experiments that directly confirm the predictions of the model. The first experiment demonstrates the dominant role of word-category frequency in resolving noun-verb ambiguities. The second experiment then presents evidence for the modularity of the mechanism, by demonstrating that immediately available syntactic context does not override the SLCMs initial decision.

Introduction

This paper reconsiders the nature of modular architectures in the light of recent empirical, theoretical and computational developments concerning the exploitation of statistical language processing mechanisms. We defend a simpler notion of modularity than that proposed by Fodor (1983). Given current conflicting theoretical arguments and empirical evidence for and against modularity, we argue for modularity strictly on computational and methodological grounds. We then apply this to a particular aspect of human language processing: the problem of lexical category disambiguation.

While previous work has often focused on the kinds of linguistic knowledge which are used in ambiguity resolution, we focus on the role of statistical, or frequency-based, knowledge. While such mechanisms are now a common element of non-modular, constraint-based models (see Tanenhaus et al., 2000), we argue that probabilistic mechanisms may be naturally associated with modular architectures. In particular, we suggest that a Statistical Lexical Category Module (SLCM) provides an extremely efficient and accurate solution to the sub-problem of lexical category disambiguation. Following a summary of the model and how it accounts for the range of relevant existing data, we review the results of two new experiments that test the predictions of both the statistical and modular aspects of the SLCM, and provide further support for our proposals.

Modularity, constraints and statistics

The issue of modularity continues to be a hotly debated topic within the sentence processing literature.[2] Parser-based models of human sentence processing led to the tacit emergence of syntactic modularity, which was then rationally defended by Fodor (1983). In particular, Fodor argued that cognitive faculties are divided into input processes, which are modular, and central processes, which are not. The divide between input and central processes is roughly coextensive with the divide between perception and cognition; in the case of language, Fodor located this divide between the subject matter of formal linguistics and that of pragmatics and discourse analysis.

Recently, their has been a shift in consensus towards more interactionist, non-modular positions. The term 'constraint-based' is often used to denote such an interactionist position. The constraint-based position is tacitly assumed to imply that all constraints can in principle apply immediately and simultaneously, across all levels of linguistic representation, and possibly even across perceptual faculties (Tanenhaus et al., 1995).

Modular and interactive positions are often associated with other computational properties. Spivey-Knowlton and Eberhard (1996) argue that modular positions tend to be symbolic, binary, unidirectional and serial. In contrast, interactive models tend to be distributed, probabilistic, bi-directional and parallel. Further, Spivey-Knowlton and Eberhard suggest that "when a model is specified in enough detail to be associated with a region in this space, that region's projection onto the continuum of modularity indicates *the degree to which* a model is modular" (pp. 39–40, their italics).

Spivey-Knowlton and Eberhard's position turns a historical accident into a definition. While existing models do pattern approximately along the lines they propose, we suggest that their characterisation inaccurately represents the underlying notion of modularity.[3] We propose a simplified definition of modularity that is independent of any commitment to orthogonal issues such as the symbolic-distributed, binary-probabilistic, unidirectional-bidirectional and serial-parallel nature of a particular theory. Rather our definition focuses purely on information-flow characteristics:

- A module can only process information stated in its own representational and informational vocabulary. For example, the syntactic processor can only make use of grammatical information.
- A module is independently predictive. That is, we do not need to know about any other component of the cognitive architecture to make predictions about the behaviour of a module (provided we know the module's input).
- A module has low bandwidth in both feedforward and feedback connections. By this we mean that it passes a comparatively small amount of information (compared to its internal bandwidth) on to subsequent and prior modules.

These three defining properties of a modular architecture overlap. If one module cannot understand the representational vocabulary of another, then information about its internal decision process is of no use; thus the cost of passing such information on would not be warranted. Similarly, a module cannot be independently predictive if its decisions depend on representations constructed by other modules that are not part of its input – independent prediction is therefore directly tied to low bandwidth feedback connections.

In sum, we propose a simple definition of modularity in which modules process a specific representation and satisfy the relevant constraints which are defined for that level of representation. Modules have high internal bandwidth and are connected to each other by relatively low bandwidth: the lower the bandwidth, the greater the modularity. This definition is independent of whether we choose to state our modules in more distributed or symbolic terms, as it should be.

Statistical mechanisms

In the previous section, we noted Spivey-Knowlton and Eberhard's (1996) claim that modularity is normally associated with binary rather than probabilistic decision procedures. This claim derives largely from the association of constraint-based architectures with connectionist implementations (Tanenhaus et al., 2000; MacDonald et al., 1994) which in turn have a natural tendency to exhibit frequency effects. We proposed a definition of modularity which is consistent with statistical mechanisms. In this section, we argue that modularity and statistical mechanisms are in fact natural collaborators.

The motivation for modularity is essentially one of computational compromise, based on the assumption that an unrestricted constraint-satisfaction procedure could neither operate in real-time (Fodor, 1983), nor could it acquire such a heterogeneous system of constraints in the first place (Norris, 1990). It is still reasonable to assume however, that modules will converge on highly effective processing mechanisms; that is, a mechanism which can accurately and rapidly arrive at the correct analysis of the input, based on the restricted knowledge available within the module. For purposes of disambiguation, the module should therefore use the best heuristics it can, again modulo any computational and informational limitations.

In the spirit of rational analysis (Anderson, 1991), one might therefore choose to reason about such a mechanism as an optimal process in probabilistic terms. This approach has been exploited both in the study of human sentence processing (Chater et al., 1998; Jurafsky, 1996) and in computational linguistics where statistical language models have been effectively applied to problems of speech recognition, part-of-speech tagging, and parsing (see Charniak (1993; 1997) for an overview). We propose a specific hypothesis, in which modules may make use of statistical mechanisms in their desire to perform as effectively as possible in the face of restricted knowledge. We define statistical modularity by introducing the 'Modular Statistical Hypothesis' (MSH):

> **The Modular Statistical Hypothesis**: The human sentence processor is composed of a number of modules, at least some of which use statistical mechanisms. Statistical results may be communicated between modules, but statistical processes are restricted to operating within, and not across, modules.

This hypothesis encompasses a range of possible models, including the coarse-grained architecture espoused by proponents of the Tuning Hypothesis (Mitchell et al., 1995; Mitchell & Brysbaert, 1998). However, it excludes interactive models such as those proposed by MacDonald et al. (1994), Tanenhaus et al. (2000) and Jurafsky (1996) – despite their probabilistic nature – since the models that fall within the MSH are a necessarily subset of those that are modular.

In the case of a statistical module we assume that heuristic decision strategies are based on statistical knowledge accrued by the module, presumably on the basis of linguistic experience. Assuming that the module collates statistics itself, it must have access to some measure of the 'correctness' of its decision; this could be informed by whether or not reanalysis was requested by later processes. The most restrictive modular statistical model is therefore one in which modules are fully encapsulated and only offer a single analysis to higher levels of processing.

The statistical measures such a module depends on are thus architecturally limited. Such measures can not directly reflect information pertaining to higher levels of processing, as these are not available to the module. Assuming very low bandwidth feedforward connections, or shallow output, it is also impossible for the module to collate statistics concerning levels of representation that are the province of modules that precede it. A modular architecture therefore constrains the representations for which statistics may be accrued, and subsequently used to inform decision making processes; this contrasts with an interactive architecture, where there are no such constraints on the decision process.

It is worth noting that we have argued for the use of statistical mechanisms in modular architectures on primarily *rational* grounds. That is, such statistical mechanisms have been demonstrated to provide highly effective heuristic decisions in the absence of full knowledge, and their use is therefore highly strategic, not accidental. Indeed, it might even be argued that such mechanisms give good approximations of 'higher-level' knowledge. For example, simple word bigrams will model those words that co-occur frequently or infrequently. Since highly semantically plausible collocations are likely to be more frequent than less plausible ones, such statistics can appear to be modelling semantic knowledge, as well as just the distribution of word types.

In contrast, constraint-based, interactionist models motivate the existence of frequency effects as an essentially unavoidable consequence of the underlying connectionist architecture (see Seidenberg (1997) for general discussion), along with other factors such as neighbourhood effects. Interestingly, this may lead to some rather strong predictions. Since such mechanisms are highly sensitive to frequency, they would seem to preclude probabilistic mechanisms that do not select a "most-likely" analysis based on these prior frequencies. Pickering et al. (2000), however, present evidence against likelihood-based accounts, and propose and alternative probabilistic model based on a rational analysis of the parsing problem (Chater et al., 1998).

Lexical category ambiguity

The debate concerning the architecture of the human language processor has typically focused on the syntax-semantics divide. Here, however, we consider the problem of lexical category ambiguity, and argue for the plausibility of a distinct lexical category disambiguation module. Lexical category ambiguity occurs when a word can be assigned more than one part of speech (noun, verb, adjective etc.). Consider, for example, the following sentence:

(1) He saw her duck.

There are two obvious, plausible readings for sentence 1. In one reading, 'her' is a possessive pronoun and 'duck' is a noun (cf. 2a); in the other reading, 'her' is a personal pronoun and 'duck' is a verb (cf. 2b).

(2) a. He saw her$_{\text{POSS}}$ apple.
b. He saw her$_{\text{PRON}}$ leave.

Lexical Category Ambiguity and Lexical Access

Lexical access is the stage of processing at which lexical entries for input words are retrieved. Evidence suggests that multiple meanings for a given word are activated even when semantic context biases in favour of a single meaning (Swinney, 1979; Seidenberg et al., 1982; but see Kawamoto (1993) for more thorough discussion). The evidence does not, however, support the determination of grammatical class during lexical access. Tanenhaus, Leiman and Seidenberg (1979) found that when subjects heard sentences such as those in (3), containing a locally ambiguous word in an unambiguous syntactic context, they were able to name a target word which was semantically related to either of the possible meanings of the ambiguous target (e.g. SLEEP or WHEEL) faster than they were able to name an unrelated target.

(3) a. John began to tire.
b. John lost the tire.

This suggests that words related to both meanings had been primed; both meanings must therefore have been accessed, despite the fact that only one was compatible with the syntactic context. Seidenberg, Tanenhaus, Leiman and Bienkowski (1982) replicated these results, and Tanenhaus and Donnenworth-Nolan (1984) demonstrated that they could not be attributed to the ambiguity (when spoken) of the word 'to' or to subjects inability to integrate syntactic information fast enough prior to hearing the ambiguous word.

Such evidence is consistent with a model in which lexical category disambiguation occurs after lexical access. The tacit assumption in much of the sentence processing literature has been that grammatical classes are determined during parsing (see Frazier (1978) and Pritchett (1992) as examples). If grammar terminals are words rather than lexical categories, then such a model requires no augmentation of the parsing mechanism. Alternatively, Frazier and Rayner (1987) proposed that lexical category disambiguation has a privileged status within the parser; different mechanisms are used to arbitrate such ambiguities from those concerned with structure building.

Finally, lexical categories may be determined after lexical access, but prior to syntactic analysis. That is, lexical category disambiguation may constitute a module in its own right.

The Privileged Status of Lexical Category Ambiguity

There are essentially three possible positions regarding the relationship between syntax and lexical category.

1. Lexical categories are syntactic: The terminals in the grammar are words and it is the job of the syntactic processes to determine the lexical category that dominates each word (Frazier, 1978; Pritchett, 1992).
2. Syntactic structures are in the lexicon: The bulk of linguistic competence is in the lexicon, including rich representations of the trees projected by lexical items. Parsing is reduced to connecting trees together (MacDonald et al., 1994; Kim and Trueswell, this volume).
3. Syntax and lexical category determination are distinct: Syntax and the lexicon have their own processes responsible for initial structure building and ambiguity resolution.

If we take the latter view of lexical category ambiguities, one possibility is that a pre-syntactic modular process makes lexical category decisions. These decisions would have to be made on the basis of a simple heuristic, without the benefit of syntactic constraints. In common with all modules, such a process will make incorrect decisions when potentially available information (such as syntactic constraints) could have permitted a correct decision. It does, however, offer an extremely low cost alternative to arbitration by syntactic and other knowledge. That is, disambiguation on the basis of full knowledge potentially entails the integration of constraints of various types, across various levels of representation. It may be the case that such processes cannot converge rapidly enough on the correct disambiguated form.

For this argument to be compelling, it must also be the case that lexical category ambiguities are frequent enough to warrant a distinct resolution process. This can be verified by determining the number of words that occur with more than one category in a large text corpus. DeRose (1988) has produced such an estimate from the Brown corpus; he found that 11.5% of word types and 40% of tokens occur with more than one lexical category. As the mean length of the sentences in the Brown corpus is 19.4 words, DeRose's figures suggest that there are 7.75 categorially ambiguous words in an average corpus sentence.

Our own investigations suggest the extent of the problem is even greater. Using the TreeBank version of the Brown corpus, we discovered 10.9% ambiguity by type, and a staggering 65.8% by token. To obtain these results, we used the coarsest definition of lexical category possible – just the first letter of the corpus tag (i.e. nouns were not tagged separately as singular, plural, etc.). Given the high frequency of lexical category ambiguity, a separate decision making process makes computational sense, if it can achieve sufficient accuracy. If category ambiguities are resolved prior to parsing, the time required by the parser is reduced (Charniak et al., 1996).

A Statistical Lexical Category Module

In this section we outline a specific proposal for a Statistical Lexical Category Module (SLCM). The function of the SLCM is to determine the best possible assignment of lexical part-of-speech categories for the words of an input utterance, as they are encountered. The model differs from other theories of sentence processing, in that lexical category disambiguation is postulated as a distinct modular process, which occurs prior to syntactic processing but following lexical access.

We argued earlier for a model of human sentence processing that is (at least partially) statistical on both rational and empirical grounds: such a model appears sensible and has characteristics which may explain some of the behaviour patterns of the HSPM. We therefore propose that the SLCM employs a statistically-based disambiguation mechanism, as such a mechanism can operate efficiently (in linear time) and achieve near optimal performance (most words disambiguated correctly, see next section), and we assume such a module would strive for such a rational behaviour.

What statistics?

If we accept that the SLCM is statistical, a central question concerns what statistics condition its decisions. Limitations of the modular architecture we are proposing constrain the choice. The SLCM has no access to structural representations; structurally-based statistics could therefore not be expressed in its representational vocabulary. We will assume that the input to the module is extremely shallow – just a word and a set of candidate grammatical classes. In this case, the module also has no access to low level representations including morphs, phonemes and graphic symbols; the module may only make use of statistics collated over words or lexical categories, or combinations of the two.

It seems likely that the SLCM collates statistics concerning the frequency of co-occurrence of individual words and lexical categories. One possible model is therefore that the SLCM just picks the most frequent class for each word; for reasons that will become apparent, we will call this the ‘unigram’ approach. The SLCM may also gather statistical information concerning prior context. For example, decisions about the most probable lexical category for a word may also consider the previous word. Alternatively, such decisions may only consider the category assigned to the previous word, or a combination of both the prior word and its category may be used.

Probability theory and the SLCM

The problem faced by the SLCM is to incrementally assign the most likely sequence of lexical categories to a given sequence of words as they are encountered. That is, as each word is input to the SCLM, it outputs the most likely category for it. Research in computational linguistics has concentrated on a (non-incremental) version of this problem for a number of years and a number of successful and accurate ‘part-of-speech taggers’ have been built (e.g. Weischedel et al., 1993; Brill, 1995). While a number of heuristic tagging algorithms have been proposed, the majority of modern taggers are statistically based, relying on distributional information about language (DeRose, 1988; Weischedel et al., 1993; Ratnaparkhi, 1996; see also Charniak, 1997 for discussion). It is this set of taggers that we suggest is most suitable for an initial model of statistical lexical category disambiguation. They provide a straightforward learning algorithm based on prior experience, are comparatively simple, employ a predictive and uniform decision strategy (i.e. don't make use of arbitrary or ad hoc rules), and can be naturally adapted to assign preferred lexical category tags incrementally.

The SLCM, as with part-of-speech taggers, is based on a Hidden Markov Model (HMM), and operates by probabilistically selecting the best sequence of category assignments for an input string of words.[4] Let us briefly consider the problem of tag assignment from the perspective of probability theory. The task of the SLCM is to find the best category sequence ($t_1 \ldots t_n$) for an input sequence of words ($w_1 \ldots w_n$). We assume that the 'best' such sequence is the one that is most likely, based on our prior experience. Therefore the SLCM must find the sequence ($t_1 \ldots t_n$) such that $\mathrm{P}(t_1 \ldots t_n, w_1 \ldots w_n)$ is maximised. That is, we want to find the tag sequence that maximises the joint probability of the tag sequence and the word sequence.

One practical problem, however, is that determining such a probability directly is difficult, if we wish to do so on the basis of frequencies in a corpus (as in the case of taggers) or in our prior experience (as would be the case for the psychological model). The reason is that we may have seen very few (or quite often no) occurrences of a particular word-tag sequence, and thus probabilities will often be estimated as zero. It is therefore common practice to approximate this probability with another which can be estimated more reliably. Corley and Crocker (2000) argue that the SLCM approximates this probability using category bigrams, as follows:

$$P(t_0, \ldots t_n, w_0, \ldots w_n) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1})$$

The two terms in the right hand side of the equation are the two statistics that we hypothesise to dominate lexical category decisions in the SLCM. $P(w_i|t_i)$ – the unigram or word-category probability – is the probability of a word given a particular tag.[5] $P(t_i|t_{i-1})$ – the bigram or category co-occurrence probability – is the probability that two tags occur next to each other in a sentence. While the most accurate HMM taggers typically use trigrams (Brants, 1999), Corley and Crocker (2000) argue that the bigram model is sufficient to explain existing data and is simpler (requires fewer statistical parameters). It is therefore to be preferred as a cognitive model, until evidence warrants a more complex model.

Estimates for both of these terms are typically based on the frequencies obtained from a relatively small training corpus in which words appear with their correct tags. This equation can be applied incrementally. That is, after perceiving each word we may calculate a contingent probability for each tag

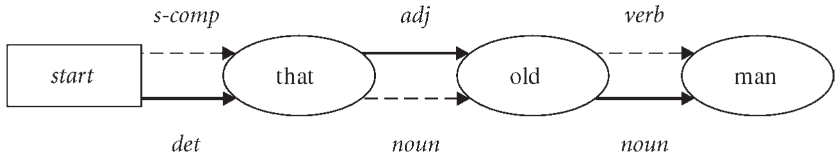


Figure 1. Tagging the sequence “that old man”

path terminating at that word; an initial decision may be made as soon as the word is seen. Figure 1 depicts tagging of the phrase “that old man”. Each of the words has two possible lexical categories, meaning that there are eight tag paths. In the diagram, the most probable tag path is shown by the sequence of solid arcs. Other potential tags are represented by dotted arcs.

The tagger’s job is to find this preferred tag path. The probability of a sentence beginning with the start symbol is 1.0. When ‘that’ is encountered, the tagger must determine the likelihood of each reading for this word when it occurs sentence initially. This results in probabilities for two tag paths – start followed by a sentence complementiser and start followed by a determiner. The calculation of each of these paths is shown in Table 1.

Table 1. Tagging “that old man”; stage 1 – “that”

Path	Probability
1 scomp	*P (“that”|scomp) P (scomp|start)*
2 det	*P (“the”|scomp) P (det|start)*

While “that” occurs more frequently as a sentence complementiser than as a determiner in absolute terms, sentence complementisers are relatively uncommon at the beginning of a sentence. Therefore tag path 2 is likely to have a greater probability.

The next word, “old”, is also category ambiguous as either an adjective or a noun. There are therefore four possible tag paths up until this point. Table 2 shows the calculations necessary to determine the probability of each of them.

Table 2. Tagging “that old man”; stage 2 – “old”

Path	Probability
1.1 scomp-adj	*P (“old”|adj) P (adj|scomp) P (path1)*
1.2 scomp-noun	*P (“old”|noun) P (noun|scomp) P (path1)*
2.1 det-adj	*P (“old”|adj) P (adj|det) P (path2)*
2.2 det-noun	*P (“old”|noun) P (noun|det) P (path2)*

In this case, "old" is far more frequently an adjective than a noun, and so this is the most likely reading. As an adjective following a determiner is more likely than one following a sentence complementiser, path 2.1 becomes far more probable than 1.1.

The process is identical when "man" is encountered. There are now eight tag paths to consider, shown in Table 3. As "man" occurs more frequently as a noun than a verb, and this reading is congruent with the preceding context, path 2.1.2 is preferred.

Table 3. Tagging "that old man"; stage 3 – "man"

Path	Probability
1.1.1 scomp-adj-verb	*P ("man"|verb) P (verb|adj) P (path1.1)*
1.1.2 scomp-adj-noun	*P ("man"|noun) P (noun|adj) P (path1.1)*
1.2.1 scomp-noun-verb	*P ("man"|verb) P (verb|noun) P (path1.2)*
1.2.2 scomp-noun-noun	*P ("man"|noun) P (noun|noun) P (path1.2)*
2.1.1 det-adj-verb	*P ("man"|verb) P (verb|adj) P (path2.1)*
2.1.2 det-adj-noun	*P ("man"|noun) P (noun|adj) P (path2.1)*
2.2.1 det-noun-verb	*P ("man"|verb) P (verb|noun) P (path2.2)*
2.2.2 det-noun-noun	*P ("man"|noun) P (noun|noun) P (path2.2)*

So far, we have assumed that it is necessary to keep track of every single tag path. This would make the algorithm extremely inefficient and psychologically implausible; as the length of the sentence grows, the number of possible tag paths increases exponentially. However, a large number of paths which will never be 'most probable' can rapidly be discarded, using a standard dynamic programming solution – the Viterbi (1967) algorithm (see Charniak, 1993 for explanation). This algorithm is linear; this means that the amount of work required to determine a tag for each word is essentially constant, no matter how long the sentence is. Indeed, this property contributes directly to the psychological plausibility of this mechanism over more complex alternatives.

We have argued that taggers such as the SLCM are, in general, extremely accurate (approaching 97% – see Charniak, 1997; Brants, 1999). However, they have distinctive breakdown and repair patterns. Corley and Crocker (2000) argue that these patterns are very similar to those displayed by people upon encountering sentences containing lexical category ambiguities. In particular, they show how the SLCM, when trained on a standard corpus of English, models the following experimental results:

'That' Ambiguity (Juliano & Tanenhaus, 1993). In this study, Juliano and Tanenhaus investigated the initial decisions of the HSPM when it encounters

the categorially ambiguous word "that", in both sentence initial and post verbal contexts. In sentence initial position, "that" is more likely to be a determiner, while post-verbally, it is more likely to be a complementiser. Corley and Crocker provide a simulation demonstrating that the proposed bigram model accounts for the findings, while a simpler unigram model does not.

Noun-Verb Ambiguities (MacDonald, 1993). Following the study of Frazier and Rayner (1987), MacDonald investigated the processing of words that are ambiguous between noun and verb categories, e.g. as in "warehouse *fires*", to determine if semantic bias affected initial decisions. Corley and Crocker show how the SLCM can straightforwardly account for the findings. This is discussed in more detail in the next section.

Post-Ambiguity Constraints (MacDonald, 1994). Reanalysis may occur in the SLCM when the most probable tag sequence at a given point requires revising an adjacent, previous tag. Corley and Crocker (2000) demonstrate how such reanalysis in the SLCM can simulate the post-ambiguity constraints investigated by MacDonald, in which reduced relative clause constructions were rendered easier to process when the word following the ambiguous verb (simple past vs. participle) made the participle reading more likely.

New evidence for the SLCM

The Modular Statistical Hypothesis posits the existence of identifiable subsystems within the human language processor, and argues for the use of statistical mechanisms within modules as optimal heuristic knowledge. For the task of lexical category disambiguation, we have presented a particular modular statistical mechanism. While our model accounts well for a range of relevant existing findings, as outlined in the previous section, many of those results were based on experiments designed to test rather different hypotheses, and as such provide imperfect and indirect support for the mechanism we have developed.

In this section we review two recent experimental results from Corley (1998) which directly test the central predictions of the theory. These predictions are:

- The **Statistical Lexical Category Hypothesis** (**SLCH**): Initial lexical category decisions are made on the basis of frequency-based statistics.
- The **Modular Lexical Category Hypothesis** (**MLCH**): Lexical category decisions are made by a pre-syntactic module.

Experiment 1 is concerned with the SLCH; it is designed to determine whether initial lexical category decisions are affected by the statistical bias of individual words. Experiment 2 more directly tests the MLCH; the experiment determines whether initial decisions are made on the basis of lexical statistical bias even in the face of strong syntactic evidence to the contrary.

Experiment 1: The statistical lexical category hypothesis

Words that are ambiguous between noun and verb readings are very common in English. Frazier and Rayner (1987) and MacDonald (1993) both employed this ambiguity in their experiments; their results were taken as support for the delay strategy and an interactive constraint-based view respectively. The SLCH simply asserts that the initial decisions of the HSPM will be strongly influenced by frequency-based statistics. For this ambiguity, all other things being equal, the HSPM will initially prefer a noun reading for a word that is frequency-biased towards a noun reading, and a verb reading for a verb-biased one.[6]

Previous studies of this ambiguity have not fully tested this hypothesis. For example, MacDonald's (1993) experimental items included only noun-biased words. In contrast, Corley (1998) produced a controlled set of experimental items in which both noun-biased and verb-biased conditions were represented. Example materials are shown below.

Experiment 1: Materials

a. The woman said that the German *makes* the beer she likes best.
b. The woman said that the German *makes* are cheaper than the rest.
c. The foreman knows that the warehouse *prices* the beer very modestly.
d. The foreman knows that the warehouse *prices* are cheaper than the others.

In (a) and (b), the ambiguous word ("makes") is biased towards a verb reading. In (a) the disambiguating region ("the beer") also favours this reading. In contrast, the disambiguating region in (b) favours a noun reading. (c) and (d) are analogous except that the ambiguous word is noun-biased.

The frequency bias of each of the ambiguous words used in this experiment was determined from the British National Corpus, chosen for both its size (100 million words) and its relatively balanced and British content. As this experiment is only designed to test whether statistical bias does have an effect, and not whether other constraints do not, only strongly biased items were used. The experimental items were further controlled to ensure that the possible noun compounds ("German makes", "warehouse prices") were plausible but infrequent and non-idiomatic. This control ensured that contextual bias effects (MacDonald, 1993) would not be expected to influence the outcome of the experiment.

If the SLCH is correct, reading times in the disambiguating region should reflect an interaction between bias and disambiguation. In other words, subjects' initial decisions should depend on the bias of the ambiguous words; we would therefore expect reading time increases reflecting reanalysis to occur only when the disambiguating region forces a reading at odds with the bias of the ambiguous word.

In contrast, a non-statistical model such as the Garden Path theory (Frazier, 1979) predicts the same initial decision in all four conditions. A main effect of disambiguation would be anticipated, but not one of bias, and no interaction between bias and disambiguation. Frazier and Rayner's (1987) delay strategy also does not predict a main effect of bias or an interaction; any main effect of disambiguation is compatible with, rather than predicted by, the strategy.

32 subjects took part in the experiment, which was performed as a self-paced reading study, using a moving window display (Just, Carpenter and Woolley, 1982). The resulting reading times were adjusted for word length using a procedure described in Ferreira and Clifton (1986).

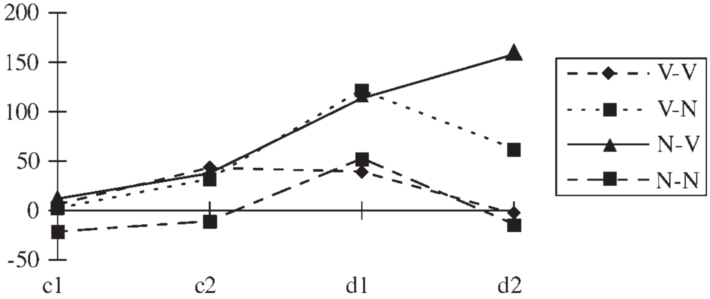


Figure 2. Experiment 1 length-adjusted reading times

Results and discussion

Average length-adjusted reading times obtained for experiment 1 are shown in Figure 2. Here, c1 is the word preceding the ambiguous word, c2 is the ambiguous word and $d_1 \ldots d_n$ is the disambiguating region. V-V indicates that c2 is verb biased, and that the item is disambiguated as a verb, and so on.

The SLCH predicts effects at the start of the disambiguating region; the results for the first word of the disambiguating region are shown in Figure 3. These results show a highly significant interaction between bias and disambiguation ($F_1 = 8.05, p < .01; F_2 = 27.99, p < .001$). A planned comparison of means also revealed a highly significant difference in reading times between the verb disambiguation conditions ($F_1 = 8.27, p < .01; F_2 = 10.86, p < .01$) and a significant difference between the noun disambiguation conditions ($F_1 = 4.72, p < .05; F_2 = 7.46, p < .02$).

These results indicate that initial lexical category decisions are strongly influenced by the frequency-bias of the individual ambiguous words; the results are exactly as predicted by the SLCH and therefore provide very strong support for it. They are not compatible with any non-statistical model, including the delay strategy.

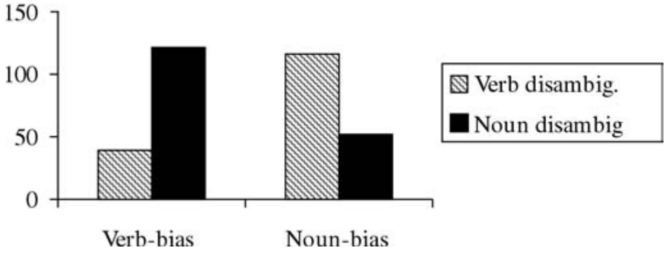


Figure 3. Experiment 1 results for the first word of the disambiguating region (d_1)

Experiment 2: The modular lexical category hypothesis

The SLCH posits that initial lexical category decisions are made on the basis of frequency-based preferences. It does not require that no other constraints influence these decisions; nor does it entail a modular architecture. If we presuppose the modular architecture argued for earlier, the SLCH still does not indicate the existence of a Statistical Lexical Category Module; lexical category decisions could be made by a statistical parser (e.g. Jurafsky, 1996).

The MLCH addresses the question of modularity, stating that a pre-syntactic module is responsible for lexical category decisions. Initial lexical category decisions should not be affected by syntax and 'higher' levels of processing. The MLCH therefore makes interesting predictions where syntactic constraints and frequency-based lexical category bias are in opposition. For example, in a syntactically unambiguous sentence containing words that display lexical category ambiguity, the MLCH asserts that reanalysis effects will be observed if the initial decision of the lexical category module is syntactically illicit.

Corley's (1998) experiment 2 examined materials of this nature, again concerning the noun – verb ambiguity. Examples are given below.

Experiment 2: Materials

a. The woman said that the German *makes* are cheaper than the rest.
b. The woman said that the German *make* is cheaper than the rest.
c. The foreman knows that the warehouse *prices* are cheaper than the others.
d. The foreman knows that the warehouse *price* is cheaper than the others.

Example (a) is identical to (b) in experiment 1 – the ambiguous word is verb-biased, but the disambiguation favours a noun reading. In contrast, (b) is unambiguous; the plural verb "make" is not syntactically licit following the singular noun "German"; "make" must therefore be a noun. If (all) syntactic constraints affect initial lexical category decisions, we would expect this decision to favour the noun reading despite the verb bias of the lexically ambiguous word.

Examples (c) and (d) both contain noun-biased ambiguous words. In (c) the disambiguating material favours a noun reading. (d) is again unambiguous – the plural verb "price" cannot follow the singular noun "warehouse"; "price" must therefore be a noun.

Experiment 1 determined initial lexical category decisions in the absence of syntactic constraints. The MLCH asserts that these preferences should not be changed by the presence of syntactic constraints. We therefore predict that in (a) and (b), a verb reading will be initially preferred whereas in (c) and (d) a noun reading will be preferred.

As all materials are (eventually) only compatible with the noun reading, we would expect processing difficulty, realised as a reading time increase, to be evidenced downstream from the ambiguous word in the verb-bias conditions. (a) is identical to the materials in experiment 1, and we would therefore predict reading time increases at the disambiguating region. In (b), reading time increases may appear on the ambiguous word itself. This is because there is sufficient evidence for higher levels of processing to demand lexical category reanalysis as soon as the ambiguous word is read. We would therefore predict that reanalysis, reflected by reading time increases, would start on the ambiguous word in the verb-biased unambiguous condition. We do not predict any reading time increases on the noun-biased conditions.

In contrast, any model in which syntax affects initial lexical category decisions, including interactive constraint-based models, must predict no reanalysis effects on the unambiguous conditions. The delay strategy predicts decreased reading times for the ambiguous word and increased reading time for the disambiguating region in the ambiguous conditions compared to the unambiguous ones.

Results and discussion

The method used was the same as that for experiment 1. Average length-adjusted reading times obtained for experiment 2 are shown in Figure 4. On the first word of the disambiguating region, a highly significant main effect of bias was observed ($F_1 = 20.1, p < .001; F_2 = 18.68, p < .001$), but there was no main effect of ambiguity ($F_1 = 0.26, p > .6; F_2 = 0.16, p > .6$). This suggests that initial decisions are based on word bias and ignore syntactic constraints. By the second word of the disambiguating region, recovery in the verb-bias unambiguous condition appears complete. In contrast, recovery in the verb-bias ambiguous condition lags into this word. This suggests that syntax does have a rapid effect on lexical category decisions, but only after the initial decision has been made.

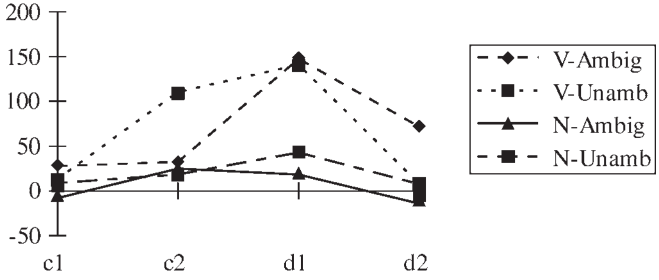


Figure 4. Experiment 2 length-adjusted reading times

A planned comparison of means for the ambiguous word (see Figure 5) reveals a significant difference in reading times for the two verb-biased conditions ($F_1 = 5.24, p < .03; F_2 = 7.16, p < .015$) but not for the noun-biased conditions ($F_1 = 0.12, p > .7; F_2 = 0.10, p > .75$). In the unambiguous verb-bias condition, subjects experience difficulty reading the lexically-ambiguous word. This is predicted by the MLCH; syntactic constraints result in a rapid reanalysis effect but do not affect the initial decision.[7]

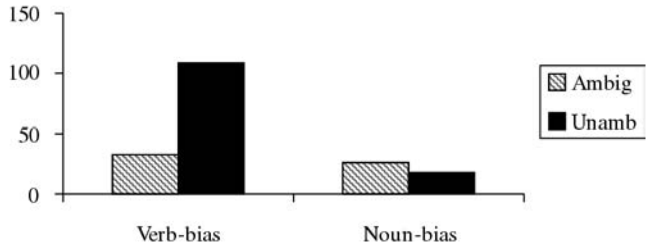


Figure 5. Experiment 2 results for the ambiguous word (c_2)

These results are predicted by and strongly support the MLCH (and the SLCH). They are not compatible with the delay strategy, which predicts a main effect of ambiguity on both the ambiguous word and the disambiguating region. These results are also incompatible with any model in which syntax determines initial lexical category decisions, including some possible interactive constraint-based models. Finally, the observed effect on the ambiguous word is not explained by a model in which reading times are sensitive to syntactic complexity (Just & Carpenter, 1980; MacDonald, 1993). Such a model might (incorrectly) predict an increased reading time on the ambiguous word in the verb-bias ambiguous condition (as a verb phrase must be constructed). However, the observed increase on the verb-bias unambiguous condition compared to the verb-bias ambiguous condition cannot arise directly from syntactic complexity.

Number agreement might have an effect even in a pre-syntactic module if contextual information affects initial decisions (as in the SLCM). This is because the lexical category sequence singular noun followed by plural verb has very low frequency. If we accept that contextual information is used, then this experiment provides evidence that it is in some ways coarse-grained. In particular, the lexical category tags used by the SLCM cannot include number.

Summary of results

Experiment 1 strongly supported the SLCH and the results were not compatible with a model in which frequency-based bias does not affect initial lexical category decisions. Non-statistical models of lexical category disambiguation are therefore disconfirmed.

One such model is the delay model, proposed by Frazier and Rayner (1987). MacDonald's (1993) study demonstrated that Frazier and Rayner's results might have arisen from an artefact in their experiment. Experiment 1 also produced results that are incompatible with the delay model. In contrast, constraint-based models tend to be frequency-based and are therefore compatible with the results reported in experiment 1.

Experiment 2 provided clear evidence that lexical category decisions are made without regard to syntactic constraints – they are therefore pre-syntactic. This result supports the MLCH. The experiment also provided initial evidence that any contextual information used alongside lexical frequency bias (such as the category bigrams of the SLCM) in determining initial lexical category decisions must be coarse-grained.

In supporting the MLCH and SLCH, the experiments reported here also provide direct support for the more general Modular Statistical Hypothesis proposed at the beginning of this paper. In particular, the results of experiment 2 do not appear compatible with interactive models in which syntactic constraints may non-modularly resolve lexical category ambiguities.

Summary and conclusions

We have argued that while statistical mechanisms are commonly taken to be the province of connectionist, constraint-based models of sentence processing, they are also highly consistent with a modular perspective. Rather than being unavoidable side effects of the computational machinery, we argue that statistical mechanisms will be rationally exploited by modular architectures precisely because they provide near optimal heuristic decisions in the absence of full knowledge. Indeed this is a central motivation for the use of statistical language models in computational linguistics. We have dubbed this general proposal the Modular Statistical Hypothesis (MSH).

To investigate the MSH, we proposed a specific theory of human sentence processing, in which lexical category ambiguities are resolved by a post-lexical access/pre-syntactic module. In particular we have argued for the Statistical Lexical Category Module, which adopts the standard tagger algorithm and exploits word-category unigrams and category bigrams to incrementally estimate the probability of the most likely category sequence for a given sentence. We have reviewed the operation of the SLCM, and how it accounts for relevant existing experimental findings.

We then reviewed the results of two new experiments from Corley (1998) designed to directly test both our modular and statistical claims concerning lexical category disambiguation. In both experiments, the predictions of the SLCM were confirmed, thereby supporting both our specific account of category disambiguation and the MSH more generally. The results also have implications for other theories of human sentence processing. While it is true that a constraint-based, interactive framework can be made to account for these findings, it does not *predict* them. That is, such a framework could equally have been made to account for the opposite findings, while such results would have disconfirmed our more predictive (and therefore, we argue, methodologically preferable) modular theory. Regardless, our findings do narrow the space of possible models, suggesting in particular the systematic priority of 'bottom-up' information (e.g. lexical frequency) over 'top-down' (e.g. syntactic and semantic) constraints. Again this follows directly from a modular account, and requires stipulation within a constraint-based framework (though it may follow from particular computational realisations of a constraint-based model).

The findings of experiment 2 also present a challenge for linguistic and psycholinguistic theories which deny the lexical-syntactic divide. These include syntactic theories such as Head-Driven Phrase Structure Grammar, Lexicalised Tree Adjoining Grammar, and Categorial Grammars, to the extent that they claim to be psychologically real (but see Kim and Trueswell (this volume) for a contrary view). Our findings suggest that category decisions are resolved prior to decisions concerning syntactic structures, and also suggest that the categories themselves are relatively coarse-grained, e.g. not including number features.

Finally, we suggest that there is undeniable evidence for the central role of statistical information in human sentence processing. This is a result which needs to be incorporated into the range of existing 'symbolically-based' models. However, such statistical mechanisms should not automatically be taken as evidence against rational and modular theories. On the contrary, statistics may be a module's best friend.

Notes

[1] This paper presents entirely joint work, and the order of authors is arbitrary. Correspondence should be sent in the first instance to M. Crocker (crocker@coli.uni-sb.de). The authors would like to express particular thanks to Charles Clifton, Jr. and Martin Corley for their invaluable assistance. The authors gratefully acknowledge the support of an ESRC Research Fellowship (to Matthew Crocker, #H5242700394) and an ESRC Studentship (to Steffan Corley, #R00429334081), both of which were held at the Centre for Cognitive Science, University of Edinburgh.

[2] See Crocker (1999) for a more complete introduction to the issues presented in this section.

3. Of course their characterisation does define a particular computational position which one might dub 'modular', but the falsification of that position crucially does not falsify the general notion of modularity, only the particular position they define.

4. See Corley and Crocker (2000) or Corley (1998) for a more thorough exposition of HMM taggers and the model being assumed here. See also Charniak (1993, 1997), for more general and more formal discussion.

5. The use of P(w|t) makes the model appear top-down. See Corley (1998, pp. 85–87) for how this (apparently generative) statistical model is actually derived from an equation based on bottom-up recognition. See also Charniak (1997) for discussion.

6. While the model we have presented uses $P(w_i|t_i)$ and $P(t_i|t_{i-1})$, the second measure has no effect in this experiment, where the ambiguous word always follows a noun. This is because P(noun|noun) and P(verb|noun) are approximately equal (as determined from the BNC and Brown corpora). This experiment therefore does not bear on the use of the bigram measure which was independently motivated in Corley and Crocker (2000).

7. Thanks to one of the reviewers for pointing out that, as all temporarily ambiguous sentences are disambiguated towards the noun reading, it might be argued that these results arise from an experimental-internal bias. However, we believe this suggestion is implausible. If the subjects developed a strategy of preferring the noun reading when encountering ambiguous items, we would not expect to observe a significant effect at the start of the disambiguating condition in both verb bias conditions. Furthermore, this strategy would not explain the crucial observation of a reanalysis effect in the verb-bias unambiguous condition on the ambiguous word. Development of such a strategy is also unlikely due to the large number of filler items (80) compared to experimental items (24) presented to each subject.

References

Anderson, J.R. (1991). Is human cognition adaptive? *Behavioural and Brain Sciences, 14*, 471–517.

Brants, T. (1999). *Tagging and Parsing with Cascaded Markov Models*, Unpublished Ph.D. dissertation, Saarland University, Germany.

Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. *Computational Linguistics, 21*, 543–566.

Charniak, E. (1993). *Statistical Language Learning*. MIT Press, Cambridge, MA.

Charniak, E. (1997). Statistical Techniques for Natural Language Parsing. *AI Magazine, 18*(4), 33–44.

Charniak, E., Carroll, G., Adcock, J., Cassandra, A., Gotoh, Y., Katz, J., Littman, M. & McCann, J. (1996). Taggers for parsers. *Artificial Intelligence, 85*(1–2).

Chater, N., Crocker, M.W., & Pickering, M. (1998). The Rational Analysis of Inquiry: The Case for Parsing. In: Chater & Oaksford (eds), *Rational Analysis of Cognition*, pp. 441–468. Oxford University Press, Oxford.

Corley, S. (1998). *A Statistical Model of Human Lexical Category Disambiguation*, Unpublished Ph.D. dissertation, University of Edinburgh.

Corley, S. & Crocker, M.W. (2000). The Modular Statistical Hypothesis: Exploring Lexical Category Ambiguity. In: Crocker, Pickering & Clifton (eds), *Architectures and Mechanisms for Language Processing*, pp. 135–160. CUP, England.

Crocker, M.W. (1999). Mechanisms for Sentence Processing. In: Garrod and Pickering (eds), *Language Processing*, Psychology Press.

DeRose, S.J. (1988). Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics, 14*, 311–339.

Ferreira, F. & Clifton Jr., C. (1986). The Independence of Syntactic Processing. *Journal of Memory and Language, 25*, 348–368.

Fodor, Jerry A. (1983). *The modularity of mind*. MIT Press, Cambridge, MA.

Frazier, L. (1978). *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. Thesis, University of Connecticut.

Frazier, L. & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language, 26*, 505–526.

Juliano, C. & Tanenhaus, M.K. (1993). Contingent frequency effects in syntactic ambiguity resolution. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 593–598. Lawrence Erblaum Associates.

Jurafsky, D.A (1996). Probabilistic Model of Lexical and Syntactic Access and Disambiguation, *Cognitive Science, 20*, 137–194.

Just, M. & Carpenter, P. (1980). A Theory of Reading: From Eye Fixations to Comprehension. *Psychological Review, 87*, 329–354.

Just, M., Carpenter, P. & Woolley, J. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General, 111*, 228–238.

Kawamoto, A.H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity. *Journal of Memory and Language, 32*, 474–516.

MacDonald, M.C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*, 692–715.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*, 157–201.

MacDonald, M.C., Pearlmutter, N.J. & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 10*(4), 676–703.

Mitchell, D.C. & Brysbaert, M. (1998). Challenges to recent theories of cross-linguistic variation in parsing: Evidence from Dutch. In D. Hillert (ed.), *Sentence Processing: A Cross-linguistic Perspective*, pp. 313–335. Academic Press.

Mitchell, D.C., Cuetos, F., Corley, M.M.B. & Brysbaert, 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.

Norris, D. (1990). Connectionism: A Case for Modularity. In D.A. Balota, G.B. Flores d'Arcais & (eds), *Comprehension Processes in Reading*, pp. 331–343. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Pickering, M., Traxler, M. & Crocker, M. (2000). Ambiguity Resolution in Sentence Processing: Evidence against Likelihood. *Journal of Memory and Language, 43*(3), 447–475.

Pritchett, B.L. (1992). *Grammatical Competence and Parsing Performance*. University of Chicago Press, Chicago, IL.

Ratnaparkhi, A. (1996). A Maximum Entropy Model for Part-Of-Speech Tagging. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, Philadelphia, PA, pp. 133–142.

Seidenberg, M., Tanenhaus, M., Leiman, J. & Bienkowski, M. (1982). Automatic Access of the Meanings of Ambiguous Words in Context: Some Limitations on Knowledge-Based Processing. *Cognitive Psychology, 14*, 489–537.

Seidenberg. M. (1997). Language Acquisition and Use: Learning and Applying Probabilistic Constraints. *Science*. 275.

Spivey-Knowlton, M. & Eberhard, K. (1996). The future of modularity. In G.W. Cottrell (ed.), *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*, pp. 39–40. Lawrence Erblaum Associates.

Swinney, D.A. (1979). Lexical Access during Sentence Comprehension: (Re)Construction of Context Effects. *Journal of Verbal Learning and Verbal Behaviour, 18*, 645–659.

Tanenhaus, M.K. & Donnenworth-Nolan, S. (1984). Syntactic Context and Lexical Access. *Quarterly Journal of Experimental Psychology 36*A, 649–661.

Tanenhaus, M.K., Leiman, J.M. & Seidenberg, M.S. (1979). Evidence for Multiple Stages in the Processing of Ambiguous Words in Syntactic Contexts. *Journal of Verbal Learning and Verbal Behaviour, 18*, 427–440.

Tanenhaus, M.K., Spivey-Knowlton, M.J. & Hanna, J.E. (2000). Modelling Discourse Context Effects: A Multiple Constraints Approach. In Crocker, Pickering & Clifton (eds), *Architectures and Mechanisms for Language Processing*, Cambridge University Press, Cambridge, England.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M. and Sedivy, J.C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science, 268*, 1632–1634.

Viterbi, A. (1967). Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory, 13*, 260–269.

Weischedel, R., Meteer, M., Schwarz, R., Ramshaw, L. & Palmucci, J. (1993). Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics, 19*, 359–382.

Encoding and storage in working memory during sentence comprehension

Laurie A. Stowe, Rienk G. Withaar, Albertus A. Wijers, Cees A.J. Broere and Anne M.J. Paans
University of Groningen / University Hospital Groningen

In this article we will discuss evidence from a number of recent neuroimaging experiments. These experiments suggest that three areas play a role in sentence comprehension: the left inferior frontal gyrus (LIFG), the left posterior superior temporal gyrus (STG), and the anterior temporal lobe (ATL). The left posterior STG appears to be important for sentential *processing*, since activation in this area increases as a function of the structural complexity of the sentences which must be comprehended. The LIFG, on the other hand, is activated by storage of lexical information as well as by sentential complexity. It is possible to explain a range of experimental results by hypothesizing that this area is responsible for *storage* of both lexical and phrasal information during comprehension. The ATL does not respond to structural complexity during sentence comprehension, but it is consistently more activated during comprehension of sentences than of word lists. On the basis of evidence which shows that the ATL is important for encoding in short-term verbal memory tasks, we suggest that it is responsible for *encoding* of information about words for use later in comprehension.

Accounts of garden paths and complexity in sentence comprehension frequently appeal to the limits of working memory (Gibson, 1998; Just & Carpenter, 1992). However, to provide substance to such explanations, a fully specified working memory model is necessary. That is beyond the scope of this paper, but we will consider neuroimaging evidence suggesting that a fully specified model must be concerned with both *what* is stored and how it is *encoded* into working memory.

The evidence that we will present comes from positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) studies. These techniques can image the blood flow in various anatomical structures in the brain. If blood flow in a particular brain area increases with processing demand, we can infer that the process in question depends on the cognitive function carried out in that area. The questions that we will consider are the following:

1. Are separate areas involved in storage and in processing?
2. If so, what factors are important in determining memory load on a particular working memory area?
3. If storage is separate from processing, can we identify the mechanisms which are involved in encoding relevant information into working memory?

Dissociating sentential working memory and sentential processing

First, we will address the extent to which processing and storage can be separated. One way to demonstrate that storage during processing and the processes themselves are dissociable is to show that they are supported by different brain areas. First, we will consider evidence suggesting that sentential working memory can be localized and articulate some hypotheses based on that evidence. In the next section we will discuss an experiment which tested these hypotheses and its results, including evidence that suggest a dissociation between storage and processing.

Frontal lobes, sentential complexity and verbal memory

It has been generally accepted by neurologists and neurolinguists that an area in the frontal operculum, including parts of the left inferior frontal gyrus (LIFG) is involved in language processing. There has been some disagreement about what function it serves, however. Earlier it was thought primarily to support production, while more recent theorists argue for a role in comprehension as well. Recent neuroimaging evidence has confirmed that the frontal operculum and underlying insula are important in language comprehension. Mazoyer, Tzourio, Frak, Syrota, Murayama, Levrier, Salamon, Dehaene, Cohen, & Mehler (1993) found that this area was more activated during sentence comprehension than during a neutral resting condition. Several studies have shown increasing activation in this area as sentential complexity increases. These experiments are summarized in Table 1. The left posterior middle temporal gyrus has also been reported to become activated as sentential complexity increases

(Just, Carpenter, Keller, Eddy, & Thulborn, 1996; Stowe, Wijers, Willemsen, Reuland, Paans, & Vaalburg, 1997). We will return to this point below.

Table 1. Overview of studies showing LIFG activation for sentence comprehension. The location of the maximal voxel, if available, is given in Talairach & Tournoux (1988) stereotactic coordinate system (in millimeters from the origin)

Study	Comparison	Maxima
Just et al. (1996)	three levels of complexity	not available
Mazoyer et al. (1993)	sentence vs. rest	not available
Stowe et al. (1994)	three levels of complexity	not available
Experiment 1, this article	three levels of complexity	–38, 14, 12
Stowe et al. (1995)	two levels of complexity	–22, 26, 12
Stromswold et al. (1996)	two levels of complexity:	–47, 10, 4
	comparisons 1 & 2	–38, 20, 8
Mean maxima location		–36, 18, 9

The activation of the IFG and underlying insula during sentence processing might suggest that sentence processing or, more specifically, syntactic processing is carried out in this area. However, the frontal lobe also maintains information for delayed processing (Petrides, 1996). The involvement of the LIFG in verbal memory has been shown in blood flow change experiments on memorizing lists, maintaining lists, and recognizing items out of a recently studied list. These tasks all involve verbal working memory. The area typically activated by these tasks is the same frontal opercular area that is active in sentence processing, cf. Table 2.

For purposes of comparison, we have calculated the mean stereotactic location of the maxima of the activations in these two sets of experiments. Although the lack of information for several of the sentence studies limits the accuracy of this calculation, it can be seen that the means are quite close to each other; the variability around the mean location is also similar in both sets of studies. We assume that it is not coincidental that these two tasks activate the same area. It appears that a cognitive function carried out in this area supports both tasks; it seems likely that this function is some form of working memory.

Hypotheses about the function of the frontal lobe

We will articulate several hypotheses about the cognitive function carried out in this area and how it relates to activation during maintenance of lexical and sentential information. The predictions generated by these hypotheses were

Table 2. Overview of studies showing LIFG activation for verbal memory tasks

Study	Comparison	Maxima
Fiez et al. (1996)	Maintenance – rest	–59, 17, 12
Grasby et al. (1994)	Correlation with list length	–22, 22, 12
Paulesu et al. (1993)	Verbal list – Korean alphabet list	–34, 2, 4
	(verbal – visual memory)	–46, 2, 16
Petrides et al. (1993)	Self-ordered – externally ordered	–43, 12, 9
Mean maxima location		–41, 11, 11

tested in Experiment 1. To test the hypotheses, we measured changes in regional blood flow across various processing conditions. Blood flow reflects a summing up of the workload over the entire scan.

Single function hypotheses

We argued that it is most parsimonious to assume that the LIFG supports a single cognitive function which is called upon by both short term memory for words and sentence comprehension. We have considered three single function hypotheses which differ in terms of information to be maintained and time of maintenance.

Unstructured Lexical Verbatim Memory. As we have seen, word memory tasks activate the LIFG. Accordingly, we will first consider the hypothesis that in both verbal memory and sentence comprehension tasks, only words are maintained in the IFG. This memory representation is used during short term memory tasks but can also be accessed to support sentence processing, although the area does not actually process or represent sentential structure. However, if maintenance is indeed in a simple, unstructured passive store, it is not clear what explanation can be offered for the difference between simple and complex sentences. Therefore, this hypothesis can be rejected. As long as words are presented at the same rate, memory load should be the same, while in fact it is clear that more complex sentences cause greater activation (cf. Table 1).

> Prediction 1: Unstructured Lexical Verbatim Memory
> Word List = Simple Sentence = Complex Sentence = Most Complex Sentence

Structured Lexical Verbatim Memory. An alternative is that only lexical information is maintained in this area, but words are only stored while they are needed. We assume that in sentences, words must be maintained at least until they can be converted into a phrasal representation (cf. Marcus, 1980); it is unnecessary to maintain them after this has occurred. Under this hypothesis, syntactic complexity affects activation in this area because it determines how many words have to be maintained and how long they have to be maintained. An extra assumption that we make here is that the process which makes use of the stored words signals when maintenance is no longer necessary.

In simple sentences, phrases are completed quickly and individual words can be dismissed, so memory load will be low for the lexical memory store. In complex sentences, phrases remain incomplete for a longer time and words therefore must be maintained longer, with greater cost to the LIFG. Word lists have the heaviest load, as there is no parsing or comprehension process to signal that they can be dismissed. Thus the store will fill to its maximum capacity.

> Prediction 2: Structured Lexical Verbatim Memory
> Simple Sentence < Complex Sentence < Most Complex Sentence < Word List

At this point, it is time to consider one central question of this section: to what extent is there an area which is involved in processing sentence structure and an area which provides a working memory which supports processing? The lexical memory proposed above supports sentence processing, but it clearly does not actually carry out the processing. This implies that there is a separate area in which phrases are constructed (and possibly also stored). In such an area, word lists should be associated with relatively little cost (activation), as they cannot be assigned a syntactic structure, while simple sentences lead to a greater amount of processing, and more complex sentences to even more processing.

> Prediction 3: Sentence Structural Processing Load
> Word List < Simple Sentence < Complex Sentence < Most Complex Sentence

As noted, this pattern of activation does not necessarily suggest a processor. A working memory for phrases is also possible. The functions of other areas supporting sentence comprehension will determine which interpretation is more likely to be correct.

Complex Working Memory. To this point, we have assumed that the function of the IFG is to maintain words. However, most discussions of working memory in sentence processing assume that the sentential working memory load is determined by the complexity of the syntactic phrases which are being constructed and/or recall of earlier information to complete phrases. From this viewpoint, the LIFG may support sentence processing by maintaining phrases. However, if we assume that this area supports phrasal memory and nothing else, it is not obvious why it also appears to store lexical items (cf. Table 2). Therefore, we propose that the IFG supports a *Complex Working Memory* in which words are maintained until they can be converted into phrases and in which phrases are then stored until the larger phrase or sentence in which they are contained is completed. Marcus (1980) used such a memory buffer to explain certain constraints on syntactic structure.

Under the *Complex Working Memory* hypothesis, memory load in the LIFG is a combination of the number of words and phrases to be remembered and of how long they must be remembered (with activation summed over whole sentences or set of sentences). We have already discussed lexical memory load. Phrasal memory load is based on the number of phrases maintained within phrases: thus the more complex the sentence, the more load. Additionally, syntactically ambiguous sentences in which two potential sets of phrases must be maintained will have a greater cost. Memory load can thus be approximated by simply combining the weights of the *Structured Lexical Verbatim Memory Load* and the *Sentence Structural Processing Load* just described.

Combining the weights affects word lists most. Word lists are low on phrasal load, but high on lexical load. Assuming that both load factors have approximately the same weights, word lists (high lexical load, low phrasal load) will be associated with a heavier memory load than the simplest sentences which have a low phrasal load and also have a low lexical load, as words are combined immediately into phrases. However, word lists entail a smaller load than extremely complex sentences, in with a high phrasal and lexical load as words have to be held in memory for several words before phrases can be formed.[1]

> PREDICTION 4: COMPLEX WORKING MEMORY
> Simple Sentence < Word List = Complex Sentence < Most Complex Sentence

The goal of the first experiment reported below was to investigate the hypotheses which we have discussed.

Experiment 1: Sources of Verbal Working Memory Load

Subjects: Twelve right-handed students (7 F, 5 M; mean age = 21.08 years) who were native speakers of Dutch served as paid subjects. All subjects had normal or corrected to normal vision and no history of neurological problems. All subjects had given informed consent under a procedure approved by the University Hospital Medical Ethics Committee before participating.

Materials: To test predictions 2–4, we must contrast lists of sentences with several levels of syntactic complexity against a word list. Simple one clause sentences (containing no clausal embeddings or center-embedded structures, with phrases occurring in their normal order) were compared with two sorts of more complex sentences.

The complex sentence condition contained embeddings and list-like constructions (four sentences containing center embedded clauses or adjectival verb phrases, one center-embedded gapping construction, two right-branching embedded clauses plus non-canonical order (passive and object relative), and one multiple adjective noun phrase). We chose these constructions to cover a gamut of syntactic complexity, since we argue that in all these complex structures, the same working memory supports sentence processing.

The most complex sentence condition contained a category ambiguity followed by a phrase which fit with either category; the sentence remained ambiguous for at least four words. Then the structure was resolved to the non-preferred structure. An example is *Zij kunnen* **bakken** *met zulk deeg niet verplaatsen. Bakken* is ambiguous between noun and verb, both of which are grammatically possible. The succeeding prepositional phrase can modify *bakken* in either of these meanings. Categorial ambiguities were selected, as there is reason to suspect that both structures are processed simultaneously with such ambiguities (Frazier & Rayner, 1987). A pretest had shown which interpretation subjects typically preferred (here the verb). The ambiguity was disambiguated to the less-preferred structure (e.g. by *niet verplaatsen*). In terms of phrasal load, these structures are therefore quite complex, whether subjects garden-path and reanalyze or parse both structures. As blood flow reflects the summed effort across the scan, it is not particularly important which of these is the case. The structures to which the sentences were disambiguated are comparable to those used in the simple sentence list.

Word lists contained both content and function words, but were ordered so that no two contiguous words formed a phrase, which prevented successful syntactic processing. This was necessary to ensure that no phrasal memory load built up. Semantic relationships were avoided to prevent strategic semantic processing. The mix of word classes was comparable to that of the complex sentences and was obtained by creating similar sentences and then scrambling the order of the words across the entire list.

Four lists were created, each of which contained only one of these conditions. Each was presented during a separate scan. The four lists were matched on number of words, mean word length, and mean logarithmic word frequency. Sentence lists were additionally matched on rated plausibility.

Procedure: In order to minimize movement during the experiment, head moulds were made for each subject. Subjects were placed lying with their heads in a Siemens CTI (Knoxville, Tennessee, USA) 951/31 positron emission tomography camera parallel to and centered 3 cm above the glabella-inion line. A computer screen was suspended in front of the camera above the subject's body with the center of the screen approximately 90 cm from the subjects eyes. Before the actual measurement started, a transmission scan was made in order to correct for attenuation. During the transmission scan, subjects were presented with a practice list so that they would become familiar with the procedure. Prior to each scan, 1.85 GBq $H_2^{15}O$ was injected as a bolus in saline into the right brachial vein via a venous canula. Simultaneously with the injection, presentation of the sentences began. Each word was presented in the center of the computer screen for 650 msec. This speed was chosen on the basis of a pretest to ensure the comprehension of the ambiguous sentences, the most difficult condition. An asterisk appeared between sentences and words strings of a similar length to allow subjects to blink. The collection of the data started 23 seconds after list presentation began, in order to give the tracer time to reach the brain, and continued for 90 seconds. Between injections, 15 minutes were allowed for decrease of activity to background level. The order of presentation of the different lists was counterbalanced across subjects.

Data analysis: For each scan, regional cerebral blood flow (rCBF) was estimated. The measurement of the tracer is affected by the subjects' skulls and head support. Transmission scans were used in order to correct for attenuation. Then, the data were resampled using a voxel size of 2.2 × 2.2 × 2.4 mm. In order to correct for subject movement between scans, a least mean squares algorithm (Woods, Cherry, & Mazziotta, 1992) was employed to align each subject's scans.

For further data analysis, we used the Statistical Parametric Mapping program (SPM95, developed by the Wellcome Institute of Cognitive Neurology, London, UK). In order to be able to compare data from different subjects, scans were translated into the brain atlas coordinate system of Talairach & Tournoux (1988) using linear and non-linear stereotactic normalization procedures (Friston, Ashburner, Frith, Poline, Heather & Frackowiak, 1995a) Then a Gaussian filter with a width of 20 mm in the x (side to side) and y (front to back) dimension and a width of 12 mm in the z (top to bottom) dimension was applied to each image. This was done since the translation into the brain atlas coordinate system does not produce perfect alignment; detection of nearly overlapping activations is maximized by the spatial smearing introduced by the filter.

A comparison of changes in rCBF between conditions was made on a voxel by voxel basis in a series of planned comparisons. An increase of rCBF is taken to reflect increased regional metabolic activity, and hence increased functional brain activity in the region, for the condition in which it occurs. Comparisons produced a Z-statistic for each voxel (Friston, Holmes, Worsley, Poline, Frith & Frackowiak, 1995b). The probability of the Z-score was then corrected for multiple comparisons (Friston, Worsley, Frackowiak, Mazziotta & Evans, 1994) in a procedure similar to the Bonferroni correction. However, this correction is fairly strict and false positives at corrected $P< 0.2$ appear to be uncommon (EU Concerted Action on Functional Imaging, 1996); weak activations will therefore be discussed here. Additionally a statistic was calculated for the spatial extent of contiguous voxels which were activated above the threshold $Z= 2.8$, since a cluster of false positives is less likely than single false positives.

Three comparisons were made, which tested predictions 2 through 4 respectively. The important aspect of each of these predictions is that load increases in a predictable way across the four conditions. It is difficult to estimate the exact degree of increase that should be expected across the four conditions, as the difference in memory load is not necessarily constant between conditions. Nevertheless, a simple linear regression (with equidistant weights) based on the relative loads predicted by each hypothesis provides a first-order approximation of the increase across the conditions under a given hypothesis.

Evidence for a Frontal Complex Working Memory

First we will discuss the two analyses which tested predictions concerning the function of the left inferior frontal lobe.

Lexical Verbatim Memory. The first correlation analysis was based on Prediction 2, derived from the *Structured Lexical Verbatim Memory* hypothesis. According to this hypothesis, Simple Sentence (= –3) < Complex Sentence (= –1)

< Ambiguous/Most Complex Sentence (= 1) < Word List (= 3).[2] There were no significant activations showing this pattern. This suggests that there is no area of the brain, including the LIFG, which supports a structured lexical memory.

Complex Working Memory. The second correlational analysis tested Prediction 4, which was based on the *Complex Working Memory* hypothesis. According to this hypothesis, Simple Sentence (= –2) < Complex Sentence (= 0.01) = Word List (= 0.01) < Ambiguous/Most Complex Sentence (= 1.98).[3] This analysis identified an area of significant activation in the LIFG (cf. Figure 1). This activation includes parts of Brodmann's Areas 44 and 45, as well as underlying insular cortex. The maximally activated voxel was located at –38, 14, 12 and had a Z value of 4.87, with a corrected P of 0.004. Note that this is very similar to the mean center of activation in Tables 1 and 2. As predicted on the basis of the previous literature and on the hypothesis that the left inferior frontal lobe provides a working memory which maintains both lexical and phrasal information in memory during sentence processing, the LIFG shows increasing memory load for the combined weights of lexical and phrasal memory load.

A separate sentence processing area: Sentence Structural Processing Load

The third correlation analysis examined Prediction 3 which states that activation should increase as Sentence Structural Processing Load increases in an area which supports sentence processing. Under this hypothesis, Word List (= –3) < Simple Sentence (= –1) < Complex Sentence (= 1) < Ambiguous/Most Complex Sentence (= 3). One area showed a significant activation with this pattern (cf. Figure 2). The activated area in this figure is in the posterior left middle temporal gyrus with an extension into the posterior superior temporal gyrus (Brodmann's Areas 21 and 22). The voxel in the activation showing the maximal activation was located at –34, –58, 4; the Z value was 4.32, which is associated with a corrected P of 0.047. This activation is similar in location to several activations in response to increasing sentential complexity reported in the literature (Just et al., 1996; Stowe et al., 1997). We noted above that this pattern of activation is also predicted for an area that supports phrasal working memory. However, since the LIFG appears to support phrasal memory as well as lexical memory, the attribution of processing to this area is a more plausible explanation of the activation.

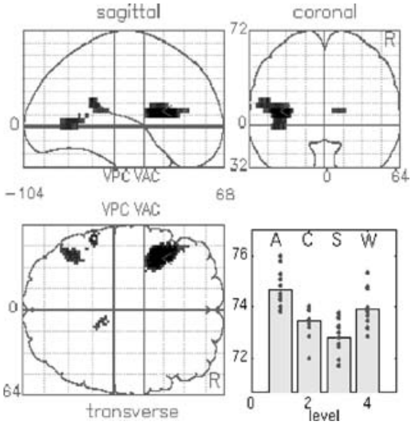


Figure 1. Complex Working Memory comparison. A = Ambiguous; C = Complex; S = Simple; W = Word List.

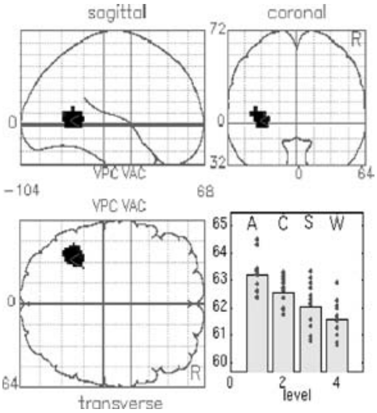


Figure 2. Sentence Structural Processing Load comparison. A = Ambiguous; C = Complex; S = Simple; W = Word List.

Note: In all figures, voxels exceeding Z = 3.0 are shown (darker voxels have higher Z-scores) as if looking through a transparent brain. In the upper left-hand corner, the brain is seen from the right side (sagittal), showing area(s) of activation in the z (top to bottom) and y (front to back) dimension; in the upper right-hand corner, the view is from the back (coronal) showing the z and x (left to right) dimensions; in the lower left-hand corner, the view is from above (transverse), showing the x and y dimensions. Smaller non-significant activations are also included to show the extent of noise in the comparison. The lower right-hand corner plots the mean rCBF and variance at the maximally significant voxel in ml. of blood per minute per dl. brain volume. Condition labels are explained below each figure.

Conclusion from Experiment 1

The results of Experiment 1 suggest that it is possible to dissociate processing from working memory anatomically. The left posterior temporal lobe shows an increasing activation as sentential complexity increases. The LIFG's activation is best characterized by a combination of lexical memory load and phrasal memory load. Word lists, with a high lexical, but low phrasal load, show more activation than simple sentences but less than ambiguous sentences, where two structures have to be stored for processing. We will return to the consequences of this result for models of sentence processing in the conclusion.

A double function hypothesis

The hypotheses in the preceding section were based on the assumption that the LIFG supports a single cognitive function subserving working memory tasks and sentence processing. This was based on the argument that a single function is most parsimonious. However, this assumption is not necessarily correct. An alternative is that the common location of the activations for verbal memory tasks and sentence complexity manipulations is misleading. Rather than one neural network, two independent, though similar, networks in the LIFG may support memory for word lists and for sentence processing respectively. The existence of multiple memory stores is consistent with some theories of working memory in sentence processing (Caplan & Waters, 1999; Martin, 1993), although the exact sorts of memory stores that are proposed may differ.

If there are two networks located in virtually the same area, the activation in the area will be determined jointly by the activity in the two networks, that is, activation will depend on a combination of the lexical and phrasal memory loads (cf. note 1). On the basis of the results of Experiment 1, we can conclude that a lexical memory network in this area cannot be an unstructured memory, however. If it were, word lists are predicted to have the least load, since the contribution of lexical memory load is constant over all conditions and only phrasal memory load changes. This pattern was tested as Structural Processing Load, and it did not produce a significant activation in the LIFG. If the lexical memory network is a structured memory, the predictions are identical to those of the *Complex Working Memory* hypothesis, given the same assumptions discussed in footnote 1. That pattern was found in Experiment 1 for the LIFG and can thus be explained equally well by either hypothesis.

However, the *Complex Working Memory* hypothesis makes an additional prediction: lexical memory and phrasal memory should share the same resource within the LIFG. That means that an increase in load for phrasal memory has effects on lexical memory, as less memory is available. Thus if these factors are orthogonally manipulated, they ought to interact. In sentences, these factors normally co-vary, since as syntax becomes more complex, lexical load and phrasal load both tend to rise. However, non-sentential lexical memory is also supported by the LIFG (cf. Table 2), so an external memory load manipulation can be used to provide the orthogonal comparison.

> Prediction 5: Complex Working Memory
> Lexical Load interacts with Phrasal Load

The *Separate Functions* hypothesis differs from the *Complex Working Memory* hypothesis in this respect. Since the networks are separate according to this hypothesis, lexical memory load and phrasal memory load are not competing for the same resources and thus varying phrasal load should not affect the resources available for lexical memory. This suggests that lexical memory load and syntactic load should not interact, barring ceiling effects.[4] Experiment 2 tested whether syntactic complexity interacts with extrinsic memory load in the frontal operculum or not.

> Prediction 6: Separate Lexical and Phrasal Working Memory
> Lexical Load does not interact with Phrasal Load

Experiment 2: Lexical Working Memory Load vs. Phrasal Memory Load

Subjects: Twelve right-handed informed volunteers (10 F, 2 M; mean age = 21.6) participated in the study. The same criteria were used as in Experiment 1.

Materials: Syntactic complexity and external memory load were orthogonally manipulated. The complex condition contained non-final embedded structures: center-embedded adverbial clauses, subject clauses, and center-embedded relative clauses modifying subject NPs or topicalized object NPs at the beginning of the main clause, all of which are more complex than single clause sentences and thus use more phrasal memory storage. An example of the relative clause on a topicalized object is *De lamp die op de grond was gevallen, repareerde de monteur* (lit. The lamp that on the ground was fallen repaired the mechanic). The sentences were assigned to lists matched as to sentence structure. Matching lists were then created containing simple sentences made by dividing the complex sentences into two main clauses, e.g. *De lamp was op de grond gevallen. De monteur repareerde de auto* (lit: The lamp has on the ground fallen. The mechanic repaired the auto.) To do so, it was necessary to add words in some clauses, but the additional length was balanced by the deletion of complementizers, relative pronouns, and subordinating conjunctions. The complex lists and simple lists were thus matched in number of words, mean logarithmic word frequency and word length. The simple lists were also matched in sentence structure; all lists were matched in rated plausibility.[5]

To manipulate lexical memory load, memory sets were selected from each list. The high load set consisted of five words, the low load of one. The memory sets were comparable in mean logarithmic frequency, mean length in letters and in syllables, morphological complexity and word category. The combinations of list and load were allocated to the subjects so that subjects saw sentences and memory list only once, but across the experiment, lexical factors were matched across the four conditions. The conditions were presented in various orders so that the conditions appeared equally frequently as first, second, third, and last scan. There is thus no reason to expect confounds due to the order of conditions, lexical content of the conditions, or lexical content of the memory sets.

To force the subjects to maintain the words in working memory, they were requested to respond as soon as they saw a word from the memory set in the sentences that they were reading. The words in the five word memory sets were evenly spread over the lists. Memory targets appeared in approximately the same position in each list. The final memory target in each five-word list was the low memory load target.

Procedure: The procedure was similar to that in Experiment 1. The words in the memory set were displayed on the monitor suspended in front of the subjects for 30 seconds before injection of the tracer. Sentences started three seconds after injection and were presented as in Experiment 1, but each word appeared for 500 msec, since this speed ensured comprehension in a pretest.

Data analysis: The data analysis was the same as in Experiment 1.

Interactions between Working Memory Components

Virtually no mistakes were made in the recognition task. In the rCBF data we tested for main effects of sentence complexity, of external memory load, as well as for interactions of these two factors. There was no significant main effect of sentence complexity in this experiment, as opposed to the experiments summarized in Table 1. External lexical memory load caused a significant activation of the left extrastriate occipital lobe centering in Brodmann's Area 18; the maximally activated voxel within this region had a corrected P of 0.011, with a Z value of 4.56, located at –16, –76, 20 (cf. Figure 3). This activation is similar to activations found in a number of experiments in which visual working memory has been manipulated, except that it is lateralized to the left (cf. Fiez, Raife, Balota, Schwarz, & Raichle, 1996, for an overview of these results).

We performed two tests for interactions between sentential complexity and external memory load. The first was for areas in which external memory load had a more positive effect for complex sentences than for simple sentences. This showed a significant activation in the LIFG in Brodmann's Area 44 and

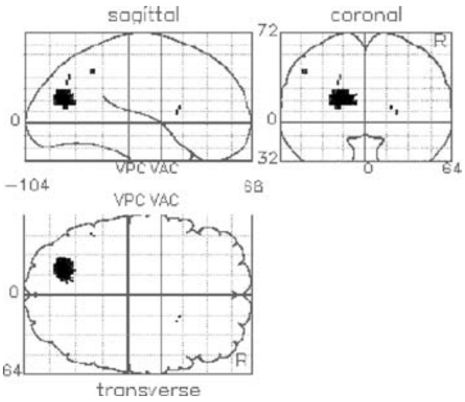


Figure 3. Voxels showing activation at Z > 3 for the Main Effect of External Memory Load.

45 but the activation also includes large portions of the left frontal dorsolateral cortex (cf. Figure 4) and a significant area was also activated in the right frontal lobe. The largest maxima was, however, located in the anterior insula underlying the frontal operculum at –28, 8, 4; it had a Z value of 4.9 and a corrected P of 0.003. The spatial extent of the activation was also significant (P = .006) at the threshold Z = 3.0; the region included 473 voxels above this threshold. The form of the interaction is shown in the lower right hand corner of Figure 4.

The opposite interaction, in which the effect of memory load is more positive for simple sentences than for complex sentences, showed a weak interaction in an area of left inferior posterior parietal lobe which was spatially contiguous to the area which showed a main effect of external memory load (cf. Figure 5). This interaction was not predicted by any of the hypotheses discussed above. The inferior posterior parietal activation was maximal at –8, –78, 32 with a Z value of 3.9 and a corrected probability (P) of 0.123. Although this effect is weak, the left inferior posterior parietal lobe and its right hemisphere homologue have been found in a number of studies of visual working memory (cf. Fiez et al., 1996, for an overview). Thus it is likely that the effect reflects real differences in cognitive processing over these conditions. The form of the interaction is shown in the lower right-hand corner of Figure 5.

Evidence for a Single Working Memory Function in the LIFG

Experiment 2 investigated whether lexical memory and phrasal memory make use of the same working memory or different working memories by testing for an interaction of these factors. The presence of an interaction suggests very

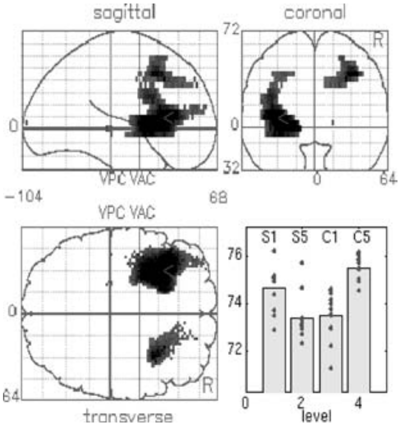


Figure 4. Interaction of Lexical Load with Sentence Complexity. S=Simple; C=Complex; 1=Low Load; 5=High Load.

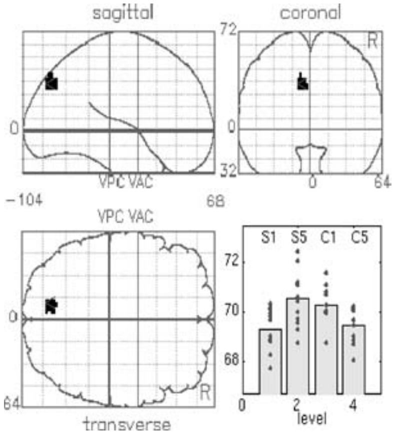


Figure 5. Interaction of Lexical Load with Sentence Complexity. S=Simple; C=Complex; 1=Low Load; 5=High Load.

strongly that both tasks make use of the same resource. The form of the interaction is unexpected however. As can be seen in the plot in Figure 4, the simple sentences with low memory load actually show nearly as much activation as the complex sentences with high memory load. The beginnings of an explanation can be seen in the other two effects found in this experiment. The main effect of load is seen as an activation in the left extrastriate occipital lobe. This area has frequently been activated by visual memory tasks, which suggests that subjects tended to treat the external memory task primarily as a visual rather than as a verbal memory task (Baddeley, 1986). The weak second interaction supports this suggestion. The result consists of an extension of the main visual memory activation, suggesting that visual memory processing was extended in some of the conditions. Those conditions are precisely the ones in which the frontal lobe activation was weakest. It seems that there was a trade-off between the visual and verbal memory components during the word-monitoring task. The lexical load in the left inferior parietal lobe is apparently dependent on the extent to which the task is carried out in the frontal lobe, which in turn depends on the available resources. In the low load/simple sentence condition, the task can be partly carried out in the verbal memory system. The resources may be limited enough to suggest using more visual resources in the high load/simple sentence condition and complex sentence conditions. This leaves the question of the high load/complex sentence condition, in which verbal memory resources are clearly limited. Nevertheless, the task is apparently partly carried out in the verbal system. It seems possible that while processing complex sentences, it is more difficult to suppress the use of the verbal memory system for the secondary task. This interaction, however, clearly needs more research.

This interpretation of the data does not weaken the original conclusion however. It rather suggests that the left frontal gyrus provides a working memory resource which can store lexical information in verbal memory tasks and in sentence processing. When two concurrent tasks make use of this working memory, the resources may not be adequate, so that one of the tasks may need to be handled in another system, if available. Here, visual memory resources are available for the word-monitoring task. This implies strategic control over this memory resource, at least for tasks involving explicit word maintenance.

Encoding verbal information during sentence comprehension

Up to this point we have focused on processing vs. storage during sentence comprehension. These functions are apparently associated with the LIFG and PTL. However, several experiments have shown that a third area, the anterior temporal lobe (ATL), is activated when sentences are compared with word lists (cf. Figure 6, from Stowe, Broere, Paans, Wijers, Mulder, Vaalburg, & Zwarts' study (1999). Several experiments with this result are summarized in Table 3.

Table 3. Studies showing ATL activations for sentence vs. word comparison

Study	Comparison	Maxima
Bottini et al. (1994)	Sentence – Word list	–48 –10 –8
Mazoyer et al. (1993)	Sentence – Rest vs.	ATL activation
	Word List – Rest	No activation
Stowe et al. (1999)	Sentence – Word List	–50 –2 –16
Tzourio et al. (1998)	Sentence – Rest	ATL

The most interesting point about this activation, in the current context, is that this area has not been reported in any of the experiments manipulating sentential complexity. Conversely, none of the experiments comparing sentence with word lists report a significant activation of the LIFG. This suggests that the area is involved in neither processing nor storage. We will first confirm this dissociation using data from Experiment 1, and then consider evidence suggesting that the ATL is involved in encoding information into storage.

Complexity and the Anterior Temporal Lobe

In Experiment 1, it is possible to test for areas in which all three sentence conditions showed equivalent activation and more activation than the word list condition. All sentence conditions activated lateral ATL bilaterally (cf. Figure 7).

The maximally activated voxel within the left hemisphere was located at –40 –2 –20; the activation was significant in extent (256 voxels; P = 0.014. The homologous area in the right hemisphere was also significant in extent, including 193 voxels, giving a probability of 0.038, with the maximum voxel at 42, 8, –16. The lower right-hand corner of Figure 7 shows the relative activation of the four conditions at the maximally activated voxel in the left ATL.

It is striking how comparable the activations for simple sentences, complex sentences and syntactically ambiguous sentences are, confirming that there is no effect of syntactic complexity in this area. The dissociation between areas activated by complexity and by reading sentences vs. word lists is very interesting. Considering only the results in Table 3, it would have been plausible to interpret the activation in the ATL as supporting syntactic or semantic processing. However, if this area were engaged in the construction of a morphosyntactic or sentential semantic representation, we would expect clear effects of structural complexity. This does not emerge. On the other hand, the cognitive process which is subserved by this area is clearly called upon more during sentence comprehension than during the processing of word lists.

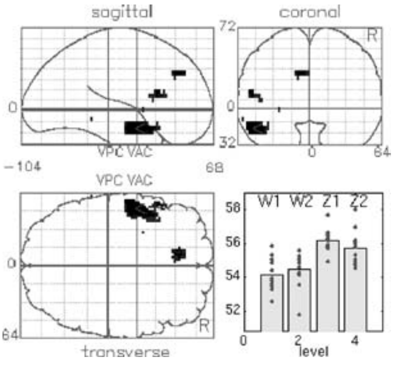


Figure 6. Sentence vs. Word List in Experiment 3. W1 = Word List 1; W2 = Word List 2; Z1 = Sentence List 1; Z2 = Sentence List 2.

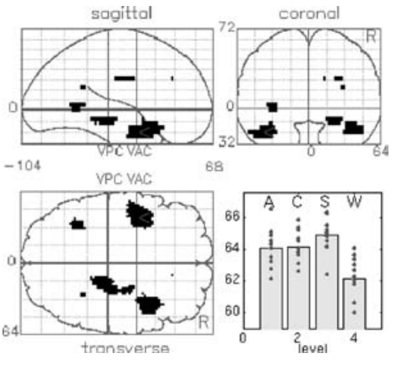


Figure 7. Sentence vs. Word List in Experiment 1; A = Ambiguous; C = Complex; S = Simple; W = Word List.

Functions of the Anterior Temporal Lobe

There is evidence in the literature showing that the ATL is important for encoding into memory. Combined with evidence that anterior temporal damage leads to some decrement in sentence comprehension, this suggests that the ATL plays a role primarily in encoding verbal information into storage for later use.

Encoding for Later Retrieval from Memory. The lateral ATL is involved in encoding of verbal information for later retrieval under some circumstances. Fedio & Van Buren (1975) found that stimulating the lateral surface of the ATL while patients were presented with pictures did not interfere with naming the pictures, but later recall was impaired. Stimulation thus did not interfere with identification or word production, but did with encoding into working memory. Stimulation during maintenance caused less problems than stimulation during the encoding phase (Ojemann & Dodrill, 1985).

Another set of data provides more problematic support for the same claim. Anterior temporal lobectomies performed on patients suffering from complex partial seizures affect some of the systems involved in recall. When normal subjects study words, delayed repetition of the word affects a negative event-related potential wave form, the N400, which peaks approximately 400 msec after the presentation of a word. Before lobectomy, epileptics show this pattern too, but after lobectomy in either hemisphere, this effect is decreased or absent (Rugg, Roberts, Potter, Pickles, & Nagy, 1991). On the other hand, their recognition of repeated items is not impaired, so the actual memory is not affected. The results just discussed are found when subjects intentionally study word lists. Without intentional study, the N400 shows a repetition effect, but it disappears quite quickly. Schnyer, Allen, & Forster (1997) show that the N400 repetition effect disappeared after several words under masked priming. With no masking, the repetition effect is diminished within six words (Karayanidis, Andrews, Ward, & McConaghy, 1991). The N400 repetition effect is not generated in the anterior temporal lobe, as words which are repeated at a short lag show N400 priming effects even after anterior temporal lobectomy (Rugg et al., 1991). It appears that the N400 is affected if the word is, in some sense, activated; intentional encoding into memory sets up circumstances under which the word remains activated for a longer period. This encoding is missing for the anterior temporal lobectomy patients.

Apparently, a memory encoding process which affects the availability of the word is carried out in the anterior (although not necessarily lateral) temporal lobe. This is the problematic aspect of the data. In anterior temporal lobectomies, normally medial structures are removed, such as the hippocampus and the amygdala which are known to play a role in memory. The effects just described may be due to either lateral or medial structures. The activation in Experiment 1 and that reported by Stowe et al. (1999), on the other hand, were lateral and did not include the medial structures. Since the effects of electrical stimulation also primarily affect lateral cortex, the conclusion that the lateral cortex plays a role in memory encoding nevertheless seems reasonable.

Anterior Temporal Lobe effects in sentence comprehension. It has not generally been noted that lesions in the ATL lead to sentence processing deficits. However, some problems may occur. Zaidel, Zaidel, Oxbury, & Oxbury (1995) tested left and right temporal lobectomy patients on syntactically and semantically ambiguous sentences. After anterior temporal lobectomy, patients found it difficult to understand ambiguous sentences, particularly the less prominent meaning of the sentence. Left temporal lobectomy affects comprehension of syntactic ambiguities much more than right temporal lobectomy; both affect comprehension of lexical ambiguities.

There are several other studies which implicate the left ATL in sentence processing. Grossman, Payer, Onishi, D'Esposito. Morrison, Sadek, & Alavi (1998) found that sentence processing impairment was correlated with hypoperfusion in the ATL and the LIFG for a group of patients suffering from frontal lobe degeneration. Dronkers, Wilkins, Van Valin, Redfern, & Jaeger (1994) showed that for a diverse group of patients who shared a deficit in morphosyntactic processing, lesions overlapped only in the anterior superior temporal gyrus, coinciding with a portion of our activation. Thus the ATL apparently plays some role in sentence processing, although it is difficult to say what exactly it does.

We suggest that sentence processing invokes encoding processes which can also be used during intentional study tasks. Lack of encoding of the relevant information into memory may account for the difficulty which anterior lobectomy patients find in interpreting the second meanings of ambiguous sentences and may affect comprehension of complex sentences as well. An observation supporting this hypothesis is that the N400 repetition effect appears to last longer while reading sentences than word lists. Van Petten, Kutas, Kluender, Mitchener, & McIsaac (1993) showed that the repetition effect remains significant even after twenty words (cf. Karayanidis et al.'s 6). This may be longer than the repetition effect seen in unstudied word lists; however this point needs further research, as no explicit comparison has been made.

Conclusion

We have discussed a number of neuroimaging studies which show that there are three areas of cortex which are involved in sentence processing. We have shown that each of the three has a different function in sentence processing, as they respond to different variables.

Left posterior middle to superior temporal cortex shows a straightforward effect of structural complexity: it shows least activation for word lists, more for simple sentences, and increasing activation as structural complexity increases further. The best characterization of the cognitive function of this area is that it is involved in processing some aspect of sentence structure (e.g. syntactic or semantic structure).

The LIFG, on the other hand, shows increasing activation under a combination of *Lexical Verbatim Memory* load and *Phrase Structure Memory* load, as shown in Experiment 1. The hypothesis that this area maintains both lexical and phrase structure information in memory during sentence processing explains this pattern of activation. In Experiment 2, we showed that lexical memory load and phrasal memory load cause interacting effects in this area. This suggests that the memory function in this area does not consist of two separate, overlapping networks. Rather both types of representation are apparently competing for the same resources.

The third area, lateral ATL, does not respond to structural complexity at all, but it does show increased blood flow relative to word lists for all sentence types. This pattern of activation suggests that the area does not actually construct a sentential structure; otherwise, we would expect to see effects of complexity. Other evidence out of the literature suggests that this area is involved in encoding lexical information under certain circumstances. We suggest that this happens automatically during sentence processing, although it can also be used during conscious study. Damage to this area does not cause dramatic impairment of comprehension, but when lexical information is necessary later in sentence processing there is an effect, such as when retrieving a second meaning of an ambiguous sentence (Zaidel et al., 1995).

Although we have claimed that lexical and phrasal information are both maintained in the frontal lobe and suggested that the information is encoded via processing in the anterior lateral temporal lobe, the actual content of what is encoded and maintained remains underspecified by these results. It could be syntactic information only, alternatively it may include semantic information. It is even possible that only the identity of the lexical item or phrase is maintained and used as a pointer to some other location where more extensive information may be retrieved if necessary. These issues will need further research for clarification.

Data from the neuroimaging studies discussed here raise several issues for models of sentence comprehension. The first concerns the dissociation of processing and storage. The data discussed here suggest that processing and working memory must be distinguished from each other as separate functions. Several recent theories of comprehension assume one common resource for both (Just & Carpenter, 1992; Gibson, 1998); these would have to be expanded to explain the existence of a separate processing mechanism. A second issue concerns encoding; the hypothesis that the ATL is primarily active in encoding suggests that this function must also be separated from maintenance more explicitly than has been done in most theories. A lot of research remains to be done to investigate these issues. Other interpretations are available for portions of the data reported here; we feel that the hypotheses presented here represent the best explanation for the entire data set. However, the main point is that researchers who are interested in developing a neurologically plausible model of working memory and sentence processing should be able to account for the pattern of data reported here.

Notes

[1] I.e., if the weights of the Lexical Memory Load are: Word List = 4, Very Complex Sentences = 3, Complex Sentences = 2, and Simple Sentences = 1, and the Phrasal Memory Load weights are: Word List = 1, Very Complex Sentences = 4, Complex Sentences = 3, and Simple Sentences = 2, the combined weights will be: Word List = 5, Very Complex Sentences = 7, Complex Sentences = 5, and Simple Sentences = 3. The actual weights used are equivalent to these, except for centering around 0.

[2] The status of the ambiguous sentences is not entirely clear, as it depends on the circumstances under which words are dismissed from the lexical store. We assumed maintenance until resolution (e.g. Frazier & Rayner, 1987).

[3] These weights are virtually, but not quite linear, chosen because zero cannot be used if the means as well as variances are to be considered in calculating the correlation and the weights must sum to zero.

[4] If the area reaches maximal blood flow in the double load condition, it might appear that the effect of both loads combined is less than the effect achieved by a single factor. This sort of interaction would thus not choose between the models.

[5] Materials for this experiment and Experiment 1 available on request.

References

Baddeley, A.D. (1986). *Working Memory*. Oxford: Clarendon Press.

Bavelier, D., Corina, D., Jezzard, P., Padmanabhan, S., Clark, V.P., Karni, A., Prinster, A., Braun, A., Lalwanim A., Rauschecker, J.P., Turner, R. & Neville, H. (1997). Sentence reading: A functional MRI study at 4 Tesla. *Journal of Cognitive Neuroscience, 9*, 664–686.

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiak, R.S.J. & Frith, C.D. (1994). The role of the right hemisphere in the interpretation of figurative aspects of language: A positron emission tomography activation study. *Brain, 117*, 1241–1253.

Caplan, D. & Waters, G. (1999). Verbal working memory and sentence comprehension. *Brain and Behavioral Science, 22*, 77–126.

Dronkers, N.F., Wilkins, D.P., Van Valin, R.D., Jr., Redfern, B.B. & Jaeger, J.J. (1994). A reconsideration of the brain areas involved in the disruption of morphosyntactic comprehension. *Brain and Language, 47*, 461–462.

EU Concerted Action on Functional Imaging. (1996). Reproducibility of PET activation studies: Lessons from a multi-centre European experiment. *Neuroimage, 4*, 34–54.

Fedio, P. & Van Buren, J.M. (1974). Memory deficits during electrical stimulation of the speech cortex in conscious man. *Brain and Language, 1*, 29–42.

Fiez, J.A., Raife, E.A., Balota, D.A., Schwarz, J.P. & Raichle, M.E. (1996). A positron emission tomography study of the short-term maintenance of verbal information. *Journal of Neuroscience, 16*, 808–822.

Frazier, L. & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory Lang, 26*, 505–526.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C. & Evans, A.C. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping, 1*, 214–220.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.B., Heather, J.B. & Frackowiak, R.S.J. (1995a). Spatial registration and normalisation of images. *Human Brain Mapping, 2*, 165–189.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frith, C.D. & Frackowiak, R.S.J. (1995b). Statistical parametric mapping and functional imaging: A general linear approach. *Human Brain Mapping, 2*, 189–210.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*, 1–76.

Grasby, P.M., Frith, C.D., Friston, K.J., Simpson, J., Fletcher, P.C. & Frackowiak, R.S.J. (1994). A graded approach to the functional mapping of brain areas implicated in auditory-verbal memory. *Brain, 117*, 1271–1282.

Grossman, M., Payer, F., Onishi, K., D'Esposito, M., Morrison, D., Sadek, A. & Alavi, A. (1998). Language comprehension and regional cerebral defects in frontotemporal degeneration and Alzheimer's disease. *Neurology, 50*, 157–163.

Just, M.A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122–149.

Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F. & Thulborn, K.R. (1996). Brain activation modulated by sentence comprehension. *Science, 274*, 114–116.

Karayanidis, F., Andrews, S., Ward, P.B. & McConaghy, N. (1991). Effects of inter-item lag on word repetition: An event-related potential study. *Psychophysiology, 28*, 307–319.

Marcus, M.P. (1980). *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.

Martin, R.C. (1993). Short-term memory and sentence processing: Evidence from neuropsychology. *Memory and Cognition, 21*, 176–183.

Mazoyer, B.M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L. & Mehler, J. (1993). The cortical representation of speech. *Journal of Cognitive Neuroscience, 5*, 467–479.

Ojemann, G.A. & Dodrill, C.B. (1985). Verbal memory deficits after left temporal lobectomy for epilepsy. *Journal of Neurosurgery, 62*, 101–107.

Paulesu, E., Frith, C.D. & Frackowiak, R.S.J. (1993). The neural components of the verbal component of working memory. *Nature, 362*, 342–344.

Petrides, M., Alivisatos, B., Meyer, E., & Evans, A.C. (1993). Functional activation of the human frontal cortex during the performance of verbal working memory tasks. *Proceedings of the National Academy of Science USA, 90*, 878–882.

Petrides, M. (1996). Lateral frontal cortical contribution to memory. *Seminars in the Neurosciences, 8*, 57–63.

Rugg, M.D., Roberts, R.C., Potter, D.D., Pickles, C.D. & Nagy, M.E. (1991). Event-related potentials related to recognition memory: Effects of unilateral temporal lobectomy and temporal lobe epilepsy. *Brain, 114*(5), 2313–2332.

Schnyer, D.M., Allen, J.J.B. & Forster, K.I. (1997). Event-related brain potential examination of implicit memory processes: Masked and unmasked repetition priming. *Neuropsychology, 11*, 243–260.

Stowe, L.A., Broere, C.A.J., Paans, A.M.J., Wijers, A.A., Koster, J. & Vaalburg, W. (1997). Working memory components in a language task. *Neuroimage, 4*, 552.

Stowe, L.A., Broere, C.A.J., Paans, A.M.J., Wijers, A.A., Mulder, G., Vaalburg, W. & Zwarts, F. (1998). Localizing components of a complex task: Sentence processing and working memory. *Neuroreport, 9*, 2995–2999.

Stowe, L.A., Paans, A.J.M., Wijers, A.A., Zwarts, F., Mulder, G. & Vaalburg, W. (1999). Sentence comprehension and word repetition: A positron emission tomography investigation. *Psychophysiology, 36*, 786–801.

Stromswold, K., Caplan, D., Alpert, N. & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language, 52*, 452–473.

Talairach, J. & Tournoux, P. (1988). *Co-Planar Stereotaxic Atlas of the Human Brain*. New York, NY: Thieme Medical Publishers.

Tzourio, N., Nkanga-Ngila, B. & Mazoyer, B. (1998). Left planum temporale surface correlates with functional dominance during story listening. *Neuroreport, 9*, 829–1233.

Van Petten, C., Kutas, M., Kluender, R., Mitchener, M. & McIsaac, H. (1993). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience, 3*, 131–149.

Woods, R.P., Cherry, S.R. & Mazziotta, J.C. (1992). A rapid automated algorithm for accurately aligning and reslicing positron emission tomography images. *Journal of Computer Assisted Tomography, 16*, 620–633.

Zaidel, D.W., Zaidel, E., Oxbury, S.M. & Oxbury, J.M. (1995). The integration of sentence ambiguity in patients with unilateral focal brain surgery. *Brain and Language, 51*(3), 458–468.

The time course of information integration in sentence processing

Michael J. Spivey, Stanka A. Fitneva, Whitney Tabor and Sameer Ajmani
Cornell University/University of Connecticut/MIT

Recent work in sentence processing has highlighted the distinction between serial and parallel application of linguistic constraints in real time. In looking at context effects in syntactic ambiguity resolution, some studies have reported an immediate influence of semantic and discourse information on syntactic parsing (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998). However, in looking at the effects of various constraints on grammaticality judgments, some studies have reported a temporal precedence of structural information over semantic information (e.g., McElree & Griffith, 1995, 1998). This chapter points to some computational demonstrations of how an apparent temporal dissociation between structural and non-structural information can in fact arise from the dynamics of the processing system, rather than from its architecture, coupled with the specific parameters of the individual stimuli. A prediction of parallel competitive processing systems is then empirically tested with a new methodology: speeded sentence completions. Results are consistent with a parallel account of the application of linguistic constraints and a competitive account of ambiguity resolution.

Introduction

For more than a couple of decades now many psycholinguists have been investing a great deal of effort into elucidating the “sequence of stages” involved in the comprehension of language. Emphasis has been placed on the question: When do different information sources (syntax, semantics, etc.) get extracted from the linguistic input? One answer to this question that has been very influential is that the computation of syntax precedes the computation of semantics and pragmatics (e.g., Frazier & Fodor, 1978; Ferreira & Clifton, 1986; McElree & Griffith, 1995, 1998). One opposing answer that is gaining support is that there are no architecturally imposed delays of information during sentence processing, that all relevant information sources are extracted and used the moment they are received as input (MacDonald, Pearlmutter, & Seidenberg, 1994; Spivey-Knowlton & Sedivy, 1995; Trueswell & Tanenhaus, 1994). Recently, however, some disillusionment has been expressed concerning the question itself:

> "Given the wide range of results that have been reported, it seems most appropriate at the moment to determine the situations in which context does and does not have an influence on parsing, rather than continue the debate of *when* context has its impact." (Clifton, Frazier, & Rayner, 1994, p. 10, italics theirs).

Perhaps one way to redirect the "when" question to better understand the mixed results in the literature would be to turn it into a "how" question. Could *the manner in which* various information sources combine during sentence processing wind up explaining why context sometimes has an early influence and sometimes a late influence? It seems clear that a treatment of this kind of question will require some theoretical constructs and experimental methodologies that are new to sentence processing, as well as some careful attention to lexically-specific variation in stimulus items. The purpose of this chapter is to describe some of these new approaches and the implications that they have for claims about the time course of information integration in sentence processing.

Nonlinear dynamics

Over the past fifteen years, a number of researchers have designed dynamical models of sentence processing (Cottrell & Small, 1983; Elman, 1991; McClelland & Kawamoto, 1986; McRae, Spivey-Knowlton & Tanenhaus, 1998; Selman & Hirst, 1985; Spivey & Tanenhaus, 1998; St. John & McClelland, 1990; Tabor & Hutchins, 2000; Tabor, Juliano, & Tanenhaus, 1997; Waltz & Pollack, 1985; Wiles & Elman, 1995; see also Henderson, 1994, and Stevenson, 1993, for hybrid models that combine rule-based systems with some fine-grain temporal dynamics). A dynamical model is a formal model that can be described in terms of how it changes. Typically, such models take the form of a differential equation,

$$\mathrm{d}\mathbf{x}/\mathrm{d}t = f(\mathbf{x}) \tag{1}$$

with an initial condition, $\mathbf{x} = \mathbf{x}_0$. Here $\mathbf{x}$ is a vector of several dimensions and t is time. The equation says that the change in $\mathbf{x}$ can be computed from the current value of $\mathbf{x}$. The behavior of such systems is often organized around *attractors*, or stable states ($f(\mathbf{x}) = 0$) that the system goes toward from nearby positions. Nearby attractors will tend to have a strong "gravitational pull," and more distant attractors will have a weaker pull. The most common strategy is to assume that initial conditions are determined by the current context (e.g., a string of words like "Alison ran the coffee-grinder") and that attractors correspond to interpretations of that context (e.g. Alison is the agent of a machine-operation event where the machine is a coffee-grinder). The model, (Eq. 1), is called *nonlinear* if f is a nonlinear function. Nonlinearity is a necessary consequence of having more than one attractor. Since languages contain many sentences with different interpretations (and many partial sentences with different partial interpretations), dynamical models of sentence processing are usually highly nonlinear. The potential for feedback in Equation (1) – the current value of a particular dimension of $\mathbf{x}$ can depend on its past value – is also important. It can cause the system to vacillate in a complex manner before settling into an attractor.

Many dynamical sentence processing models are implemented in connectionist models (i.e., artificial neural networks). The "neural" activation values correspond to the dimensions of the vector $\mathbf{x}$ and the activation update rules correspond (implicitly) to the function, f. In some such cases (e.g., Elman, 1991; St. John & McClelland, 1990; Wiles & Elman, 1995), Equation (1) is replaced by an iterated mapping (Eq. 2):

$$\mathbf{x}_{t+1} = f(\mathbf{x}_t) \tag{2}$$

which makes large discrete, rather than continuous, or approximately continuous, changes in the activation values. Typically, such discrete models are designed so that words are presented to the model one at a time and activation flows in a feedforward manner upon presentation of a single word. This architecture makes no use of the feedback potential of Equation (1), so the dynamics of single word-presentations are trivial; but over the course of several word presentations, activation can flow in circuits around the network, and feedback (as well as input) can contribute significantly to the complexity of the trajectories (Wiles & Elman, 1995). Other proposals allow feedback to cycle after every input presentation. Some such proposals present all the words in a sentence at once (Selman & Hirst, 1985), while others use serial word presentation and allow cycling after each word (Cottrell & Small, 1983; McRae et al., 1998;

Spivey & Tanenhaus, 1998; Tabor & Hutchins, 2000; Tabor et al., 1997; Waltz & Pollack, 1985; Wiles & Elman, 1995).

Models which allow feedback to cycle after each input make fine-grained predictions about the time course of information integration in sentence processing. In fact, several existing dynamical models of sentence processing exhibit at least simple forms of vacillation. For example, when presented with the string, "Bob threw up dinner," Cottrell and Small (1983)'s model shows a node corresponding to the *purposely propel* sense of "throw" first gaining and then losing activation (see also Kawamoto, 1993). Tabor et al. (1997) define a dynamical system in which isolated stable states correspond to partial parses of partial strings. At the word "the" in the partial sentence, "A woman insisted the…", for example, they observe a trajectory which curves first toward and then away from an attractor corresponding to the (grammatically impossible) hypothesis that "the" is the determiner of a direct object of "insisted," before reaching an (grammatically appropriate) attractor corresponding to the hypothesis that "the" is the determiner of the subject of an embedded clause. Syntax-first models of sentence processing (Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1998) are typically designed to restrict vacillation to a very simple form: first one constraint system (syntax) chooses a parse instantaneously and then another one (e.g., semantics) revises it if necessary.

In *lexical* ambiguity resolution, there is evidence for another simple form of vacillation. Tanenhaus, Leiman, and Seidenberg (1979, see also Swinney, 1979, and Kawamoto, 1993), found that ambiguous words exhibit temporary (approx. 200 ms) priming of both meanings (e.g. "rose" as *flower* and "rose" as *moved up*) even in a context where only one meaning is appropriate (e.g. "She held the rose"). Soon thereafter, the contextually inappropriate meaning ceases to exhibit priming. Recent constraint-based models of parsing predict effects in syntactic ambiguity resolution that significantly resemble the effects in lexical ambiguity resolution (MacDonald et al., 1994; Spivey & Tanenhaus, 1998; Trueswell & Tanenhaus, 1994). In contrast, typical syntax-first models of sentence processing posit syntactic parsing strategies that immediately select a single structural alternative (Frazier & Fodor, 1978; Frazier, 1987). To test these two types of models, what we need are experimental methodologies that provide access to the moment-by-moment representations computed during syntactic parsing. Do we see early vacillation between syntactic alternatives, as is seen between lexical alternatives? In this chapter, we will discuss two experimental methodologies that show promise for revealing the temporal dynamics of syntax-related information during sentence processing: speeded grammaticality judgments (McElree & Griffith, 1995, 1998), and speeded sentence completions. Results from these methodologies are simulated by a nonlinear competition algorithm called Normalized Recurrence (Filip, Tanenhaus, Carlson, Allopenna, & Blatt, this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, & Hanna, 2000).

Normalized Recurrence is a relatively simple dynamical system in which the alternative interpretations that a given stimulus might map onto are treated as localist units in the network. The multiple information sources that might give evidence for these different interpretations are then given localist units representing their support for the various stimulus-interpretation mappings. See Figure 1. First, each of the information sources has their previous activations normalized to a sum of 1.0:

$$S_{c,a}(t) = S_{c,a}(t-1)/\sum_{a} S_{c,a}(t-1) \tag{3}$$

where $S_{c,a}(t)$ is the activation of the cth information source supporting the ath alternative at time t. Next, the information sources combine in a weighted sum at the interpretation units:

$$I_a(t) = \sum_{c} [w_c * S_{c,a}(t)] \tag{4}$$

where $I_a(t)$ is the activation of the ath alternative interpretation at time t, and the weights, w_c – one for each information source – sum to 1.0. When an interpretation unit reaches a criterion activation, some appropriate output is stochastically triggered, such that the activation function across the different interpretation units is treated as a probability density function describing the likelihood of each interpretation triggering its preferred action (e.g., looking at the object corresponding to that interpretation, Spivey-Knowlton & Allopenna, 1997). The final computation that completes a cycle of competition is feedback from the integration units to the information sources, where an information source's weighted activations are scaled by the resulting interpretation node's activation and sent as cumulative feedback to the information source (Eq. 5). This feedback is how the model gradually approaches a stable state, coercing not only the interpretation units to settle on one alternative, but also coercing the information sources to conform.

$$S_{c,a}(t+1) = S_{c,a}(t) + I_a(t) * w_c * S_{c,a}(t) \tag{5}$$

It should be noted that, unlike many connectionist models, this network does not "learn" its weights. Instead, they are each set to $1/n$ (where n is the number

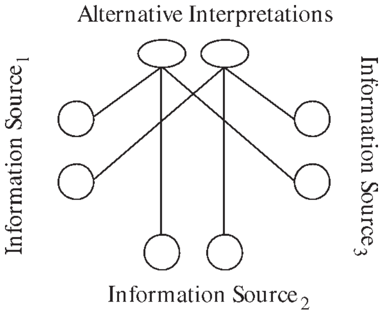


Figure 1. A schematic of the Normalized Recurrence competition algorithm with three information sources competing over two alternatives.

of information sources, Spivey & Tanenhaus, 1998), or the entire weight space is sampled and the weights with the best fit to the data are used. For example, McRae et al. (1998) designed a Normalized Recurrence network to simulate sentence completion data and self-paced reading data on the Reduced Relative/Main Clause ambiguity. Initially combining three information sources (a general main-clause bias, thematic fit information, and verb tense frequency), and sampling the entire range of weights, it was found that the best weights for fitting that data set were the following: main-clause bias =.5094, thematic fit =.3684, and verb tense frequency =.1222. However, with different stimulus sets and different presentation circumstances that emphasize their information sources differently, the weights for these constraints are likely to vary somewhat.

Highly simplified in comparison to attractor networks that use distributed representations (e.g., Tabor et al., 1997), Normalized Recurrence thereby allows an easily interpreted "peek" into the system's state at any point in time. Panels A and B of Figure 2 show some generic examples of the activation of two alternative interpretations competing over time. Nonlinear trajectories through the state-space on the way toward settling on one alternative can produce complex behavior in the model. In fact, when several information sources compete over three or more interpretations, an alternative whose initial activation starts out in "second place" can sometimes wind up usurping the most active alternative and eventually become the final interpretation (Figure 2C).

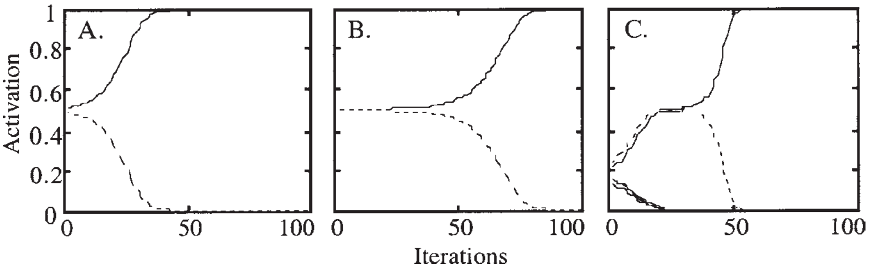


Figure 2. Example results from Normalized Recurrence. Panels A and B are from a network with an architecture like that in Figure 1. Panel C is from a network with four information sources competing over six alternatives. Note that the alternative that starts out with the highest activation (dashed line) ends up losing.

Measures of the activation of linguistic representations

While modeling allows a kind of "x-ray vision" into the internal working parts of a system that might be functioning in a fashion similar to that of the mind, psycholinguists are typically more interested in getting that kind of "x-ray vision" for the *actual* mind – not an idealized set of formulas intended to simulate the mind. To this end, a number of experimental methodologies have been used over the past couple of decades to tap into the salience of certain linguistic representations *during real-time language processing*. Most of them have been using differences in reaction times to infer relative activations of linguistic representations. It is assumed that a faster reaction time implies a representation with some unspecified amount of greater activation. Although this assumption seems fair enough, determining the mapping from latencies to activations has been largely ignored. What would be preferable would be to see experimental data reflecting the activation of a linguistic representation changing over time, much like those in Figure 2.

One recent example of this kind of "window" into the moment-by-moment activation of different linguistic representations is research with headband-mounted eyetracking (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). In experiments looking at spoken word recognition, it was observed that when participants were instructed to "pick up the candy" they tended to briefly fixate a *candle* before finally fixating and grasping the candy. In fact, plotting the probability of fixating the various objects across time produced curves that were surprisingly close to the lexical activation functions from the TRACE model of speech perception (McClelland & Elman, 1986) – not unlike those in Figure 2C.

Another recent example that shows similar time-slices in the temporal dynamics of linguistic representations is McElree and Griffith's (1995, 1998) use of the speed-accuracy trade-off (SAT) procedure with speeded grammaticality judgments. When the last word in a sentence makes it grammatical or ungrammatical, a rushed decision on this grammaticality is likely to be based on only partially complete representations. By applying signal detection theory to these rushed decisions over various time intervals, McElree and Griffith show a smooth, gradual increase in the detectability of the grammaticality over time – as measured by d-prime, which provides an index of a subject's sensitivity to a stimulus irrespective of his/her response criteria. In the following sections of this chapter, we will review some of McElree and Griffith's findings and conclusions, test the Normalized Recurrence competition algorithm on their results, as well as introduce some results from a new speeded response methodology: speeded sentence completions. We wish to illustrate how, with the recurrent interplay between experimental data and model simulations, we can iteratively refine a sound theory of the time course of information integration in sentence processing.

Serial stages in sentence processing

The first question that arises in understanding how a serial system might work is the size of the unit of computation. In this kind of treatment, a particular processing stage does not send output to the next stage until it has received (and performed its operations on) an entire unit of computation. In the case of sentence processing, a number of proposals have been forwarded for the size of such units. The temporally-extended unit of serial computation has been suggested to be as large as entire clauses (Fodor, Bever & Garrett, 1974) or as small as individual words (Frazier & Fodor, 1978). Alternatively, the serial system could be smoothly cascading, but have a kind of "raw transmission time" between modules (McClelland, 1979). For example, McElree and Griffith (1995) have postulated a $\sim$ 100 ms delay between the initial computation of subcategory information and the initial computation of thematic role information. More recent work has suggested a 200–400 ms delay between syntactic information and lexical information (McElree & Griffith, 1998).

McElree and Griffith's SAT analysis of speeded grammaticality judgments is particularly exciting in that it provides a glimpse into the activation of certain linguistic representational formats (syntax, thematic roles, subcategory constraints, etc.) in real time. In this task, subjects are presented grammatical and ungrammatical sentences, and instructed to, as quickly as possible, judge their grammaticality. As our interest is in *when* various information sources begin to affect the grammaticality judgment, our primary focus will be on the *ungrammatical* sentences. According to McElree and Griffith, sentences like (1a) become ungrammatical at the final word due to a subcategorization violation, because the verb *agreed* is intransitive. In contrast, sentences like (2a) become ungrammatical at the final word due to a thematic role violation, because the Agent of the verb *loved* must be animate (and books are inanimate). In order to compute d-primes via signal detection theory (Green & Swets, 1966) for the SAT task, the ungrammatical sentences (1a & 2a) provided the signal+noise trials and the grammatical sentences (1b & 1b) provided the noise trials. (Thus, the SAT analysis actually treats the task as one of "ungrammaticality detection," rather than grammaticality judgment.)

(1) a. Some people were agreed by books. (Subcategory Violation Sentence)
b. Some people were agreed with rarely. (Subcategory Control Sentence)

(2) a. Some people were loved by books. (Thematic Violation Sentence)
b. Some books were loved by people. (Thematic Control Sentence)

In the SAT version of this speeded grammaticality judgment task, the target sentences were presented to subjects one word at a time in the center of the screen in a noncumulative fashion. Immediately, or shortly, after presentation of the last word in the sentence, a tone would signal to the subject that she/he must respond as to the grammaticality of the sentence within 300 ms. The temporal interval between the onset of the last word and the presentation of the tone was either, 14, 157, 300, 557, 800, 1500, or 3000 ms. (After a couple hours of practice, subjects eventually became skilled at forcing themselves to respond within 300 ms of the tone, even though their processing of the sentence, at the very short intervals, was incomplete.) As seen in Figure 3, mean d-prime values (across six subjects) at the shortest intervals were at or near chance performance. However, at the intermediate and later intervals, performance clearly improved in a smooth, graded fashion. Interestingly, detection of ungrammatical sentences was slightly better for subcategory violations (filled circles) than for thematic role violations (open circles).

One possible interpretation of the data in Figure 5 is that they come from two different exponential functions, each with its own x-intercept. For example, if one extended the left hand portions of the two curves in the simple downward direction implied by the data points at those first few intervals, they would reach a d-prime of zero at slightly different places along the horizontal time axis. If one assumes a dual-process serial processing system, one could infer from these different x-intercepts (as long as the variability in processing time is equal across conditions) that subcategorization information "becomes operative," and informs the detection of ungrammaticality, about 100 ms before thematic role information does. In fact, using an exponential equation (Eq. 6) to fit the data points, McElree and Griffith (1995) suggest exactly that.

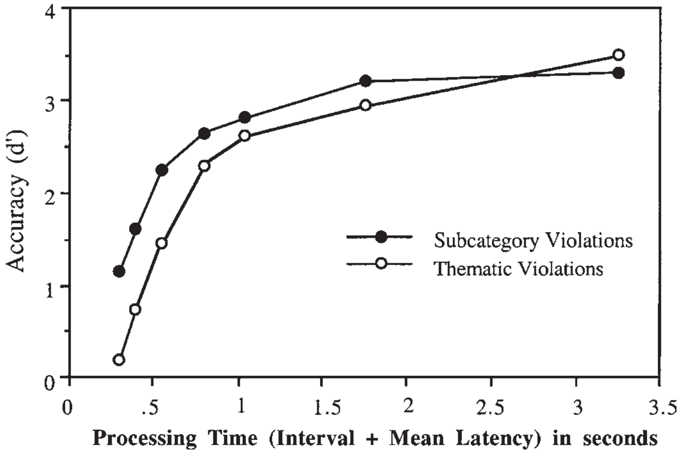


Figure 3. Accuracy of grammaticality detection for subcategory and thematic violations. (Adapted from McElree & Griffith, 1995.)

$$d'(t) = \lambda(1 - e^{-\beta(t-\delta)}), \text{ for } t > \delta, \text{ else } 0 \qquad (6)$$

In Equation 6, accuracy (d') at each fraction of a second t is determined by three free parameters: λ, β, and δ. As the scalar of the entire equation, λ determines the asymptote of the curve, where improvement in accuracy over time tapers off and total accuracy "maxes out." As the scalar in the exponent of e, β determines the rate of rise in d-prime over time, or the slope of the curve as it departs from zero. Finally, as the time relative (because it is subtracted from t) portion of the exponent of e, δ determines the x-intercept of the curve, or the point in time immediately before accuracy climbs above chance. Thus, at the point in time where the curve is to reach zero, t and δ will be equal to one another, and t–δ will equal zero, making the entire equation equal zero.

As negative d-primes would imply a perverse pattern in the data, the last part of the equation insures that for values of t that would produce negative d-primes, d-prime is instead rectified to zero. To fit these parameters to the

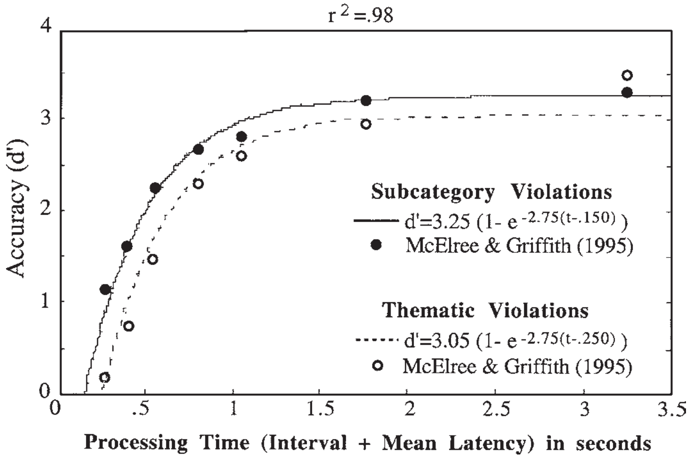


Figure 4. Accuracy of grammaticality detection and approximated fits from McElree and Griffith's (1995) dual-process serial processing account: Equation 6. The fit to the data accounts for 98% of the variance. (Adapted from McElree & Griffith, 1995.)

data curves, McElree and Griffith apply Chandler's (1967) Stepit algorithm that searches the parameter space to find the best-fitting parameter values – somewhat similar to that carried out by McRae et al. (1998) in setting the weights for the Normalized Recurrence competition algorithm. Figure 4 shows an example of the data being fit by the equation, using different δ values (and different λ values) for subcategory violations and thematic violations.

Importantly, this equation provides a standardized method of estimating where the d-prime curves over time would reach zero if they had been sampled from an exponential function that actually had an x-intercept. However, it is certainly possible, in principle, that the data points in Figure 3 do not come from a function with a real x-intercept, but instead come from a function that never actually touches the x-axis, such as the logistic in Figure 5. With no actual x-intercepts (instead, each curve's y-intercept signifies a nonzero d' at timestep 1 – and is rectified to zero at timestep zero, similar to the rectification done in McElree and Griffith's equation), it would be impossible to make any claims about separate processes "becoming operative" at different discrete points in time.

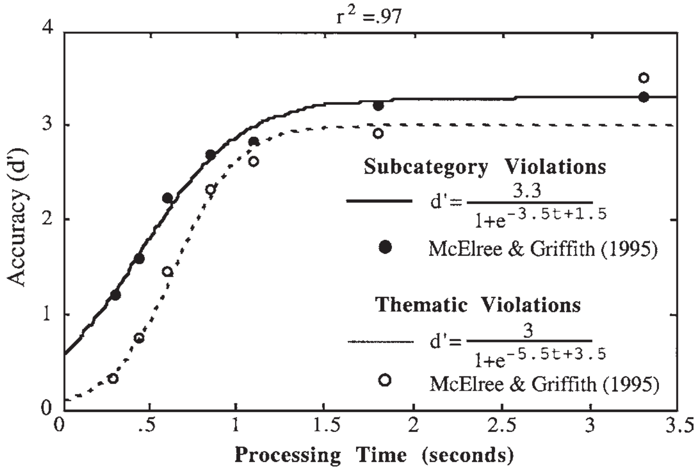


Figure 5. Accuracy of grammaticality detection and approximated fits from a logistic function. The fit to the data accounts for 97% of the variance.

Parallel integration in sentence processing

McElree and Griffith (1995) anticipated that, far from requiring a serial-stage account of sentence processing, their results might in fact be accommodated by certain parallel models of information processing. As the sigmoidal function in Figure 5 is a natural result of competition in Normalized Recurrence, we decided to test Normalized Recurrence on McElree and Griffith's results. To apply the Normalized Recurrence competition algorithm to this grammaticality judgment task, the two information sources (subcategorization and thematic roles) were each condensed into two values: one for the probability of the sentence being grammatical, and one for the probability of the sentence being ungrammatical, based on that information source's strength of constraint. Thus, rather than becoming operative at an earlier point in time, subcategorization information may simply provide a probabilistically stronger constraint on grammaticality than thematic role information does. That is, it may be the case that thematic fit is more violable in our typical language experience (e.g., "This computer hates me.") than subcategorization constraints (e.g., "I slept the day away."). Figure 6 shows a schematic diagram of the Normalized Recurrence model, with bidirectional connections between the information sources and the integration layer (where grammaticality judgment takes place) allowing converging/conflicting biases to be passed back and forth.

As in other Normalized Recurrence simulations, competition between mutually exclusive representations ("grammatical" and "ungrammatical," in this case) proceeded with three critical steps for each iteration of the model: 1) Normalization of information sources (Eq. 3), 2) Integration of information sources (Eq. 4, where $w=1/n$), and 3) Feedback from the integration layer to the information sources (Eq. 5). An important difference between this Normalized Recurrence simulation and previous ones is that the model was not allowed to iterate until reaching a criterion, because duration of competition (e.g., reaction time) was not the measure of interest. Rather, the model was stopped at various intervals and the activations of the interpretation units were treated as probabilities of "grammatical" and "ungrammatical" responses. In order to prevent unnaturally high d-primes, each interpretation unit has a maximum of .95 activation in this first simulation.

With each iteration, the model gets more and more "confident" in one of these decisions. Of course, in the case of only two competing alternatives, the moment one decision is greater in activation than the other, it is obvious that (in this deterministic version of the competition algorithm) the *current* winner will be the *ultimate* winner. However, for simulating the time course of information integration, we need to allow the model to settle toward some criterion activation, especially if we consider the possibility that different response mechanisms (e.g., manual response, vocal response, or eye movements) may have different criteria for execution.

In the first simulation of McElree and Griffith's (1995) SAT version of the speeded grammaticality judgment task, the model was given input values indicating either a grammatical sentence, subcategory violation sentence, or the-

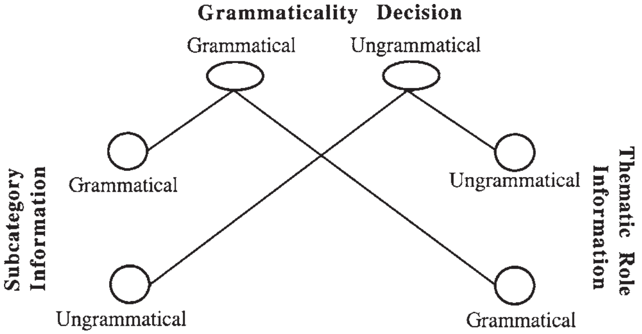


Figure 6. Schematic diagram of the Normalized Recurrence model designed to simulate the results of McElree and Griffith (1995).

matic violation sentence. The model was then allowed to iterate, gradually converging toward a decision on the grammaticality of the input, until an interruption point was reached, at which time the integration layer's values were recorded for the probability of a correct response (as though the model were being interrupted and forced to make a decision). For grammatical sentences, the input value for the grammatical node in each constraint was .51, and thus the input value for the ungrammatical node in each constraint was .49. When each iteration is treated as 50 ms, these values produce "grammatical" response times that approximate those from McElree and Griffith (1995). For a subcategory violation, the input values for the subcategory nodes were .2 grammatical and .8 ungrammatical, whereas for a thematic violation, the input values for the thematic role nodes were .4 grammatical and .6 ungrammatical.

To compute d-primes at each time step of the model, the activation of the "grammatical" integration node after a grammatical input was treated as the percentage of *hits*, and the activation of the "ungrammatical" integration node after ungrammatical input was treated as the percentage of *correct rejections*. Figure 7 compares McElree and Griffith's data to the model's d-prime values as a function of processing time. The first thing to notice is that the model reaches asymptote much more abruptly than in the human data. This is primarily due to a .95 maximum imposed on the activations in order to prevent d-primes of 4+. Much of the smooth, graded approach to asymptote exhibited by Normalized Recurrence actually takes place between .95 and 1.0 activation. With that range omitted, this first simulation rather suddenly hits a sharp maximum before it is through with the steeply rising portion of its sigmoid function over time. Despite this obvious weakness of the first simulation, the critical portion of the data, where the early measurements for subcategory and thematic role violations are dissociated, is well accounted for by the model. Whereas McElree and Griffith's (1995) account of the data assumes that the curves for subcategory and thematic violations must depart from zero d-prime (or "become operative") at different points in time, Normalized Recurrence accounts for this portion of the data using two sigmoidal curves that "become operative" at the same time, but one has a stronger initial bias backing it up.

Improvements on this first simulation can be achieved in a number of ways. There are essentially six parameters in this model that can be manipulated: 1) the "grammatical input" value, 2) the "subcategory violation" value, 3) the "thematic violation" value, 4) the weights (since each pair must sum to 1, each of these first four terms counts as a single model parameter), 5) the activation rectification limit, and 6) the amount of time each iteration corresponds to. In the first simulation, the space of parameters 2 and 3 was searched (in steps of .05) to converge on an approximate fit to the data.

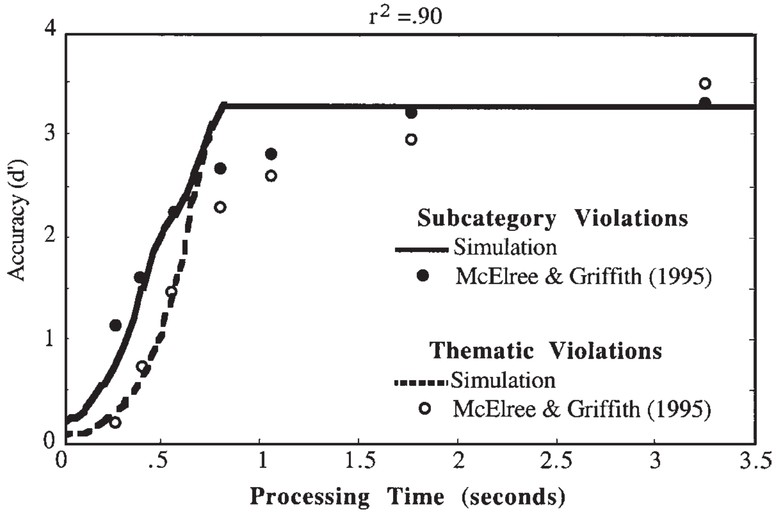


Figure 7. Accuracy of grammaticality detection (McElree & Griffith, 1995) and results of the first simulation with Normalized Recurrence. The fit to the data accounts for 90% of the variance.

In this next simulation (Figure 8), parameters 5 and 6 were modified to converge on a fit to the data. The input values for the different experimental conditions were identical to those of the first simulation. However, instead of a strict activation rectification, the normalization function (Eq. 3) added a small uniformly random value between 0 and .2 to the denominator at each time step (cf. Heeger, 1993). Also, the time constant was reduced to 30 ms per iteration.

The experimental results of McElree and Griffith's (1995) SAT version of the speeded grammaticality judgment task are certainly intriguing. Unlike most experimental methodologies in the field of sentence processing, the SAT procedure provides a window into preliminary incomplete representations that are in the process of being computed as information continuously accrues. However, attempting to extrapolate from the sampled d-primes to the underlying function's x-intercept via an exponential function may prematurely imply separate discrete points in time at which different linguistic processors "become operative." Instead, the results of these simulations suggest that the sampled d-primes over time may come from a system that integrates its different information sources simultaneously but with differing strengths. A weaker signal (e.g., thematic constraints) that "becomes operative" *at the same time* as a stronger signal (e.g., subcategory constraints) will still take longer to rise above the noise inherent in a probabilistic information processing system (Figure 8). Although the performance of the Normalized Recurrence model is encouraging, the simulations presented here did not quite account for as much of the variance in the data as did McElree and Griffith's (1995) six-parameter Stepit-driven exponential fit (Eq. 5). Moreover, McElree and Griffith's (1998) more recent findings hold still more challenges for a parallel processing system, such as syntactic island constraints having higher d-primes than lexically-specific constraints, and crossings between different curves of d-prime over time. Future work with this model will explore further manipulation of the parameters of this network.

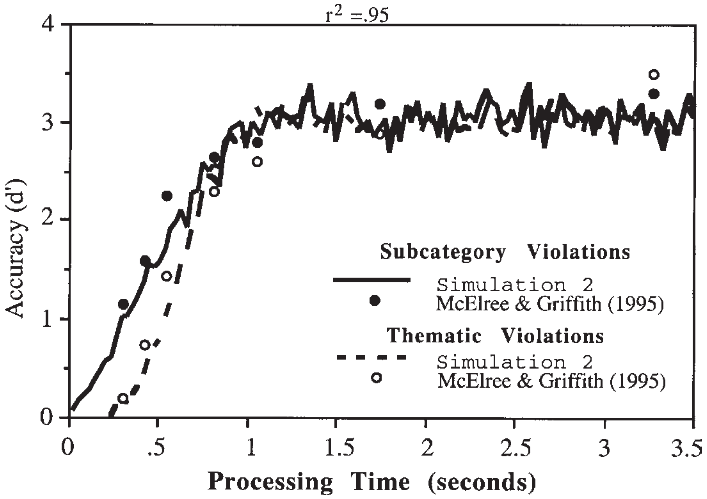


Figure 8. Accuracy of grammaticality detection (McElree & Griffith, 1995) and results of the revised simulation with Normalized Recurrence. The fit to the data accounts for 95% of the variance.

A prediction from competition

Many competition-based models (and other dynamical models) of sentence processing assume that a representation's activation will have a relatively non-extreme value during early moments of processing, and will gravitate toward an extreme value (e.g., minimum or maximum) as time proceeds – modulo the occasional nonmonotonic vacillation. Note that, since its representations are localist nodes, Normalized Recurrence's attractors are corners in the state-space, and therefore a single run of the model with only two competing interpretations cannot exhibit vacillations. (Nonmonotonic behavior on one run of the model, such as that in Panel C of Figure 2, can only happen when several information sources compete over several interpretations, or when some stochasticity is added to the normalization function.)

When only two interpretations are competing in Normalized Recurrence, as one begins to increase in activation, the other must decrease, and they will continue on these trajectories monotonically. For example, with the ambiguity between a Main Clause (MC) and a Reduced Relative (RR) (3), input to the model that averages just barely in favor of the MC will cause the model to start out with equibiased representations for the MC and RR that gradually settle entirely in favor of the MC interpretation. In contrast, a model that posits a separate processing stage for syntactic biases followed by a stage for thematic role biases (e.g., Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1998) might predict zero activation of the RR representation early on, regardless of what thematic fit information suggests. If thematic role information strongly biases an RR interpretation (such as a *prisoner* being a good Patient and a poor Agent of a *capturing* event), the activation of the RR representation will, at later points in time, eventually accrue some positive activation.

(3) a. The prisoner captured *a rat and kept it as a pet.* (Main Clause)
b. The prisoner captured *by the guards was tortured.* (Reduced Relative)

Thus, the prediction made by Normalized Recurrence, and ruled out by the two-stage models, is the following: With sentence fragments of the form "The" -noun-verb "-ed", in which thematic role information strongly biases the RR structure, even early moments of processing should show nonzero activation of the RR representation. A further, more specific, prediction from Normalized Recurrence is that those particular sentence fragments in which all constraints conspire *just barely* in favor of the MC, should in fact elicit greater positive activation of the RR representation during the *early* moments of processing than during the *later* moments of processing.

To test these predictions, we have designed a novel experimental methodology: speeded sentence completions, in which participants read sentence fragments, one word at a time, and complete these sentences under various time constraints. They are allowed 300 ms, 600 ms, 900 ms, or 1200 ms to prepare the completion. The results from 63 participants are presented herein.

Speeded sentence completions

Each trial proceeded as follows: a red circle appeared in the center of the screen indicating where the words would be presented, each word of the sentence fragment was presented in a noncumulative fashion in the center of the screen for 500 ms, three periods then appeared indicating that the participant should start preparing a completion, then a green circle appeared indicating that a completion must begin within 300 ms. Participants found the task difficult at first, but a 20-trial practice session (with a 500 ms processing interval) was typically enough to acquaint the participant with the task.

In this experiment, the three periods were on the screen for 300, 600, 900, or 1200 ms. These four different processing-interval conditions were run as separate blocks. In each block of 20 trials, the first 10 were fillers(ranging from two to four words in length), allowing the participant to get accustomed to that particular processing-interval condition. The remaining 10 trials in the block had four critical sentences embedded among 6 fillers items. The order of these blocks was randomized for each participant.

Participants were instructed to speak into the microphone what first came to mind and not to censor themselves. They heard a beep if they started responding too soon (i.e., while '. . . ' was still on the screen), and saw a "Respond faster!" sign on the screen if they began their response more than 300 ms after the green circle appeared. After finishing a sentence, they pressed a key on the button box to advance to the next trial.

The critical sentences were constructed from sixteen verbs, each with a typical Agent and a typical Patient for that particular event. Agenthood and Patienthood ratings were taken from norms collected in the work of McRae et al. (1998), and the verb form frequencies (Simple Past Tense, Past Participle, and Base frequency) were taken from Kucera & Francis (1982). See Table 1.

To utilize as many data as possible, all responses that began 300–1500 ms after the onset of the three periods were included in the analysis. Responses that were too early or too late for their condition were counted as belonging to the temporally accurate processing-interval. For example, if during a block of trials with the 600 ms delay a response occurred at 950 ms, it was counted as belonging to the 900–1200 ms processing-interval bin. Responses that began after 1500 ms (3%) and responses that were incomplete and/or still ambiguous (6%) were excluded from analysis.

The overall results of this study are compelling. At the earliest measured point in time, the 300–600 ms bin, sentence fragments with Patient-like nouns show significantly more reduced relative completions than those with Agent-

Table 1. Stimuli used in Speeded Sentence Completions

Verb	Verb Frequency			Noun1	Thematic Fit		Noun2	Thematic Fit	
	SPast	PPart	Base		Ahood	Phood		Ahood	Phood
arrested	4	15	27	police	6.45	1.46	suspect	1.40	5.49
audited	0	1	3	government	6.17	3.00	taxpayer	2.72	6.16
captured	2	15	33	troops	5.97	3.87	prisoner	1.76	5.03
convicted	1	13	16	juror	6.61	1.32	criminal	1.45	5.87
cured	1	6	20	doctor	6.76	3.78	patient	1.37	6.14
executed	1	13	22	terrorists	6.05	4.03	hostages	1.66	4.95
graded	0	2	3	teacher	6.94	2.60	student	2.42	6.81
instructed	2	14	23	coach	6.74	2.11	trainee	1.66	6.22
investigated	2	16	38	auditor	6.25	2.22	theft	1.22	6.78
paid	1	95	256	man	5.50	3.65	tax	1.63	5.43
punished	1	8	14	parent	6.50	1.54	child	1.53	5.78
rescued	1	5	14	knight	5.97	1.68	victim	1.21	4.89
sent	1	74	172	manager	5.55	2.95	package	1.58	6.16
sentenced	1	8	9	judge	6.94	1.27	defendant	1.25	6.35
tortured	1	8	10	kidnapper	5.68	1.60	slave	1.29	5.57
worshipped	1	2	12	priest	6.67	4.05	goddess	1.50	6.73

like nouns (25% vs. 2%; p<.05). Figure 9 shows the percentage of RR completions for both good Patients and good Agents at the four processing-interval bins. When the sixteen items are averaged for each curve, the temporal dynamics from individual items cancel each other out, resulting in relatively flat curves that are consistently about 25% apart from one another. We do not see in the good Patient condition an initial near-zero percentage of RRs that gradually increases over time, as would be most naturally predicted by a syntax-first model. Nonetheless, although this result seems most consistent with a simultaneous integration of constraints account of sentence processing, a syntax-first model can always accommodate these findings by restricting the purely syntactic processing stage to the first 300 ms of processing.

The more specific prediction made by competition models is also borne out: that a particular sentence fragment in which all constraints conspire *just barely* in favor of the MC will elicit greater positive activation of the RR representation during the *early* moments of processing than during the *later* moments of processing. For example, "The prisoner captured..." elicited 35–40% RRs during the early delay conditions, and 0–10% RRs during the latter delay conditions. Syntax-first models are fundamentally incapable of explaining such a result, whereas Normalized Recurrence predicts this result quite naturally.

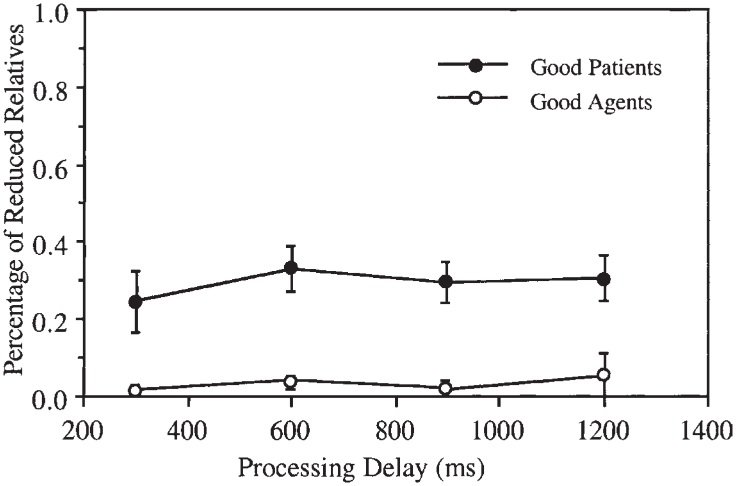


Figure 9. Overall results of the speeded sentence completion task.

Normalized Recurrence

As the Normalized Recurrence competition algorithm emerged in the context of the constraint-based lexicalist framework in sentence processing (e.g., Filip et al., this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998), it makes sense to apply the model to the lexically specific stimuli used in this experiment and average the two groups of 16 runs of the model for comparison with the averaged human data (Figure 9). (In fact, very different and inappropriate results would arise from instead averaging the stimulus parameters in the two groups of 16 items and running the model twice with those averaged values.)

Figure 10 shows a schematic diagram of the Normalized Recurrence simulation of the speeded sentence completions. We used the same three information sources as in McRae et al. (1998): SVO bias, Thematic Fit, and Verb Form Frequency. For the SVO bias, MC=.92 and RR=.08 (McRae et al., 1998). For the lexically specific biases, the values were taken from Table 1. Thematic fit ratings were entered "as is," and the verb form frequencies were entered as MC=SPast/Base, RR=PPart/Base. (For the two verbs where SPast=0, the values were entered as MC=.01 and RR=.99, instead of MC=0 and RR=1.) After searching the weight-space for this network (in steps of .05), the best approximate fit to the data was found with the SVO Bias being weighted at .45, Thematic Fit weighted at .3, and Verb Form Frequency weighted at .25. Al-

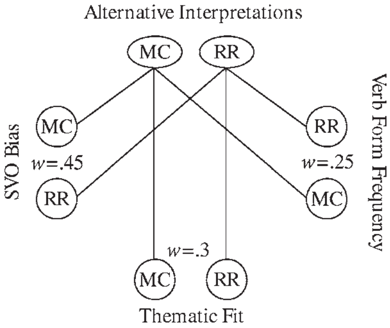


Figure 10. A schematic diagram of the Normalized Recurrence model that simulates the speeded sentence completions data.

though the weight for Verb Form Frequency is notably greater here than in McRae et al. (1998), the ordinal ranking of McRae et al's weights is preserved.

The model was given input from all 32 noun-verb pairs, and allowed to iterate for 120 cycles of competition, treating each iteration as equivalent to 10 ms of processing time. Thus, the activation of the RR Interpretation node from cycle 30 to 120 provided the model's prediction of the probability of an RR completion during the four processing-interval bins in Figure 9.

When the model's results from the 16 good Patient items were averaged, they slightly overestimated the percentage of RR completions, at around 40%. See Figure 11. Similarly, at the first processing-interval bin, the model slightly overestimated the percentage of RR completions for good Agent items as well. Notably, however, just as the temporal dynamics of the individual items canceled each other out when averaged in the human data, so did the temporal dynamics of the individual items in the model simulations cancel each other out when averaged. The model's fit to the human data when averaged across participants and items is close: r^2=.92.

Future work will need to break down these two curves into their item-by-item effects, and test the model's account of the behavior of individual sentence fragments. Since the constraint-based lexicalist framework predicts systematic item-by-item variation, the ultimate challenge for this account of sentence processing is to simulate the temporal dynamics of individual stimulus items. As the typical dataset in a sentence processing experiment contains perhaps 4–5 data points per stimulus item per condition, this new goal will require a much larger than usual dataset.

Until now, serial stage accounts of sentence processing have enjoyed the position of needing only to demonstrate effects averaged across items. How-

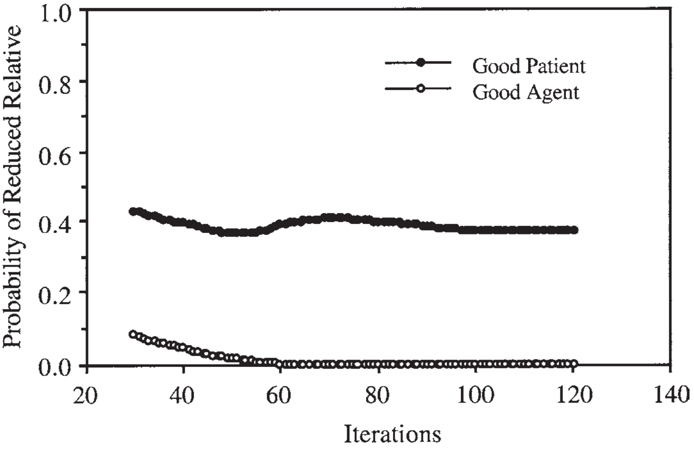


Figure 11. Results of Normalized Recurrence's simulation of the speeded sentence completion task, averaged across the 16 verbs.

ever, as these theories become more explicit in their account of how the later stages work, they too will need to make predictions about item-by-item variation, e.g., handled by a late constraint-based stage or by a rule-based reanalysis system.

General discussion

In this chapter, we have discussed the benefits of a few new tools in sentence processing, both theoretical and methodological. Nonlinear dynamics provides a new perspective for understanding the simultaneous existence of systematic, rule-like behavior in language, via nearby strong attractors, and sporadic, probabilistic behavior in language, via distant or weak attractors (cf. Tabor & Hutchins, 2000). Most dynamical models of sentence processing generally posit that all available constraints on interpretation are active simultaneously, but with varying strengths – and the results of these strength differences, as the system gravitates toward an attractor, can be quite nonlinear. Clearly, the best way to test this kind of account of language is to explore the temporal dynamics of language processing at a fine-grain scale, and look for the kinds of nonlinearities that are predicted.

In contrast to dynamical models, serial stage models of sentence processing tend to account for rule-like constraints and more probabilistic constraints with completely separate processing systems that apply their constraints at different points in time (Frazier & Fodor, 1978; Frazier, 1987; McElree & Griffith, 1995, 1998). In support of this kind of account, the results of McElree and Griffith's (1995, 1998) SAT procedure with the speeded grammaticality judgment task show what look like differential "start times" for syntactic processing, verb-subcategory processing, thematic role processing, etc. However, simulations with the Normalized Recurrence competition algorithm demonstrate that McElree and Griffith's functions of d' over time can be approximated by a model that integrates all information sources simultaneously, just with different input strengths. Essentially, this amounts to an existence proof, showing that data that might have been interpreted as consistent only with a serial stage account of sentence processing may in fact be accommodated by a parallel, integrative dynamical model of information integration.

The next step comes when this "existence proof" makes a specific prediction: that at a point of syntactic ambiguity, *early* moments of processing will show partial activation of the non-preferred alternative – and in some circumstances may even show greater activation of that alternative during early moments of processing than during later moments of processing. In order to test this prediction, a new methodology was introduced. Participants were instructed to complete sentence fragments (that were ambiguous between beginning a main clause or reduced relative clause) under varying time pressure. Results indicated that when semantic information supported the reduced relative, participants exhibited a substantial salience of the reduced relative alternative even at the earliest measured point in time. Moreover, with sentence fragments for which the constraints just barely favored the main clause, a reduced relative completion was more likely early on than later on. A simulation of Normalized Recurrence approximated these results rather well.

In sum, the evidence for serial stage models of sentence processing is waning. Many of the findings that were once treated as evidence that the influence of semantic information on parsing is delayed are being accommodated by models that apply syntactic and semantic biases simultaneously (e.g., Filip et al., this volume; McRae et al., 1998; Spivey & Tanenhaus, 1998; Tabor et al., 1997; Tanenhaus et al., 2000). Moreover, we report here suggestive evidence in the salience of syntactic alternatives for a type of temporal dynamics – early activation of the non-preferred alternative which then decreases over time – that is typically ruled out by serial stage models of sentence processing.

The goal here is not (not yet, anyway) to make it impossible to delineate what information sources are fundamental to sentence processing and what information sources are better treated as belonging to "the rest of perception and cognition." It is relatively clear that syntax is "fundamental," verb-subcategory information is "crucial," thematic role information is "pretty important," etc. As vague as those descriptors sound in distinguishing the relative import of each information source for sentence processing, so perhaps should the distinctions between the importance of these information sources in our models of sentence processing be vague. Instead of seeking evidence for discrete, qualitative architectural differences between these information sources, such as differential "start times," we advocate seeking quantitative strength differences between them, such as graded constraint weights, and a generic integration algorithm that they follow.

Acknowledgments

Supported by a Sloan Fellowship in Neuroscience to the first author. The authors are grateful to Mike Tanenhaus, Brian McElree, Daniel Richardson and one anonymous reviewer for helpful discussions, Melinda Tyler for assistance with data collection, and Jim Magnuson for help with the verb frequency data.

References

Allopenna, P., Magnuson, J. & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Chandler, J. (1969). Subroutine STEPIT – finds local minimum of smooth function of several parameters. *Behavioral Science, 14*, 81–82.

Clifton, C., Frazier, L., & Rayner (Eds.). (1994). *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

Cottrell, G. & Small, S. (1983). A connectionist scheme for modeling word sense disambiguation. *Cognition and Brain Theory, 6*, 89–120.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Filip, H., Tanenhaus, M., Carlson, G., Allopenna, P. & Blatt, J. This volume. Reduced relatives judged hard require constraint-based analyses.

Fodor, J.A., Bever, T., & Garrett, M. (1974). *The psychology of language*. New York: McGraw-Hill.

Francis, W. & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Frazier, L. & Fodor, J.D. (1978). The sausage machine: A two-stage parsing model. *Cognition, 6*, 291–325.

Frazier, L. (1987). Theories of syntactic processing. In J. Garfield (Ed.), *Modularity in knowledge representation*. Cambridge, MIT Press.

Green, D. & Swets, J. (1966). *Signal detection theory and psychophysics*. NY: Wiley.

Heeger, D. (1993). Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *Journal of Neurophysiology, 70*, 1885–1898.

Henderson, J. (1994). Connectionist syntactic parsing using temporal variable binding. *Journal of Psycholinguistic Research, 23*, 353–379.

Kawamoto, A. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language, 32*, 474–516.

MacDonald, M., Pearlmutter, N. & Seidenberg, M. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

McClelland, J. & Elman, J. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*, 1–86.

McClelland, J. & Kawamoto, A. (1986). Mechanisms of sentence processing: Assigning roles to constituents of sentences. In McClelland & Rumelhart (Eds.), *Parallel Distributed Processing, vol. 2*. Cambridge: MIT Press.

McClelland, J. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86*, 287–330.

McElree, B. & Griffith, T. (1995). Syntactic and thematic processing in sentence comprehension: Evidence for a temporal dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 134–157.

McElree, B. & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 432–460.

McRae, K., Spivey-Knowlton, M. & Tanenhaus, M. (1998). Modeling the effects of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 37*, 283–312.

Selman, B. and Hirst, G. (1985). A rule-based connectionist parsing system. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Spivey, M. & Tanenhaus, M. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1521–1543.

Spivey-Knowlton, M. & Allopenna, P. (1997). *A computational account of the integration of linguistic and visual information in spoken word recognition*. Paper presented at the Computational Psycholinguistics Conference.

Spivey-Knowlton, M. & Sedivy, J. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55*, 227–267.

St. John, M. & McClelland, J. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217–257.

Stevenson, S. (1994). Competition and recency in a hybrid network model of syntactic disambiguation. *Journal of Psycholinguistic Research, 23*, 295–322.

Swinney, D. (1979). Lexical access during sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 18*, 645–659.

Tabor, W. & Hutchins, S. (2000). Mapping the syntax/semantics coastline. In L. Gleitman & A. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 511–516. Erlbaum: Mahwah, NJ.

Tabor, W., Juliano, C. & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12*, 211–271.

Tanenhaus, M., Leiman, J. & Seidenberg, M. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior, 18*, 427–440.

Tanenhaus, M., Spivey-Knowlton, M., & Hanna, J. (2000). Modeling the effects of discourse and thematic fit in syntactic ambiguity resolution. In M. Crocker, M. Pickering, & C. Clifton (Eds.) *Architectures and Mechanisms for Language Processing*, pp. 90–118. Cambridge: Cambridge U. Press.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 268*, 1632–1634.

Trueswell, J. & Tanenhaus, M. (1994). Toward a lexicalist approach to syntactic ambiguity resolution. In C. Clifton, L. Frazier & K. Rayner (Eds.), *Perspectives on sentence processing*. Hillsdale, NJ: Erlbaum.

Waltz, D. & Pollak, J. (1985). Massively parallel parsing: A strongly interactive model of language interpretation. *Cognitive Science, 9*, 51–74.

Wiles, J. & Elman, J. (1995). Learning to count without a counter: A case study of dynamics in recurrent networks. *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

The lexical source of unexpressed participants and their role in sentence and discourse understanding*

Gail Mauner, Jean-Pierre Koenig, Alissa Melinger and Breton Bienvenue

University at Buffalo

This chapter presents preliminary evidence that bears on the issue of how unexpressed (agent) participants are represented and when they are included in the representations of agentless passive sentences using two experimental paradigms – self-paced reading and eye-monitoring. The results of our first experiment suggest that the logical necessity of an unexpressed agent in a described event is insufficient for it to be available for interpretation. Instead it must be lexically specified by a verb to be included in the representation of a sentence. The second and third experiments provide evidence that unexpressed agents are encoded when a passive verb is integrated into a sentence's representation. The latter part of this chapter presents evidence that the representation of event participant information may best be characterized as sets of fine-grained entailments, rather than as categorical primitives and that lexically specified event participants help in establishing local discourse coherence.

One of the most fundamental aspects of understanding a sentence is figuring out the "who did what to whom" component of its meaning. To do this, readers and hearers have to identify not only the events described by a sentence, but who the necessary participants of each event are and the role that each plays. Much of the time, the information needed to identify an event's participants is readily available from the explicit content of a sentence. For example, both of the sentences in example (1) introduce a kissing event with two participants, an agent and a patient. In both sentences, the agent (Wilma) and the patient (Fred) are explicitly mentioned.

(1) a. Wilma kissed Fred.
b. Fred was kissed by Wilma.

(2) Fred was kissed.

But, participant information cannot always be derived from the explicit content of a sentence. For example, the short passive sentence in (2) also describes a kissing event in which Fred is the patient. But in this case no agent is explicitly mentioned. Nevertheless, the usual understanding of this sentence is that Fred was kissed *by someone*. What this example demonstrates is that event participant information which is part of the typical understanding of a sentence cannot always be extracted from a sentence's explicit content. Clearly, it must be derived from other sources.

In this chapter, we focus on how readers' sentence representations come to include unexpressed event participant information. We examine the encoding of implicit agents in short passive sentences and two sources from which they could be derived. One source is general conceptual knowledge. Alternatively, implicit agents could be derived from schematic semantic information associated with the lexical representations of verbs (e.g., verb argument structures). The encoding of an implicit participant from either of these sources has associated with it clear processing consequences. Implicit participants are either inferred via general conceptual processing mechanisms or they are encoded when we access a verb's semantic argument structure. If, as we argue, this second alternative is correct, then a more subtle processing issue is that of *when* covert semantic argument information is used during comprehension. We provide evidence that implicit agents are derived from lexical sources and that they are accessed and rapidly used in interpreting a sentence. In the last section, we discuss the processing consequences of recent linguistic proposals that suggest that participant information is best characterized by sets of fine-grained entailments rather than categorical primitives such as agent and patient.

1. Conceptual vs. lexical encoding of event participants

Although there are many proposals for how our understanding of a sentence comes to include unexpressed participant information derived from general conceptual knowledge, we focus on a particularly influential model, proposed by Kintsch and his colleagues (e.g., Graesser, Singer, & Trabasso, 1994; Kintsch, 1988). In this model, forming a representation for a sentence is divided into three stages. Readers first construct a linguistic (i.e., syntactic) *surface structure* from the verbatim information of a sentence. From this, conceptual representations called *propositions* are constructed and added to a *text-base*. Propositions represent the gist of the information encoded in the surface structure. Finally, it is assumed that a reader's representation of a sentence or text is partly determined by her specific and generic real-world knowledge of the situation(s) evoked by a proposition (e.g., Garnham, 1981). It is at the level of a situation model that abstract conceptual knowledge, stored for instance as schemata (e.g., Rumelhart & Ortony, 1977), is included in a reader's representation of a sentence. Under this approach, the knowledge that someone kissed Fred in short passive sentences such as (2), comes from our situational knowledge of kissing events.

Linguistic theories suggest an alternative source for the unexpressed agent in our understanding of sentences like (2). It is widely assumed that the lexical representations of verbs include schematic semantic information known as argument structures, thematic roles, or case roles (e.g., Fillmore, 1968; Gruber, 1965; Jackendoff, 1990). While similar to situation models, in that they are abstract representations of event participants, argument structures and situation models differ in two crucial ways. First, while participant information that is derived from conceptual schemata is part of one's general situational knowledge, participant information encoded in verb argument structures is not *directly* derived from general knowledge sources but instead, is part of a verb's lexico-semantic representation. Furthermore, while situation models can include highly specific information about event participants (e.g., that *Wilma* kissed Fred), the participant information encoded in a verb's argument structure is less specific and more role-like. Thus, the representation of the passive verb *kissed* includes an agent or "kisser" but not a listing of individuals, such as Wilma, as potential kissers.

Argument structures have typically been associated with a verb's explicit syntactic dependents. Consider the full passive in sentence (3a). This sentence has two explicit syntactic dependents, a subject NP, *Fred*, and a prepositional phrase, *by Wilma*. These dependents correspond respectively to the patient and agent arguments in the argument structure of the passive verb *kissed*, in (3b).

(3) a. Fred was kissed by Wilma.
b. K <X, Y> (where X = PATIENT, Y = AGENT, Fred = X, and Wilma = Y)
c. Fred was kissed.
d. K <X, Y> (where X = PATIENT, Y = AGENT, Fred = X)

Some linguists have suggested that even when an argument of a verb is not associated with an explicit syntactic dependent, it might still be included in one's interpretation of a sentence (e.g., Carlson & Tanenhaus, 1988; Roeper, 1987; Williams, 1987). Under these proposals, the argument structure associated with the passive verb in sentence (3c) includes both an agent and a patient argument, even though this sentence has only one syntactic dependent. The argument structure for this sentence, given in (3d), is identical to that of the full passive sentence in (3b). The only salient difference in their semantics is that a referent for the agent argument is specified for the full passive sentence while it remains unspecified for the short passive sentence. Crucially, it is this unspecified agent that corresponds to our intuition that Fred was kissed by some unspecified individual.

Much of the evidence for the encoding of unexpressed participants has come from experiments that interrogate people's memory for what they have read (c.f., Graesser, et al., 1992 and Keenan, Potts, Golding, & Jennings, 1990 for discussion). Typically, these studies have shown that people's recollections include information from both text and background knowledge. However, subsequent work has suggested that when readers encode covert participant information, it is often done during the recollection of a text rather than during initial comprehension (c.f., Keenan, et al. for brief review and discussion). Results of studies using more on-line methods that interrogate readers' immediately-formed sentence representations also suggest that *specific* participants are rarely encoded (e.g., encoding a hammer after reading *John pounded in the nail*) (Keenan, et al.).

By contrast to earlier studies that examined whether readers encode *specific* participants from background knowledge, more recent research has examined whether readers encode unexpressed arguments of verbs (e.g., an agent) whose referential values are *unspecified* (Carlson & Tanenhaus, 1988; Roeper, 1987). Mauner, Tanenhaus, and Carlson (1995) used rationale clauses (e.g., (4d)) to probe readers' representations of full passive, short passive, and intransitive sentences such as (4a), (4b), and (4c) respectively. (Rationale clauses are infinitives whose successful interpretation depends on their understood subjects being anaphorically linked with a volitional agent introduced by an adjoining clause.) They found that make-sense judgments and reading times to rationale clauses following short passives whose verbs were hypothesized to include an agent patterned with those of explicit agent control sentences. In contrast, rationale clauses following intransitive sentences elicited anomaly effects in both judgments and reading times.

(4) a. The ship was sunk by its owners
b. The ship was sunk
c. The ship sank
d. … to collect a settlement from the insurance company.

(5) a. #The ship was sunk, but it wasn't sunk by anyone/anything.
b. The ship sank but it wasn't sunk by anyone/anything.

These results demonstrate that readers encode implicit agents in their understanding of short passive sentences. But agents may have been encoded from conceptual knowledge rather than from verb argument structures as Mauner et al. had assumed. Notice that an agent-denying clause results in a contradiction following one of Mauner et al.'s short passive sentences (e.g., (5a)) but not following one of their intransitives (e.g., (5b)). This indicates that an agent is logically required only in the short passive.

To determine whether implicit agents are derived from linguistic or conceptual sources, we have examined sensicality judgments to rationale clauses following short passive (e.g., (6a)) and intransitive sentences (e.g., (6b)) that *both* logically require an agent. (The logical necessity of an agent participant was determined in a separate study.) Since most of the agent-entailing intransitive verbs were "middle" verbs, all matrix clauses ended with a manner adverb. If the implicit agents that readers encode in their representations of short passive sentences are derived from verb argument structures, then rationale clauses (e.g., (6c)) should be difficult to process following intransitive sentences that only logically require an agent relative to when they follow short passives that both logically and linguistically require one.

(6) a. The antique vase was sold immediately
b. The antique vase had sold immediately
c. … to raise some money for the charity.

Figure 1 presents the mean cumulative percentages of "No" judgments to the first four word positions of rationale clauses. Readers found rationale clauses equally felicitous following passive and intransitive clauses at the infinitive marker *to*. Moreover, there were virtually no rejections to subsequent word positions in rationale clauses following passive clauses (2% "No" judgments at the noun word position (e.g., *money*)). Rationale clauses following intransitive clauses elicited significantly more "No" judgments than short passives at all subsequent word positions.[1] (These differences, and all others reported in this chapter, unless noted otherwise, were significant at conventional levels (i.e., $p < .05$) in analyses of variance.)

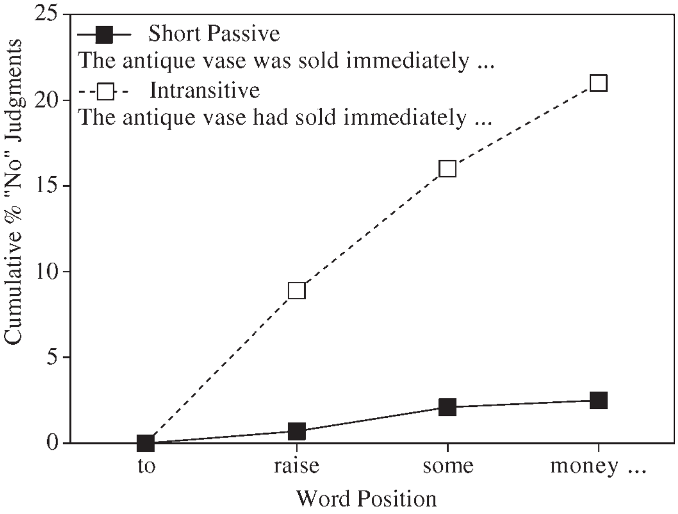


Figure 1. Cumulative percentages of "No" judgments to the first four word positions of rationale clauses following agent-entailing short passive and intransitive sentences.

The results of this study replicate Mauner et al.'s original findings, but with passive and intransitive materials equated for the logical necessity of an agent. It is therefore unlikely that earlier results were due to differences in the logical necessity of an agent. Instead, current and prior results are most plausibly interpreted as demonstrating that the unexpressed agent included in a reader's understanding of a short passive sentence is derived from the semantic argument structures associated with passive verbs, and not from more general conceptual knowledge.

2. When is participant information encoded?

While our results suggest that implicit agents, and more generally, the semantic arguments of verbs, are lexically encoded, they do not address when semantic argument information is used. This issue has played an important role in theories of sentence processing which often differ in their predictions of when semantic argument information influences processing. Of most relevance for the current discussion are studies which have provided evidence for the early influence of semantic argument information. In many, the availability of verb argument structures has been correlated with syntactic cues such as subcategory information (e.g., Trueswell, Tanenhaus, & Kello, 1993), or with the preview of an additional syntactic constituent (e.g., Tabossi, Spivey-Knowlton, McRae, & Tanenhaus, 1994). In others, argument information has been correlated with conceptual factors such as the plausibility of an NP as a filler of a given thematic role (e.g., Boland, 1997; Ferreira & Clifton, 1986; Pearlmutter & MacDonald, 1992; Trueswell, Tanenhaus, & Garnsey, 1994). Examining when implicit agents are encoded in short passive sentences may avoid some of these drawbacks because encoding can be evaluated at the verb and because it is uncorrelated with both pragmatic and syntactic cues in this construction. Although it is typical for semantic argument and subcategory information to be correlated, in most grammatical frameworks, passive verbs do *not* subcategorize for agent *by*-phrases (e.g., Bresnan, 1982; Grimshaw, 1990; Van Valin and Lapolla 1997).

We have conducted a series of experiments to examine whether readers encode implicit agents as soon as they encounter a passive verb. The logic underlying these studies is similar to that used in some filler-gap research in which readers, after encountering a clause-initial WH-filler, expect a gap with an appropriate semantic role that must be satisfied later in the clause (e.g., Boland, 1997; Clifton & Frazier, 1986; Crain & Fodor, 1985). Rationale clauses that occur in sentence-initial position are analogous to fronted WH-fillers in that their understood subjects must be associated with a volitional agent in the next clause. We used sentence-initial rationale clauses, such as (7a), to engender an "expectancy" for an agent in the linguistic representation of a subsequent clause. If our assumptions about rationale clauses are correct, comprehenders should have no difficulty processing a short passive clause such as (7b) whose verb introduces an implicit agent for the interpretation of the rationale clause. In contrast, comprehenders should experience difficulty processing an intransitive clause such as (7c), even though it describes an event in which an agent is logically required. This is because its verb does not lexically introduce an agent into the semantic representation of the clause. Crucially, if lexical argument information is used to interpret sentences as soon as a verb is recognized, then difficulty with intransitive sentences should emerge at the main verb.

(7) a. To raise money for the charity,
    b. the antique vase was sold immediately to a collector.
    c. the antique vase had sold immediately to a collector.

The first study used self-paced reading with a stops-making-sense task. We recorded "No" judgments and also reading times to sentences that were judged felicitous (i.e., responded "Yes" to) for short passive and intransitive clauses at the auxiliary verb (e.g., *had* or *was*), the main verb (e.g., *sold*), adverb (e.g.,

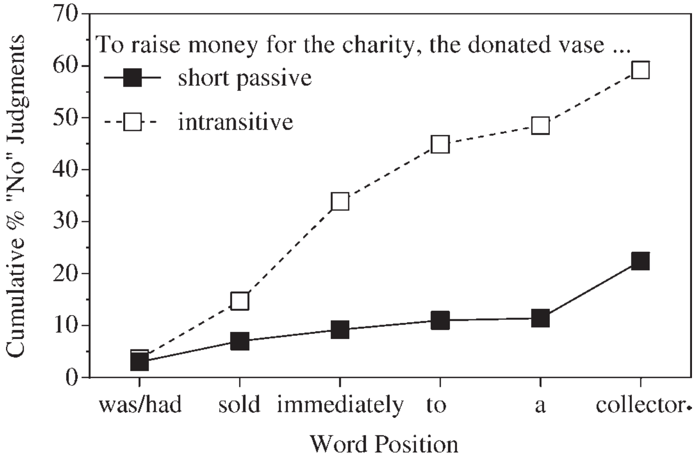


Figure 2. Cumulative percentages of "No" judgments to agent-entailing short passive and intransitive sentences following rationale clauses.

immediately), and at the three words in the sentence-final prepositional phrase (e.g., *to*, *a*, and *collector* respectively). Judgments to the critical regions of short passive and intransitive clauses are presented in Figure 2. As one can see, short passive clauses elicited practically no "No" judgments in the critical region. There was also no difference in judgments to auxiliary verbs. But, at the main verb position, intransitive clauses began to elicit significantly more "No" judgments than short passive clauses, and continued to do so through the end of the critical region.

For reading times analyses, only the auxiliary, main verb, and adverb word positions provided enough data for stable cell means. Sentence-final short passive and intransitive clauses did not differ significantly at either the auxiliary *had* or *was* (537 ms and 519 ms respectively), or main verb (e.g., *sold*) word positions (550 ms and 655 ms respectively). But, at the adverb word position (e.g., *immediately*), intransitive clauses (927 ms) elicited reliably longer reading times than short passive clauses (630 ms).

This pattern of judgments and reading times indicate that readers encode implicit agents as part of their understanding of short passive sentences as soon as they encounter a passive verb. However, the sensicality judgment task may induce readers to engage in early or additional semantic processing. Moreover, since self-paced reading times are longer when a sensicality judgment is imposed, this additional processing time may allow semantic argument information to be accessed at earlier word positions than would be the case had no judgment task been used. To address these concerns, we have examined the time course for encoding implicit agents in two eye-monitoring experiments.

Our first eye-monitoring study replicates the self-paced reading study just described. Because our two-clause sentences were too long to be presented on a single line, participants clicked a mouse button to replace rationale clauses with either a short passive or intransitive continuation. We recorded eye-movements for three regions in twelve short passive and intransitive clauses: a subject NP, a verb phrase (VP) which included an auxiliary *was* or *had* and the main verb, and a post-verb region which included adverbs and prepositional phrases. Examples of a rationale clause, and regioned short passive and intransitive continuations are provided in (8a), (8b) and (8c) respectively.

(8) a. To raise money for the charity,
b. | the antique vase | was sold | immediately to a collector.|
c. | the antique vase | had sold | immediately to a collector.|

We analyzed both unadjusted first pass and total reading times and residual first pass and total reading times, as suggested by Ferreira and Clifton (1986) and Trueswell, Tanenhaus, and Garnsey (1994). Because there were no differences in these two sets of analyses, we present only the more intuitive unadjusted reading times.

Figure 3 illustrates the mean first pass and total reading times to short passive and intransitive sentences for the three scoring regions. There were no differences in first pass reading times to short passive and intransitive sentences at either the verb or post-verb region. However, total reading times were significantly longer for intransitive than short passive sentences in the verb region and marginally so in the post-verb region. Given this data pattern, it would be reasonable to conclude that readers did not access argument structure information during their first pass through the verb regions. However, we think that this interpretation is incorrect. Readers rarely reread short passive verbs (18.6% of trials) and on average, total reading times were only 57 ms longer than first pass reading times. By contrast, readers reread intransitive verbs on 45% of trials, and total reading times were on average 232 ms longer than first pass reading times. These differences indicate that semantic argument structure information must have been processed on the first pass. Otherwise, equivalent amounts of rereading for intransitives and short passives would have been expected. One explanation for why there was no difference in first pass reading times in the verb region lies in the fact that the verb region was typically quite short. Readers may have begun programming an eye-movement to exit the verb region almost as soon they entered it. This is plausible, given that a con-

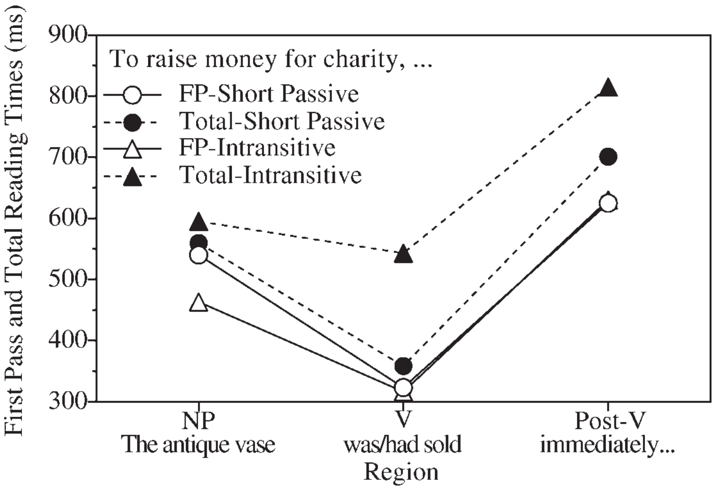


Figure 3. Mean first pass and total reading times (ms) for NP, VP, and post-VP scoring regions of short passive and intransitive sentences following rationale clauses.

servative estimate for programming an eye-movement is 150–200 ms (Matin, Shao, & Boff, 1993; Reichle, Pollatsek, Fisher, & Rayner 1998). Thus, readers may have accessed argument structure information on their first pass through a short verb region, or during the saccade exiting the region (Irwin, 1998), but realized too late to derail an eye-movement to the next region that there was no agent for the rationale clause. Support for this interpretation comes from a comparison of first pass reading times at the post-verb region to first pass reading times when first fixation times that terminated in a regression are removed. Reading times that consist only of a single fixation that terminates in a regressive eye-movement can significantly depress first pass reading times when averaged together with trials that include several fixations. This could mask potential processing difficulty or spillover effects from a previous region. This is what seems to have occurred in our data. Five of the twenty first fixations that terminated in a regression occurred in short passive sentences. Removing the 5 of 20 first fixations that terminated in a regression from passive post-verb regions increased First Pass reading times by 52 ms to 768 ms. In contrast, removing the remaining 75% of terminating first fixations from intransitive post-verb regions increased First Pass reading times from 674 ms to 787 ms. This pattern suggests that effects of argument structure information were present in first pass reading times in post-verb regions, but were masked by the high proportion of first and later fixations that terminated in a regression in intransitive sentences.

The results of this study suggest that readers interpret verb argument information at the earliest possible point, that is, while they are processing a verb. Since this study did not require readers to make any kind of judgment, it is unlikely that our earlier findings, obtained with a judgment task, were due to task demands that encouraged early semantic processing or allowed more time to access semantic argument information. Moreover, because this study used short passive and intransitive materials that were equated for the logical possibility of an agent, these results also suggest that readers are unlikely to access agent information that is conceptual rather than lexical in origin during online language processing. Finally, to the extent that the lexical representations of passive and intransitive verb participles do not subcategorize for *by*-phrases, these results represent evidence of the immediate encoding of semantic argument information that is disentangled from subcategorization information. There is nothing in the syntactic frame of a passive verb that could mediate the encoding of an agent. However, the addition of the auxiliary verb *had* in intransitive sentences, which was used to equate string length across verb phrase regions, is somewhat awkward in that it requires readers to accommodate a temporal presupposition. This could have led to more anomaly effects in intransitive sentences for reasons unrelated to differences in argument structure. Moreover, longer intransitive reading times could also have been due to difficulty in accessing the argument structures of rarer "middle" verbs relative to less rare passive verbs. We have conducted a control experiment to rule out these possibilities.

In this experiment, we examined twenty passive and intransitive sentence pairs whose intransitive forms did not require a middle interpretation (e.g., did not require a manner adverb for felicity), when preceded by either a rationale clause, as shown in (9a) and (9b), or by a control clause whose interpretation did not require an agent for interpretation, such as those shown in examples (9c) and (9d). We included a sentence-final adverb and prepositional phrase so that readers would not be forced to complete processing at the main verb.

(9) a. To raise money for the charity, | the antique vase | was sold | immediately to a collector.|

b. To raise money for the charity, | the antique vase | had sold | immediately to a collector.|

c. The detective told the museum director that | the antique vase | was sold | immediately to a collector.|

d. The detective told the museum director that | the antique vase | had sold | immediately to a collector.|

First pass and total reading times for short passive and intransitive sentences following rationale clauses are shown in Figure 4 and following control clauses plotted in Figure 5. As Figure 4 shows, there were longer total reading times at both the verb and post verb regions when intransitive rather than short passive clauses followed rationale clauses. Additionally, intransitives elicited longer first pass reading times than short passives at the post-verb region. As Figure 5 shows, there were no differences in either first pass or total reading times to short passive and intransitive sentences following clauses that did not require an agent for their interpretation. These null differences make it unlikely that anomaly effects in earlier eye-tracking or self-paced judgment studies were due to either the markedness of our intransitive verbs, or to an auxiliary verb that requires readers to accommodate a temporal presupposition. Moreover, the fact that longer reading times to intransitive sentences were obtained with non-middle verbs lessens the possibility that longer times to intransitives in the previous study arose because readers required more time to access the argument structures of rare middle verbs as compared to more frequent passive verbs. Taken together, the results of our eye-monitoring studies provide strong

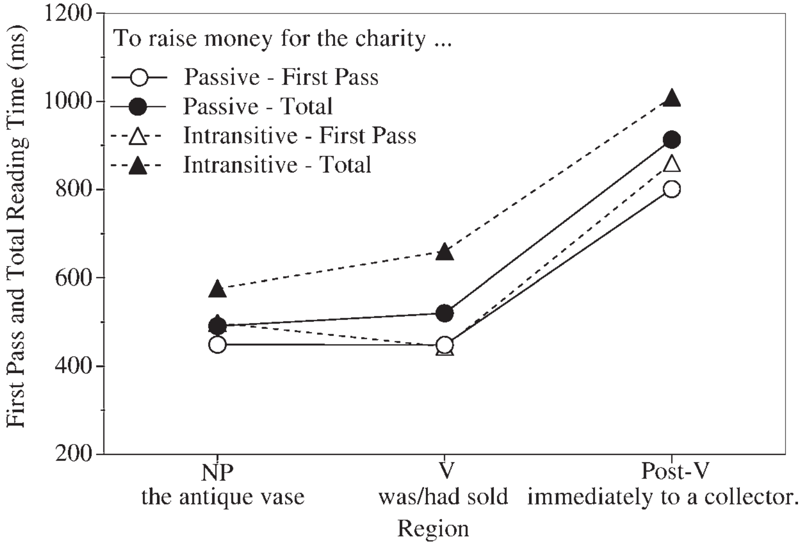


Figure 4. First pass and total reading times at three scoring regions for short passive and intransitive sentences following rationale clauses.

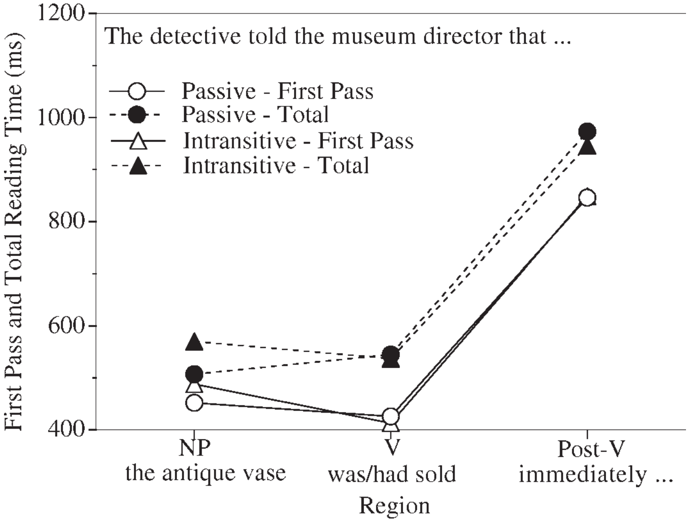


Figure 5. First pass and total reading times at three scoring regions for short passive and intransitive sentences following control clauses.

evidence for the rapid encoding of verb argument information from the lexical representations of verbs.

3. Entailment-based representations, accessibility, and discourse status

Up to this point, we have characterized lexically-encoded participant information in terms of categories such as agent and patient, etc. However, this characterization is an oversimplification that does not fully reflect how participant information is represented. As recent work in linguistics has stressed, more fine-grained distinctions are needed (e.g., Dowty, 1991; Levin, 1993). This recent work has focused on the logical entailments associated with the semantic arguments of verbs and defines thematic roles in terms of sets of entailments associated with classes of verbs. For example, Dowty (1991) has suggested that *individual thematic roles* (i.e., thematic roles that are specific to an individual verb) are clusters of lexical entailments associated with specific arguments of individual verbs. Dowty reserves the term *thematic role types* for the more general notion of thematic role that can be associated with many verbs. Thematic role types are prototypes that represent the entailments associated with an argument position across many verbs with similar clusters of entailments. Within such a representational scheme, verbs which involve an agentive participant, for example, might be defined as a subclass of verbs involving a participant that merely initiates an event (an effector in Van Valin and Wilkins' (1996) terminology). Members of the *agentive* subclass denote predicates whose participants not only initiate an event, but bear the further entailment that they willfully initiate it. If this approach is on the right track, we should expect the processing of unexpressed arguments to reflect the distinction between willful and involuntary effectors.

An example of this more fine-grained semantic distinction can be found in full passive sentences such as example (10a). Note that this sentence has an event and a state reading. On the event reading, the young woman is volitionally responsible for tormenting the priest. On the state reading, the priest is in a state of inner turmoil, and even though the young woman is the source of this turmoil, she need not be volitionally responsible for bringing this state about. One of the interesting aspects of this event-state ambiguity is that the interpretation of the "agent" argument in a *by*-phrase is subtly different depending on whether the interpretation of a full passive sentence is biased toward an event or a state. This intuitive difference can be sharpened by coupling state- and event-biased full passive sentences with a sentence-final intentional adverb (e.g., *intentionally*), as shown in (10b) and (10c) respectively. While the intentional adverb is perfectly acceptable in the event-biased sentence, it is anomalous in the state-biased sentence.[2]

(10) a. The rebel priest was tormented by the young woman.
b. #The rebel priest was profoundly tormented by the young woman intentionally.
c. The rebel priest was being tormented by the young woman intentionally.

Mauner (1996) argued that this difference in interpretation can be explained on the assumption that some thematic entailments are specific to particular classes of eventualities. This can be related to Dowty's (1991) proposals regarding entailments that are typically associated with a Proto-Agent. According to Dowty, the typical properties of agents are that they are sentient, cause events or changes of state, exist independently of the event named by a verb, and that their behavior is typically volitional or intentional. All these properties are entailed of the young woman in the sentences in (10) except intentionality, which only *necessarily* holds of the event-biased sentence (10c). What this means for the stative sentence in (10b) is that while the young woman may be interpreted as the cause of the priest's torment, she cannot be interpreted as having intentionally brought about this state. Consequently, the intentional adverb is anomalous.

If a difference in the aspectual environment of a passive sentence leads to a difference in the interpretation of an explicit agent, it may also lead to a difference in the interpretation of an implicit agent. Moreover, altering the interpretation of an implicit agent might also play a role in how accessible it is to serve as an antecedent of an implicit anaphor, such as the understood subject of a rationale clause, or an explicit anaphor, such as a pronoun, when the pronoun requires an intentional antecedent. Mauner (1996) compared the processing of rationale clauses such as (11d) following state-biased, event-biased, and unambiguously eventive short passive sentences, such as (11a), (11b), and (11c) respectively, in a self-paced reading, sensicality judgment task. If the type of eventuality introduced by biased sentences is correlated with an entailment of volition, then readers should have difficulty processing a rationale clause following state-biased (11a) but not event-biased (11b) short passive sentences.

(11) a. The rebel priest was profoundly tormented
b. The rebel priest was being tormented
c. The rebel priest was tortured
d. … to gain some information about the insurgent's hideout.

Figure 6 presents the cumulative percentages of "No" responses to the main verb in the matrix clause and the first four word positions of rationale clauses following event- and state-biased and unambiguously eventive short passive sentences. Note that at the matrix verb and at the *to* of the rationale clause, there were no differences in judgments across the three sentence types. Although the "No" judgments to unambiguous controls and event-biased short passives continued to rise in the critical region, event-biased short passives did not elicit more "No" judgments than control sentences. In contrast, "No" responses to state-biased short passive sentences began to diverge from event-biased and control sentences at the verb in the rationale clause. By the end of the scoring region, state-biased sentences elicited significantly more "No" judgments than either event-biased or control sentences. Reading times were not analyzed because there were too few data points to form stable means.

These results suggest that simple categories such as agent and patient do not completely capture the subtle semantic participant information that is computed in understanding a sentence. More fine-grained properties are used in on-line sentence comprehension. Specifically, the felicity of a rationale clause is dependent on whether an implicit agent carries an entailment of volition. When preceded by an eventive short passive that introduces a volitional agent,

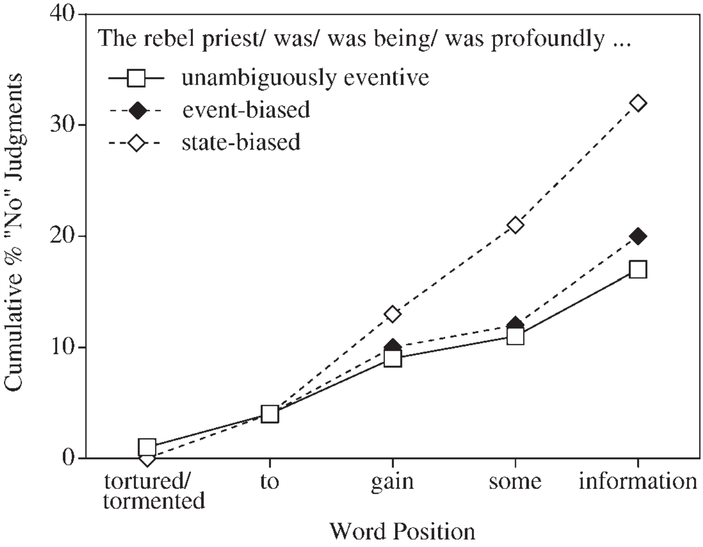


Figure 6. Cumulative percentages of "No" judgments to rationale clauses following unambiguously eventive, event-biased, and state-biased short passive main clauses.

they are easy to process. In contrast, when preceded by a stative short passive that at most introduces a nonvolitional agent,[3] they are difficult to process. There is a further interesting aspect of these results; namely, they suggest that the availability of an implicit agent to serve as the antecedent for an implicit anaphor, such as the understood subject of a rationale clause, depends in part on the kinds of entailments that can be ascribed to it. As we discuss next, the same holds for explicit anaphors.

Mauner (1996) investigated how well readers process a target sentence containing an unspecific pronominal subject (e.g. *they* or *someone*) which was the intentional agent of its own sentence (e.g., (12d)), when it follows a short passive context sentence introducing either a volitional or nonvolitional implicit agent as a likely referent for the pronoun. Participants read sentences one sentence at a time, and judged whether each target sentence made sense given its context sentence. Three kinds of context sentences were possible: state-biased short passives that introduced a non-volitional agent (e.g., (12a)), event-biased short passives (e.g., (12b)), or unambiguously eventive short passives (e.g., (12c)).

(12) a. The rebel priest was profoundly tormented for days.
b. The rebel priest was being tormented for days.

c. The rebel priest was tortured for days.
d. They wanted him to reveal where the insurgents were hiding out.
e. Was the rebel priest tortured/tormented by the ones who wanted to find out where the insurgents were hiding?

Mauner predicted that readers would find it easier to process target sentences following context sentences that provided a volitional agent to serve as an antecedent for the pronoun. Table 1 illustrates the percentages of "No" judgments and reading times for target sentences following the three types of context sentences. As one can see, targets following state-biased sentences which did not provide volitional implicit antecedents for unspecific pronouns elicited significantly more "No" judgments and longer "Yes" reading times than either type of eventive context. Targets following eventive sentences did not differ from each other in judgments or reading times.

Table 1. Mean percentages of "No" judgments, reading times and respective standard errors for target sentences following unambiguously eventive, event-biased, and state-biased short passive context sentences.

Sentence type	% "No" judgments	Reading times (ms)
Unambiguously eventive	17.9 (2.6)	3180 (183)
Event-biased	23.6 (3.1)	3300 (203)
State-biased	32 (3.5)	3484 (248)

To anticipate a potential objection, it is well known that pronouns can refer arbitrarily and are not grammatically constrained to find their antecedents in a linguistic context. For this reason, a different group of participants was asked to rate, on a five-point scale ranging from 1 ("Definitely Yes") to 5 ("Definitely No"), how probable it was that the antecedent of the target sentence's pronominal subject was the unexpressed agent of the context sentence. A sample rating study question is provided in example (12e). Ratings revealed that participants were predisposed to find the referent of the indefinite pronominal subjects of target sentences to be the implicit agents of context sentences, even when these agents were involuntary. Studies are currently under way to test a further prediction that readers will not accomodate definite pronominal subjects in the same way.

The results of these studies with ambiguous short passives show that implicit agents can serve as the antecedents to both implicit and explicit anaphoric expressions. Moreover, the anaphoric accessibility of implicit agents is affected by the kinds of entailments that are associated with them. Finally, these results demonstrate, as Carlson and Tanenhaus (1988) have argued, that implicit arguments are represented as unspecified entities in a discourse model. As such, like explicit NPs, they play a role in establishing local discourse coherence. Whether the discourse status of implicit arguments is established via the same processes that establish links between anaphors and antecedents, or instead, a link between linguistically introduced covert participants and explicitly expressed entities in a discourse model is established through coercive or accommodative processes, is part of ongoing research in our laboratory.

4. Summary and conclusions

This chapter extends the basic finding, established by Mauner et al. (1995), that readers encode implicit agents as part of their understanding of short passive sentences, in a number of directions. First, we have provided evidence that implicit agents are derived from linguistic rather than conceptual sources. This evidence suggests that theories of language comprehension and lexical representation that do not distinguish between conceptual and lexical semantic levels of representation (e.g., Graesser, Singer, & Trabasso, 1994; Kintsch, 1988) are not rich enough to capture differences in the encoding of covert participant information in people's understanding of agent-entailing short passive and intransitive sentences. We have also provided convergent evidence from self-paced reading and eye-monitoring experiments that implicit agents are encoded as soon as a passive verb is recognized. The results from eye-monitoring studies lessen the likelihood that early encoding is due to task or materials factors. Our results also provide support for the claim that implicit arguments aid in establishing local discourse coherence by introducing discourse entities that ease the integration of subsequent sentences into a discourse model. We have shown that readers more easily integrate sentences with unspecified pronominal subjects that are volitional agents of their own sentences when they are preceded by a short passive sentence that introduces a volitional implicit agent rather than an implicit effector. This is so even when readers judge implicit agents and implicit effectors to be equally probable referential candidates for the interpretation of the pronoun.

We end with some speculations regarding the potential range of lexically encoded implicit participant information. Thus far, we have focused on the syntactically most active kinds of participants: effectors and agents. One natural question that arises out of this research is: What are the boundaries of argument information within the representation of verbs? This question is particularly interesting in cases in which the linguistic and psychological evidence is mixed. Consider, for example, the instrument phrase in (13a).

(13) a. The burglar pried open the door with a piece of wood.
b. The burglar pried open the door.
c. #The burglar pried open the door, but he didn't use anything to pry it open.

It is often assumed that phrases such as *with a piece of wood* are not arguments of verbs like *pry*, but rather are adjuncts (c.f., Carlson & Tanenhaus, 1988; Speer & Clifton, 1998), as suggested in part by their omissibility. But, as we have seen with short passives, the fact that an argument does not receive overt syntactic expression is no guarantee of its absence from a verb's representation. Intuitively, even in sentences like (13b) an unexpressed instrument seems to be required in the described prying event. This intuition is confirmed by the anomaly of an instrument-denying clause in example (13c). Thus, "implicit instruments" in sentences like (13b) seem to pass a requirement on the inclusion of participant information in a verb's representation; namely, that any situation of which the verb can be predicated entails the presence of that participant.

Preliminary evidence suggests that the argument structures of verbs like *pry* may include implicit instruments. Bienvenue, Mauner, and Roehrig (1998) examined continuations for sentences like (13b), and sentences whose verb argument structures were not hypothesized to include instruments, but which could be completed with an instrument phrase (e.g., *Jordan drank a soda*). Pilot testing with similar materials regularly elicited seven semantically different continuations. We used those seven categories augmented with an "other" and an ungrammatical category to determine a chance level of responding. Sentences with verbs like *pry* elicited more instrument continuations than expected by chance as well as significantly more instrument continuations than control sentences. Similar results were obtained for sentences with hypothesized implicit goals such as *Marc drove*. A plausible explanation for these results is that verbs like *pry* and *drive* include instrument and goal participants in their respective argument structures. While these results by no means unequivocally show that instrument and goal participant information is lexically encoded in a verb's argument structure, we can rule out at least one type of conceptual information as a possible source for these continuations. With the exception of one or two items, the content of the instrument and goal phrases differed across participants. This suggests that participants were not accessing default schematic conceptual knowledge (e.g., a crowbar for sentence (13b)).

Our aim in this chapter has been to present evidence that at least some types of participant information are encoded as part of the lexical representations of verbs and are distinct from conceptually-encoded schematic knowledge of events. The work we have presented demonstrates both the need to distinguish between these sources as well as to establish how they articulate with each other in language understanding. Whether other types of unexpressed participants have similar representational sources and functions remains a challenge for future research.

Notes

* The Center for Cognitive Science of the University at Buffalo provided support for some of the work reported here. We thank David Akinbami, Wendy Baldwin, Tiffany Blasczak, Pamela Fitzgerald, Yotam Hod, Amanda Licht, and especially Patrick Hoyle for assistance in data collection and analysis, and John Trueswell, Neal Pearlmutter and two anonymous reviewers for helpful comments.

[1] There were virtually no "No" judgments to short passive main clauses (1.8%). Half of the participants also judged intransitive main clauses sensible. The other half rejected one or more intransitive main clauses (on average, 18.6% were rejected). Most of these rejections were to one item (*The new carpet installed rapidly*) on one presentation list which, when presented word-by-word, is unacceptable as an intransitive at the main verb, but becomes acceptable as a middle at the adverb. Since including the relativily high rejection rate for intransitive main clauses would have artificially inflated the initial level of "No" judgments to rationale clauses following intransitives, we have excluded main clause rejections from the cumulative percentages in Figure 1. However, these rejections contributed to the adjusted percentages that were submitted to statistical analyses.

[2] Mauner (1996) has demonstrated that aspectual cues are correlated with eventive and stative readings. For example, it is possible to reliably bias the interpretation of an ambiguous short passive towards a stative reading with degree adverbials such as *profoundly* or towards an eventive interpretation with the addition of progressive morphology (i.e., verb + *ing*). A rating study confirmed that readers interpret an ambiguous short passive such as (ia) as being more stative when modified with a degree adverbial (e.g., (ib)) and more eventive when modified with progressive morphology (e.g., (ic)).

i. a. The rebel priest was tormented.
   b. The rebel priest was profoundly tormented.
   c. The rebel priest was being tormented.

[3] State-biased short passives have often been referred to in the literature as adjectival passives. See Mauner (1996) for arguments that the predicator in state-biased passives is a verb and not an adjective.

References

Bienvenue, B., Mauner, G. & Roehrig, A. (1998, March). *Guess where and what with: The encoding of implicit instruments and locative goals*. Poster session presented at the 11th Annual CUNY conference on Human Sentence Processing. New Brunswick, NJ.

Boland, J. (1997). The relationship between syntactic and semantic processes in sentence comprehension. *Language and Cognitive Processes, 12*, 423–484.

Bresnan, J. (1982). The passive in lexical theory. In J. Bresnan (Ed.), *The mental representation of grammatical relations* (p. 3–86). Cambridge, MA: MIT Press.

Carlson, G. & Tanenhaus, M. (1988). Thematic roles and language comprehension. In W. Wilkins (Ed.), *Syntax and semantics, volume 21: Thematic relations* (p. 263–288). New York: Academic Press.

Clifton, C., Jr., & Frazier, L. (1986). The use of syntactic information in filling gaps. *Journal of Psycholinguistic Research, 15*, 209–224.

Crain, S., & Fodor, J. (1985). How can grammars help parsers? In L. Karttunen, D. Dowty, & A. Zwicky (Eds.), *Natural language processing: Psychological, computational, and theoretical perspectives* (p. 94–128). New York: Cambridge University Press.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*, 547–619.

Ferreira, F., & Clifton, C., Jr. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Fillmore, C. (1968). The case for case. In E. Bach & R. Harms (Eds.), *Universals in linguistic theory* (p. 1–87). New York: Holt, Rinehart and Winston.

Garnham, A. (1981). Mental models as representations of text. *Memory and Cognition, 9*, 560–565.

Graesser, A., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Grimshaw, J. (1990). *Argument structure*. Cambridge, Mass: MIT Press.

Gruber, J. (1965). *Studies in lexical relations*. Unpublished doctoral dissertation, MIT, Cambridge, MA.

Irwin, D. (1998). Lexical processing during saccadic eye movements. *Cognitive Psychology, 36*, 1–27.

Jackendoff, R. (1990). *Semantic structures*. Cambridge: MIT Press.

Keenan, J., Potts, G., Golding, J., & Jennings, T.M. (1990). Which elaborative inferences are drawn during reading? a question of methodologies. In D. Balota, G. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (p. 377–402). Hillsdale, NJ: Lawrence Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A constructive-integration model. *Psychological Review, 95*, 163–182.

Levin, B. (1993). *English verb classes and alternations*. Chicago: Chicago University Press.

Matin, E., Shao, K., & Boff, K. (1993). Saccadic overhead: Information processing time with and without saccades. *Perception and Psychophysics, 53*, 372–380.

Mauner, G. (1996). *The role of implicit arguments in sentence processing*. Unpublished doctoral dissertation, University of Rochester, Rochester, NY.

Mauner, G., Tanenhaus, M., & Carlson, G. (1995). Implicit arguments in sentence processing. *Journal of Memory and Language, 34*, 357–382.

Pearlmutter, N., & MacDonald, M. (1992). Plausibility effects in syntactic ambiguity resolution. In *Proceedings of the 14th annual conference of the cognitive science society* (p. 498–503). Bloomington, IN.

Reichle, E.D., Pollatsek, A., Fisher, D., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review, 105*, 125–157.

Roeper, T. (1987). Implicit arguments and the head-complement relation. *Linguistic Inquiry, 18*, 267–310.

Rumelhart, D., & Ortony, A. (1977). The representation of knowledge in memory. In R. Anderson, R. Spiro & W. Montague (Eds.), *Schooling and the acquisition of knowledge*, 99–135. Hillsdale, NJ: Lawrence Erlbaum Associates.

Speer, S.R., & Clifton, C., Jr. (1998). Plausibility and argument structure in sentence comprehension. *Memory and Cognition, 26*, 965–978.

Tabossi, P., Spivey-Knowlton, M., McRae, K., & Tanenhaus, M. (1994). Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process. In M. Moscovitch (Ed.), *Attention and performance XV* (p. 589–615). Hillsdale, NJ: Lawrence Erlbaum Associates.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 32*, 285–318.

Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preferences from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 528–553.

Van Valin, R., & Lapolla, R. (1997). *Syntax: form, meaning, and function*. Cambridge: Cambridge University Press.

Van Valin, R., & Wilkins, D. (1996). The case for 'effector': case roles, agents, and agency revisited. In M. Shibatani & S. Thompson (Eds.), *Grammatical constructions* (p. 289–322). Oxford: Oxford University Press.

Williams, E. (1987). Implicit arguments, the binding theory, and control. *Natural Language and Linguistic Theory, 5*, 151–180.

Reduced relatives judged hard require constraint-based analyses

Hana Filip*, Michael K. Tanenhaus+*, Gregory N. Carlson*, Paul D. Allopenna+ and Joshua Blatt+[1]
University of Rochester

We take as our point of departure Stevenson and Merlo's (1997) observation that the differences in the processing difficulty of sentences with reduced relative clauses (RRs) are strongly determined by the inherent lexical semantic class of the verbs used as passive participles in RRs: namely, the unaccusative vs. unergative class. Our main claim is that among the linguistic variables responsible for the relevant differences a crucial role is played by semantic variables, rather than just category-level syntactic complexity and/or complexity associated with word-internal lexical structure of verbs (see Hale and Keyser, 1993). First, we observe a considerable overlap in the distributions of acceptability judgments between sentences with RRs based on unaccusative verbs and those based on unergative verbs, and even more importantly, clear gradient effects with respect to acceptability judgments for both types of sentences that are influenced by the lexical semantics of the main verb in the matrix clause. Second, such data can be successfully motivated, if we characterize the crucial unaccusative-unergative distinction in terms of thematic Proto-Role properties (Dowty, 1988, 1991). Third, the linguistic analysis is consistent with recent constraint-based grammars, most notably HPSG, and our constraint-based model that uses the integration-competition architecture developed by Spivey (1996) and applied to reduced relatives by McRae et al. (1998) and Spivey and Tanenhaus (1998).

1. Introduction

Beginning with Bever's (1970) classic article, sentences with reduced relative clauses, such as *The horse raced past the barn fell*, have served as an important empirical testing ground for evaluating models of sentence processing. Bever observed that sentences with reduced relative clauses are difficult to understand, with people often judging the sentences to be unacceptable, because they initially assume that the 'NP V PP' sequence is a main clause. In subsequent decades one of the central controversies revolved around the question of whether structural complexity plays a primary causal role in processing difficulty of sentences with reduced relative clauses, and other sentences with temporary ambiguities. For example, in recent constraint-based models, the difficulty of reduced relative clauses is argued to arise from an interaction of multiple constraints, many of which are lexically-based (e.g., MacDonald, Pearlmutter and Seidenberg, 1994; Tanenhaus and Trueswell, 1995; Boland, 1997), but which do not include any factors directly attributable to intrinsic ease or difficulty of processing syntactic structures. Important empirical evidence in support of constraint-based approaches has come from gradient effects in the processing difficulty of reduced relatives. For example, *The eggs cooked in butter tasted delicious* is clearly much easier to process than *The horse raced past the barn fell*. In constrast to *raced*, *cooked* is much more often used transitively, it is more frequently used as a passive, and *eggs* is a very poor Agent, but a very good Theme, in a cooking event. Due to such constraints the active intransitive reading of *The eggs cooked with butter…* is less likely and the passive participle reading more likely. Gradient effects in processing difficulty for reduced relative clauses have been successfully modeled using computational implementations of multiple constraint models (McRae, Spivey-Knowlton and Tanenhaus, 1998; Spivey and Tanenhaus, 1998).

Recently, Stevenson and Merlo (1997) made the important observation that the processing difficulty of sentences with reduced relative clauses is strongly determined by the inherent lexical class of the verbs used as passive participles in reduced relatives. Sentences with reduced relatives headed by passive participles derived from unergative[2] verbs are "all mostly or completely unacceptable" (p. 355). In particular, manner of motion verbs "lead to a severe garden path in the RR construction" (p. 353), as is shown in Stevenson and Merlo's (p. 353) examples, here repeated in (1). In contrast, "unaccusative RRs are all completely acceptable or only slightly degraded" (p. 355). Stevenson and Merlo's examples are repeated here in (2):

(1) a. The clipper sailed to Portugal carried a crew of eight.
b. The troops marched across the fields all day resented the general.
c. The model planet rotated on the metal axis fell off the stand.
d. The dog walked in the park was having a good time.

(2) a. The witch melted in the Wizard of Oz was played by a famous actress.
b. The genes mutated in the experiment were used in a vaccine.

c. The oil poured across the road made driving treacherous.
d. The picture rotated 90 degrees was easy to print.

Stevenson and Merlo propose that the unergative/unaccusative difference can be explained using Hale and Keyser's (1993) syntax-in-the-lexicon model, couched within Government and Binding Theory, in which important aspects of lexical-conceptual structure are mirrored by syntactic structures within the lexicon. Unergative verbs are syntactically characterized (among other things) by having an external argument, but no direct internal argument, while unaccusative verbs have no external argument, and a direct (non-clausal, non-PP) internal argument. Due to such lexical properties, transitive and passive structures, including those in reduced relative clauses, which are derived from inherently unergative verbs are significantly more complex than those derived from unaccusative verbs "in terms of number of nodes and number of binding relations, and in having the embedded complement structure" (Stevenson and Merlo, 1997: 364). When these linguistic assumptions are implemented in Stevenson's (1994a, b) competitive attachment parser, a kind of symbolic/connectionist hybrid, it turns out that the parser cannot activate the structure needed for a grammatical analysis of reduced relatives headed by passive participles with unergative verbs, "because of its limited ability to project empty nodes and to bind them in the structure" (Stevenson and Merlo, 1997: 397). Hence, the parser is viewed as confirming the earlier judgment data, namely that there are "sharp distinctions between unergative RR clauses and RR clauses with other verbs" (p. 396).

In contrast to previous structural theories which attribute the difficulty of reduced relatives solely to category-level syntactic complexity differences, Stevenson and Merlo propose that lexical constraints play a central role in determining the processing difficulty of reduced relative clauses. However, in contrast to constraint-based models, they argue that differences among classes of lexical items are due to differences in structural complexity associated with their lexical structures. They argue that reduced relatives with participles based on unergative verbs are uniformly difficult to process, regardless of factors such as frequency and plausibility, that is, "structural complexity alone can cause failure to interpret a sentence, even when all other factors would help its correct interpretation" (Stevenson and Merlo, 1997: 392).

If correct, Stevenson and Merlo's claims would have a number of important implications for theories of sentence processing. First, they would provide the clearest evidence to date for structural complexity effects in sentence processing, due to the internal syntactic structure of words, thus helping to resolve a long-standing controversy in the field. Second, since there are both syntactic and semantic[3] aspects of the unergative/unaccusative distinction, Stevenson and Merlo's results would strongly support an approach in which syntactic correlates of semantic distinctions play the primary causal role in accounting for variation in processing difficulty.

In this chapter we evaluate Stevenson and Merlo's claims in light of additional empirical data and modeling within a constraint-based framework. Section 2 presents the results of a questionnaire study which replicates Stevenson and Merlo's finding that reduced relatives with passive participles derived from unergative verbs are, as a class, more difficult than reduced relatives with passive participles based on unaccusative verbs. However, the results also show that there is a considerable overlap in the distributions of acceptability judgments and parsing difficulty, as would be expected on a constraint-based account. Section 3 shows that the processing difficulty that is due to the unergative/unaccusative difference falls out of a computational implementation of a constraint-based model, using only those constraints that recent constraint-based theorists have claimed account for processing differences among reduced relatives (MacDonald et al., 1994; Tanenhaus and Trueswell, 1995). Thus the unergative/unaccusative difference does not require appeal to structural complexity differences. In section 4, we argue that a semantic approach based on thematic roles presents a promising alternative to the syntax-in-the-lexicon approach. The thematic properties, which characterize the two fuzzy cluster concepts Proto-Agent and Proto-Patient (Dowty, 1988, 1991), can account for a great deal of processing differences between sentences with reduced relative clauses based on unergative verbs, on the one hand, and on unaccusative verbs, on the other hand. One advantage of this novel way of looking at the garden-path phenomenon is that it allows us to understand the similarities between these two types of sentences in exhibiting clear gradient effects with respect to acceptability judgments and parsing difficulty that are influenced by the lexical semantics of the main verb in the matrix clause. The influence of the main predicate in a sentence on the magnitude of the garden-path effect has so far gone unnoticed and it is problematic for structure-based accounts that assume either category-level syntactic complexity and/or complexity associated with word-internal lexical structure of verbs. We also show that a constraint-based approach incorporating these semantic notions can be naturally embedded within recent constraint-based approaches to grammatical representation. We conclude by describing a rating study that shows the effect of the main verb on the processing difficulty of whole sentences with reduced relative clauses, as is predicted by our linguistic analysis.

2. Gradient effects

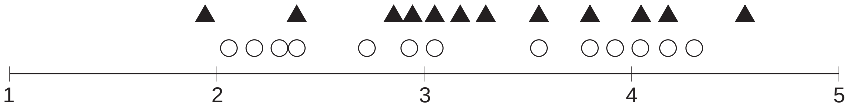
We observed that sentences with reduced relatives based on unergative verbs, including manner of motion of verbs, manifest a considerable degree of variability in acceptability, and, in fact, perfectly acceptable sentences of this type are easy to find. Examples are given in (3). At the same time, some reduced relatives with unaccusative verbs are relatively hard, such as those in (4).

(3) a. The victims rushed to the emergency room died shortly after arrival.
b. The pig rolled in the mud was very happy.
c. The Great Dane walked in the park was wearing a choke collar.
d. The prisoners paraded past the mob were later executed.[4]

(4) a. The theatre darkened for the movie frightened some preschoolers.
b. The Klingon disintegrated during the battle had launched a rocket.
c. The solution crystallized in the oven burned a hole into the petri dish.
d. The plaster hardened in the oven cracked with loud popping sounds.

In a questionnaire study we had twenty-four University of Rochester undergraduates recruited in introductory courses use a five point scale (1 = very easy, 5 = very difficult) to rate the difficulty of a mix of sentences that included reduced relative clauses with inherently unaccusative and unergative verbs, as well as transitive and passive main clause sentences using the same verbs. The full set of materials used in the rating studies are available by request from either of the first two authors. Table 1 presents the mean ratings. There was a significant effect of construction type, F1(1,23)=62.00, p<.01; F2(2,32)=82.02, p<.01. Reduced relatives were significantly harder than passives or transitives, regardless of verb type (all planned comparisons were significant at p<.01). We replicated Stevenson and Merlo's finding that sentences with reduced relatives headed by passive participles based on unergative verbs are harder to process than sentences with reduced relatives headed by participles derived from unaccusative verbs. For reduced relatives with passive participles derived from unaccusative verbs, the mean was 2.95; and for those with unergative verbs, the mean was 3.45. This difference was reliable in the analysis by subjects, F(1,23)=5.51, p<.05. However, there was substantial overlap in the distributions, and in fact the difference between the unaccusatives and unergatives was only marginally reliable in an item analysis, F(1,32)=3.15, p=.085. Four of the eighteen unergative verbs used as passive participles in reduced relatives were rated as yielding sentences with reduced relatives judged easier than the mean rating for sentences with reduced relatives based on unaccusative verbs. The sentences with these verbs are in (3) above. In addition, some sentences with

Table 1. Judged difficulty of reduced relatives



Class	RR	Trans	Pass
○ Unaccusatives	2.95	1.63	1.60
▲ Unergatives	3.45	1.52	1.81

reduced relatives headed by passive participles derived from unaccusative verbs were rated as more difficult than the mean rating for sentences with unergative-based reduced relatives (3.45). Six of the sixteen unaccusative verbs fell into this category, including the sentences in (4).

To summarize, the ratings showed that sentences with unergative-based reduced relatives were on the whole more difficult to process than sentences with unaccusative-based reduced relatives, but also that there was a considerable degree of overlap between these two types of sentences with respect to the processing difficulty. The overlap in the distributions and the continuum of difficulty is problematic for an account in which the inherent structural complexity of unergative verbs predicts "sharp distinctions between unergative RR clauses and RR clauses with other verbs" (Stevenson and Merlo, 1997: 396). They do not, however, provide definitive evidence against such a proposal, however, because measurement error or other differences among materials could lead to overlap in the data even if the underlying distributions did not overlap.

3. A constraint-based model

We implemented a constraint-based model using the integration-competition architecture developed by Michael Spivey and applied to reduced relatives by McRae et al. (1998) and Spivey and Tanenhaus (1998). In this model alternative syntactic structures compete within a probability space with multiple constraints providing probabilistic evidence for the alternatives. This model is not a fully implemented parser; rather, it is an architecture for predicting the difficulty of ambiguity resolution using principles common to constraint-based approaches. The question we addressed was whether an unergative/unaccusative

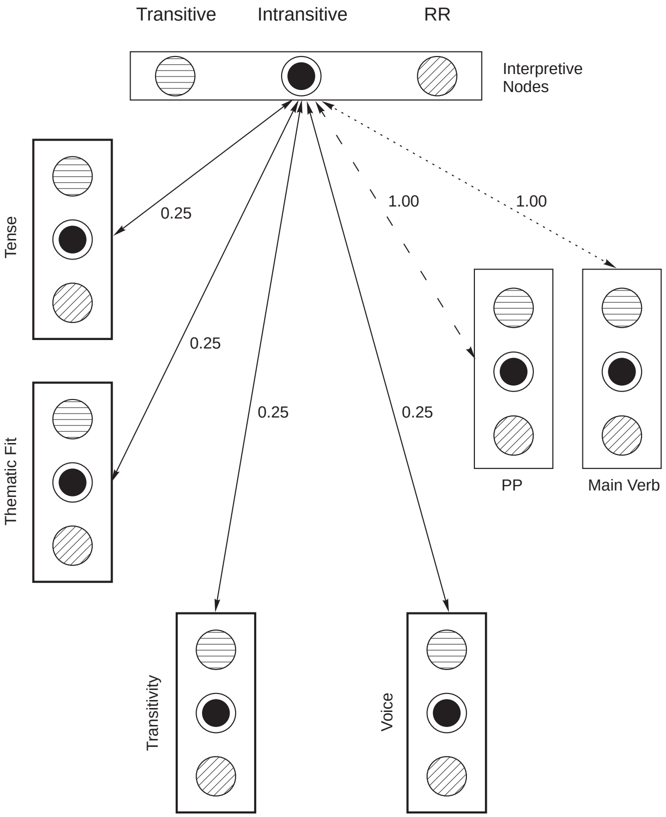


Figure 1. The Integration and Competition model used in the current simulations. Each vertical rectangle represents the input of a particular constraint. The horizontal rectangal represents the output of the model at any point in time for the three integration nodes for the embedded clause: the active transitive, active intransitive and passive in a reduced relative. Constraints for tense, voice, transitivity and thematic fit were introduced at the embedded verb. The PP constraint was introduced after the model completed cycles (i.e., until the dynamic criterion was reached) for each input at the embedded phrase. Similarly, the main verb constraint was introduced after processing was completed for the PP. The weights shown are those that were used when a constraint was first introduced, i.e., before normalization at the next input.

difference would fall out of such a model using just those constraints that have been previously identified in the constraint-based literature.

Figure 1 presents a schematic representation of the model. In the model, three constructions competed, beginning with the first verb in a sentence with a reduced relative clause: NP V(*-ed*) PP V. The constructions were: active transitive, active intransitive, and passive in a reduced relative. The full passive was ruled out at the *-ed* verb form because of the absence of a preceding copula, and thus was not included.

The constraints used were those identified by MacDonald and colleagues (e.g., MacDonald et al. (1994)) and by Tanenhaus and his colleagues (e.g., Tanenhaus and Trueswell, 1995). The following four constraints came into play at the *-ed* verb form: (1) The frequency with which a verb was used transitively or intransitively; (2) the frequency with which it was used in tensed vs. tenseless constructions; (3) the frequency with which the *-ed* verb form was used in the passive and active voice, and (4) the plausibility with which the first NP could function as subject of an active transitive, subject of an intransitive, and subject of a passive ("thematic fit"). An additional frequency constraint came into play at the PP, and another at the main verb.

In the integration-competition model, each constraint provides probabilistic support for the syntactic alternatives. The normalized bias on the constraint is multiplied by the weight assigned to the constraint. The weights of all the constraints applying at a given input are normalized so that they sum to 1.0. The model works in three steps. First the biases are multiplied by the weights to determine the evidence (activation) each provides in support of the competing interpretation (integration) nodes. Activations are summed at each integration node. Second, feedback to the constraints is provided by multiplying the probability of each integration node by its weight and adding that value to its previous bias. Third, the biases for each constraint are then renormalized. The model continues cycling until a designated criterion; the criterion is lowered after each cycle. (For details, see McRae et al., 1998; Spivey and Tanenhaus, 1998.) When the criterion is reached, the model moves onto the next region of the text, in this case the PP. The new constraint provided at the PP, namely, strong evidence for either an intransitive or a passive, was assigned a weight of 1.0, following the procedure used in McRae et al. (1998). All of the weights were then renormalized, resulting in a weight of .5 for the PP and .125 for tense, voice, thematic fit and transitivity. The same procedure for normalizing weights was followed when the model moved on to the main verb.

Because we did not have an independently motivated way of setting the weights on the four constraints at the *-ed* verb form, we assigned each an equal weight of .25. Biases for transitivity, tense, and voice were determined from corpus analyses using the ACL/DCI corpus, comprising the Brown corpus and 64 million words of the Wall Street Journal that were kindly provided to us by Paola Merlo and Suzanne Stevenson. The biases for thematic fit were determined by typicality ratings collected using the procedure developed by

McRae and colleagues (cf. McRae et al., 1998). Ratings were collected using a five point scale. Questions we used are here exemplified using the verb *melt* as an example: 'How common is it for ice to melt someone or something?' (Active Transitive), 'How common is it for ice to melt?' (Active intransitive), 'How common is it for ice to be melted by someone or something?' (Passive in RR). We tested the model on six unergative verbs, *danced*, *raced, paraded, rushed, marched*, *hurried*, and on four unaccusative verbs, *dissolved, cracked, hardened* and *melted*. This small subset of verbs represents those for which we had corpus counts, difficulty ratings and ratings for thematic fit. Table 2 presents the biases used in the model for each of the four constraints that applied at the *-ed* verb form.

Table 2.

Word	Constraint	Bias		
		Transitive	Intransitive	RR
Cracked	Tense	0.31	0.31	0.38
	Thematic Fit	0.12	0.38	0.50
	Transitivity	0.37	0.45	0.18
	Voice	0.41	0.41	0.19
Danced	Tense	0.40	0 40	0.21
	Thematic Fit	0.21	0.56	0.24
	Transitivity	0.15	0.77	0.08
	Voice	0.43	0.43	0.14
Dissolved	Tense	0.16	0.16	0.68
	Thematic Fit	0.17	0.43	0.41
	Transitivity	0.50	0.25	0.25
	Voice	0.21	0.21	0.58
Hardened	Tense	0.05	0.05	0.91
	Thematic Fit	0.21	0.49	0.30
	Transitivity	0.43	0.36	0.21
	Voice	0.23	0.23	0.55
Hurried	Tense	0.32	0.32	0.37
	Thematic Fit	0.31	0.35	0.34
	Transitivity	0.39	0.42	0.19
	Voice	0.34	0.34	0.31
Marched	Tense	0.45	0.45	0.09
	Thematic Fit	0.22	0.43	0.35
	Transitivity	0.06	0.91	0.03
	Voice	0.49	0.49	0.01
Melted	Tense	0.15	0.15	0.71

Table 2. *(continued)*

Word	Constraint	Bias		
		Transitive	Intransitive	RR
Jewelry	Thematic Fit	0.16	0.31	0.53
	Transitivity	0.34	0.49	0.17
	Voice	0.28	0.28	0.44
Melted	Tense	0.15	0.15	0.71
Witch	Thematic Fit	0.35	0.32	0.33
	Transitivity	0.34	0.49	0.17
	Voice	0.28	0.28	0.44
Paraded	Tense	0.25	0.25	0.50
	Thematic Fit	0.26	0.28	0.46
	Transitivity	0.30	0.55	0.15
	Voice	0.31	0.31	0.39
Raced	Tense	0.50	0.50	0.01
	Thematic Fit	0.10	0.45	0.45
	Transitivity	0.05	0.93	0.02
	Voice	0.50	0.50	0.01
Rushed	Tense	0.40	0.40	0.20
	Thematic Fit	0.26	0.33	0.41
	Transitivity	0.14	0.80	0.07
	Voice	0.44	0.44	0.12

As can be seen from Table 2, unergative verbs tend to be used more often than unaccusative verbs in intransitive constructions and less often as passives. For unergative verbs these factors mean that the active intransitive reading of an 'NP V(-*ed*) PP' fragment will be more strongly biased relative to the reduced relative clause reading.

In order to evaluate the output of the model, we considered three measures. The first was the total number of cycles until the criterion was reached at the main verb (cycles at the -*ed* verb form, + cycles at the PP, + cycles at the main verb). The second was the probability assigned to the reduced relative structure at the main verb. The third was the number of cycles it would take the model to assign the reduced relative a probability of .9 at the main verb. We assumed that each of these measures should correlate with the difficulty of the sentence. All three measures predicted that as a class reduced relatives with passive participles derived from unergative verbs would be more difficult than reduced relatives with passive participles derived from unacccusative verbs: for total number of cycles, t(9)=3.16, p<.01; for probability at the main verb t(9)=2.95, p<.02; and for cycles to a criterion of .9, t(9)=2.99, p<.02, all tests two-tailed. The model also correctly predicted some gradient effects. For example, the reduced relative beginning with *The witch melted…* was correctly predicted to be harder than the reduced relative beginning with *The jewelry melted…* . In addition, the reduced relative with *paraded* was predicted to be easier than the reduced relatives with *danced, raced* or *marched*. However, *The victims rushed to the hospital died* was incorrectly predicted to be quite difficult even though it was rated as fairly easy by subjects.

It is important to note that the model we presented is incomplete in important ways. There are constraints that are not included and as a result the model generally overestimates the availability of the reduced relative analysis. Moreover, we were working with only a few verbs for which we had data. Nonetheless, it is clear that the processing distinction between reduced relatives headed by passive participles derived from unergatives and unaccusatives falls out of a small set of constraints, primarily verb-based frequencies, that have been independently argued for by proponents of constraint-based models.

In the light of the results we reached so far, a proponent of the syntax-in-the-lexicon approach might appeal to two types of counterarguments. The first might be that frequencies reflect the unergative/unaccusative distinction; however, the structural complexity associated with the lexical structures of these two classes of verbs results in those frequencies and actually plays the causal role (but cf. MacDonald, 1997). The second argument is that the syntax-in-the-lexicon approach implemented in Stevenson's parser is superior because it presupposes a full-fledged linguistic theory, namely, Government and Binding Theory, whereas the constraint-based approach is not supported by independent linguistic assumptions in a similar way. In the next two sections we address these issues in turn. First, we explore and motivate the claim that among the linguistic variables responsible for the processing distinction a crucial role is played by semantic variables, rather than just syntactic variables. Second, we show that the ideas implemented within our simple model are broadly consistent with recent constraint-based grammars, most notably HPSG.

4. The linguistic basis of unaccusative/unergative distinction in processing

Our primary observation, and one that has so far gone unnoticed, is that both types of sentences with reduced relatives exhibit similar gradient effects in acceptability judgments that are crucially influenced by the lexical semantics of the main verb in a matrix clause. To put it in the simplest terms, the fewer agent-like properties and the more patient-like properties the main verb assigns to its subject, the easier the whole sentence with a reduced relative clause is judged. This idea will be discussed in detail in Section 4.2, but let us illustrate it here with a few examples. In (5a) the subject of *complained*, *the patients*, is a volitional agent in the denoted event, and we see that the whole sentence is less acceptable than (5b) with *died* as the main verb, whose subject undergoes a change of state. A similar contrast can be found in (6):

(5) a. The patients rushed to the emergency room *#complained to the nurse.*
b. The patients rushed to the emergency room *died.*

(6) a. The Great Dane walked in the park *#tugged at the leash.*
b. The Great Dane walked in the park *wore a choke collar.*

Similarly in reduced relatives with passive participles derived from unaccusative verbs, such as *darkened* in (7), we see that the use of *frightened* as opposed to *smelled* in the matrix clause is correlated with a difference in the acceptability of the whole sentence. The reason is that *frightened*, but not *smelled*, presents the subject *the theatre* as the cause of the change of the psychological state in the referent of the direct object *some preschoolers.* Other similar examples are given in (8):

(7) a. The theatre darkened for the movie *#frightened* some preschoolers.
b. The theatre darkened for the movie *smelled* like popcorn.

(8) a. The genes mutated in the experiment *#attacked* their host.
b. The genes mutated in the experiment *were used* in a new vaccine.

Most importantly, different degrees of acceptability observed in (5)–(8) resist an explanation in structure-based terms as well as explanations couched in the syntax-in-the-lexicon approach of Stevenson and Merlo (1997). Recall that the latter predict that *all* sentences with reduced relatives headed by inherently unergative verbs are predicted to pose 'sharp difficulty' (p. 392) for an interpreter, and they cannot be assigned a grammatical analysis by the parser. In order to account for unaccusative-based reduced relatives that are *not* easy to interpret, such as those in (9), Stevenson and Merlo resort to the semantic distinction between 'internal causation' and 'external causation' (see Levin and Rappaport Hovav, 1995: 210–211) to argue that they are unergative. According to them, verbs like *caramelise, solidify* and *yellow* entail 'internal causation' in their semantic description, a feature that distinguishes unergative verbs from unaccusative ones, the latter being 'externally caused' (see ibid.). Since unaccusative verbs have one internal direct object argument, the external subject argument position is unfilled, and it can be filled by an ‘external cause’ argument, when they are used transitively. This does not hold for unergative verbs, because they already have one external subject argument. By this test, *yellow* in (10a) and *solidify* in (10b) are unergative, while *harden* in (10c) and *yellow* in (10d) are unaccusative. (Examples in (9) and (10) are taken from Stevenson and Merlo, 1997: 365.)

(9) a. #The candy caramelised in an hour burned.
b. #The wax solidified into abstract shapes melted.
c. #The paper yellowed in the sun shrank.

(10) a. #The chain-smoker yellowed the papers.
b. #The sculptor solidified the wax.
c. The sculptor hardened the wax.
d. The sun yellowed the paper.

The problem with this test is that unergative verbs, including agentive manner of motion verbs, when used transitively *require* their subject argument to be an Agent: cp. **The explosion jumped the horse* vs. *The jockey jumped the horse.* (This observation was made by Cruse, 1972; Jackendoff, 1972; Levin and Rappaport Hovav, 1995; see also Stevenson and Merlo, 1997: 357 and footnote 4 below.) This inconsistency clearly indicates that a test based on the possibility of the overt expression of an Agent argument cannot be the right diagnostic for deciding the membership of verbs in the unaccusative and unergative class. The main source of confusion stems here from correlating ‘external causation’ and ‘possibility of an overt expression of an external agent’, on the one hand, and ‘internal causation’ and ‘prohibition against an overt expression of an external agent’, on the other hand. What is lacking is a precise characterization of the notions ‘internal causation’ and ‘external causation’, introduced by Levin and Rappaport Hovav (1995), and the motivation for the correlation of these semantic notions with the syntactic structures associated with unergative and unaccusative verbs. Moreover, (10a) is claimed to be less acceptable than (10d), because its subject referent may be intentionally involved in the denoted event, while in (10d) the denoted change of state is “indirectly brought about by some natural force” (p. 365). However, it is not shown how such a fine-grained distinction between ‘(volitional) Agent’ and ‘natural force’, and the suggested difference in acceptability judgments, can be viewed as being correlated with the external subject argument in the case of unergative verbs, and with the internal object argument in the case of unaccusative verbs.

The fact that Stevenson and Merlo do resort to rather subtle semantic criteria in order to account for difficult cases is instructive, because it shows that explanations in terms of categorical differences between syntactic configurations in the lexicon are insufficient. Indeed, one may ask to what extent syntactic factors are necessary in addition to semantic ones in order to account for the garden-path phenomenon. If we focus on the differential semantics of the verbs in the material discussed here, we can begin to acount for the overlapping distribution of sentences with reduced relatives as well as the great deal of variability with respect to how good or bad they are judged to be, leaving open the question of what role, if any, a word-internal syntactic differences are left to play. We now turn to characterizing those semantic constraints more precisely.

4.1 Thematic Proto-Roles

The idea that argument positions of verbs are associated with certain "thematic roles" (Case Roles, Case Relations) such as Agent, Patient, Instrument, and so forth, has received varying characterizations in the linguistic literature. Here, however, we follow the analysis of David Dowty (1988, 1991), who proposes that the only thematic roles are two cluster concepts, Proto-Agent and Proto-Patient, each characterized by a set of verbal entailments, given in (11) (see Dowty, 1991: 572). "[A]n argument of a verb may bear either of the two proto-roles (or both) to varying degrees, according to the number of entailments of each kind the verb gives it" (Dowty, 1991: 547).

(11) Contributing properties for the Agent Proto-Role:

a. volitional involvement in the event or state
b. sentience (and/or perception)
c. causing an event or change of state in another participant
d. movement (relative to the position of another participant)
(e. referent exists independent of action of verb)

Contributing properties for the Patient Proto-Role:

a. undergoes change of state
b. incremental theme
c. causally affected by another participant
d. stationary relative to movement of another participant
(e. does not exist independently of the event, or not at all)

The Argument Selection Principle determines the direct association of clusters of Proto-Agent and Proto-Patient properties with grammatical relations in a many-to-one fashion:

(12) Argument Selection Principle (Dowty 1991: 576)

In predicates with grammatical subject and object, the argument for which the predicate entails the greatest number of Proto-Agent properties will be lexicalized as the subject of the predicate; the argument having the greatest number of Proto-Patient properties will be lexicalized as the direct object.

4.2 Compatibility between subjects in sentences with reduced relative clauses

In reviewing the contrasts found in examples, such as (5)–(8), it appears that the following is a reasonable description of one effect of the main verb on a reduced relative clause:

(13) Hypothesis

The acceptability of sentences with reduced relative clauses, headed by passive participles derived from unergative and unaccusative verbs, increases when the passive participle and the main verb of a matrix clause assign their subject-NPs more Proto-Patient, and fewer Proto-Agent, properties.

The intuition behind the hypothesis (13) is that sentences are easier to interpret when there is an internal coherence among the interpretations of their constituents. One way this coherence can be achieved is in terms of compatible assignments of thematic properties to different NP arguments that are associated with one and the same participant in the domain of discourse. In sentences with a reduced relative clause the internal coherence depends in part on how well the thematic make up of the subject NP in the matrix clause matches the thematic make up of the PRO-subject of the reduced relative clause: namely, the passive participle in the reduced relative requires that its PRO subject be a "very good" Patient. Let us take (1a) #*The horse raced past the barn fell.* At the point when *raced* is processed, the preferred syntactic-semantic pattern is that of the main clause with an agentive subject-NP. However, when *fell* is processed, *raced* must be understood instead as a passive participle. Passive participles typically presuppose the existence of corresponding active transitive verbs whose subjects correspond to active direct objects (see Sag and Wasow, 1997: 164, for example; however, passive subjects do not always correspond to active direct objects, see Zwicky, 1987; Postal, 1986, and others). Let us now look at the assignment of thematic properties by the verb *raced* in its intransitive (unergative) and transitive (lexical causative) use. ('PA' stands for Proto-Agent properties and 'PP' for Proto-Patient ones.)

(14) *The horse RACED past the barn. The rider RACED the horse past the barn.*

PA	PA	PA	and	PP
(+ volition)	+ volition	(+ volition)		+ **causally**
+ sentience	+ sentience	+ sentience		**affected**
+ movement	+ **causing change**	+ movement		

A causative form of an unergative is not a "usual" transitive in that it semantically departs from prototypical transitives. Intuitively, prototypical transitives can be understood in terms of a 'billiard ball model', as Langacker (1986) calls it, which involves two participants that interact in an asymmetric and unidirectional way, whereby one of them is directly affected by some action (possibly involving movement, contact, effect, and the like) instigated or caused by the other participant. In Dowty's terms, this means that the direct object has many Proto-Patient (and a few Proto-Agent) properties, and the subject has many Proto-Agent (and a few Proto-Patient) properties. A typical unergative verb used transitively does not fit the semantics of a transitive prototype, because its direct object has a thematic make up of a "good" Agent: in our example (14) the subject *the horse* of the intransitive *raced* corresponds to the object of the transitive *raced* and they share three Proto-Agent properties. At the same, *the horse* is assigned one Proto-Patient property 'causally affected' by the transitive *raced*. The awkwardness often related to the transitive use of unergative verbs may be seen as stemming from having to reconcile these two different roles or two different perspectives (an Agent-like and a Patient-like) on one and the same participant in the denoted complex eventuality. This carries over to passive participles derived from inherently unergative verbs. The reason is that a prototypical passive construction requires its subject to have a high number of Proto-Patient properties, yet a passive participle of an unergative verb supplies a subject argument that carries a number of Proto-Agent properties, given that it corresponds to the direct object of an active transitive verb (*The rider raced the horse*), which in turn corresponds to the subject of the active intransitive verb (*The horse raced*). To return to our lead example, in (15) we see that the PRO subject of the passive participle has the same thematic properties as the corresponding active object in (14), hence it is not a "good" Patient. The main verb *fell* assigns the property 'movement' to its subject *the horse.* In so far as this can be interpreted in terms of 'movement relative to the position of another participant', and given that *the horse* in (15) is a sentient being with a (potentially) certain volitional involvement in the racing event, 'movement' can be here taken as the Proto-Agent property. (This is not uncontroversial. However, *fell* does not assign clear Proto-Patient properties to its subject either. A candidate might be 'undergoes a change of state', but here it would not mean a permanent change, rather just a change in bodily posture, and hence ultimately 'movement'.) Hence, the thematic make up of the subject NP in the matrix clause does not match the thematic constraint of the reduced relative clause which requires that its PRO subject be a "very good" Patient.

(15) **The horse**$_i$ [< **PRO**$_i$ > **RACED past the barn**] **fell.**
| |
PA PA and **PP**
+ movement (+ volition) + **causally affected**
+ sentience
+ movement

If, on the other hand, the main verb of a matrix clause assigns Proto-Patient, rather than Proto-Agent, property (or properties) to its subject, the magnitude of the garden path effect is diminished, as (16) shows: the subject of *died* is clearly a "better" Patient then the subject of *fell*, as it is entailed to undergo a permanent change of state. Hence, (16) is somewhat easier to interpret than (15).

(16) **The horse**$_i$ [<**PRO**$_i$ > **RACED past the barn**] **died.**
| |
PP PA and **PP**
+ **undergoes change of state** (+ volition) + **causally affected**
+ sentience
+ movement

Of course, not all transitive and passive uses of inherently unergative verbs are odd. Other factors, such as expectations related to the occurrence of highly conventionalized combinations of words and general world knowledge, may come into play and override the semantic mismatch described above. For example, *John walked his dog* and *Fido was walked by John tonight* sound highly natural.

Let us now look at sentences with reduced relatives headed by passive participles derived from unaccusative verbs. In (17a) the subject of the unaccusative *melted, the butter*, corresponds to the object of the active transitive *melted* in (17b), they are both entailed to have at least two Proto-Patient properties: 'change of state' and 'Incremental Theme'. Hence, they are "very good" Patients, and we can expect that both the transitive and passive uses of *melted* are perfectly acceptable.

(17) a. The butter MELTED in the pan.
b. The cook MELTED the butter in the pan.

(18) a. The butter MELTED in the pan was fresh.
b. #The butter MELTED on the stove dripped onto the kitchen floor.

As is predicted by the hypothesis in (13), (18a) is judged easier to process than (18b). (18b) contains the matrix verb *dripped* that entails that the referent of its subject argument moves relative to the position of another participant, and hence can be viewed as entailing one Proto-Agent property in its subject argument. This, however, is inconsistent with the requirement stated in our hypothesis (13) that the subject NP in the matrix clause matches in its Proto-Patient properties the thematic make up of the PRO-subject of the reduced relative clause. (18a) contains the stative predicate *be fresh* in the matrix clause, which entails no Proto-Agent properties in its subject argument, and hence (18a) is more acceptable than (18b).

4.3 An HPSG approach

In the past ten years or so there has been a growing convergence of results and methodological assumptions coming from psycholinguistics and theoretical linguistics in the domain of constraint-based approaches to natural language description (e.g., Pollard and Sag, 1987; Pollard and Sag, 1994; Sag, 1998, for example; see Tanenhaus and Trueswell, 1995 and MacDonald, 1997 for a review of constraint-based approaches in psycholinguistics). They share two main assumptions: First, a sentence's interpretation requires satisfaction of multiple (possibly differentially weighted) constraints from various domains of linguistic and non-linguistic knowledge. Second, the integration of such diverse constraints is facilitated by the information contained in lexical entries. Verb-based syntactic and semantic patterns provide a guide for interpreting key aspects of the sentence's structure and meaning, whereby semantic constraints often have a privileged status.

The lexical constraint-based approach proposed here has all the main hallmarks of recent versions of HPSG (see Sag, 1998, for example). Assumptions about lexical semantics of verbs and linguistic information directly associated with extra-linguistic context and general world knowledge are influenced by Fillmore's work and Construction Grammar (see Fillmore and Kay, in press). The grammar assumed here is monostratal, non-derivational and non-modular. It is characterized declaratively by specifying types of well-formed linguistic expressions (e.g., words, phrases, part of speech classes, argument structure classes, and traditional morphological classes, for example) and constraints on those types. All properties of linguistic expressions are represented as feature structures. Language-particular rules and universal principles are characterized as systems of constraints on feature structures. The main explanatory mechanism is unification in the narrow sense of structure sharing of token-identical feature structures (cf. Pollard and Sag, 1994).

Since lexical entries constitute the key ingredient for interpreting the main aspects of the sentence's structure and meaning, and facilitate integration of diverse types of knowledge, let us introduce their main features using a simplified lexical entry for the transitive active *raced* in (19):

$$\text{(19)}\quad \begin{bmatrix} \text{PHON} & \textit{raced} \\ \text{SYN} & \begin{bmatrix} \text{HEAD} & \textit{verb} \\ \text{CAT} & < [1]\text{NP}, [2]\text{NP} > \end{bmatrix} \\ \text{SEM} & \begin{bmatrix} \theta & < e, [1]_i, [2]_j > \\ \text{CONTENT} & \begin{bmatrix} \text{psoa} & \\ \text{PRED} & \begin{bmatrix} \text{REL} & \text{race} \\ \text{racer} & \text{i} \\ \text{racee} & \text{j} \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \text{CONTEXT} & [\ \ldots\ ] \end{bmatrix}$$

(19) contains phonological, syntactic, semantic and pragmatic information, encoded as values of the feature attributes PHON, SYN, SEM and CONTEXT, respectively. The value of SYN encodes syntactic information required for constructing syntactic projections headed by *raced*. The linking between the syntactic (SYN) and semantic (SEM) structure in the lexicon is mediated via co-indexation of syntactic arguments and thematic argument slots, and motivated by Dowty's Argument Selection Principle (here given in (12)). Each argument slot in the thematic structure of a verb corresponds to a cluster of Proto-Agent and/or Proto-Patient properties (cf. Dowty, 1991). Thematic argument slots in turn are co-indexed with individuals in the predication feature structure PRED, which together with 'psoa' (parametrized state of affairs) constitutes the value of CONTENT. The feature structure PRED captures the assumption that verbs semantically express relations between individuals. The attributes 'racer' and 'racee', which correspond to 'frame-specific participants' in Fillmore (1986) or 'individual thematic roles' in Dowty (1989), include properties that we associate with the individuals 'i' and 'j' on the basis of knowing that the statement

'i raced j' is true. In a given single-clause predication, further semantic restrictions on participants are imposed by the interpretation of noun phrases. For example, '[racer i]' will be constrained by the content of the NP filling the '[1]NP' place. PRED does not provide an exhaustive account of all that we know about the meaning of a given verb. What role an individual plays in a given situation depends on a number of other factors, including world knowledge, which is encoded under 'psoa'. (For a related, though not identical, use of 'psoa' see Pollard and Sag, 1994; Sag and Wasow, 1997.) Lexical entries of verbs also include frequency information about the occurrence of a given verb form in the language, about its argument structures, and the like.

Apart from the lexicon, the grammar will minimally include the level of verb forms and the syntactic level with phrasal templates. This is illustrated in a highly simplified Figure 2.

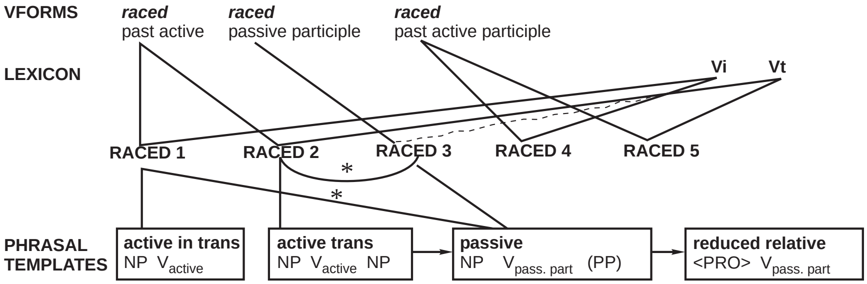


Figure 2. A simplified outline of a constraint-based model.

In general, types at each level of representation are cross-classified in multiple inheritance hierarchies according to their shared information. (Due to the limitation of space, this is not represented in Figure 2.) The information shared by a given class of objects is associated with a general type and is automatically passed down from the general type to specific members of the class. For example, RACED2 and RACED5 inherit information from the generic lexical entry for transitive verbs, here represented by the node **Vt**. Types directly subsumed under the same supertype represent mutually inhibitory alternatives, which often represent multiple interpretation alternatives and differ in frequency of occurrence in the language. For example, RACED2 (active past tense) and RACED3 (passive participle) are mutually exclusive, here indicated by the thick starred line between RACED2 and RACED3. The active intransitive use of *raced* is more frequent than the active transitive one. We assume that such frequency information is encoded in the lexical entries of verbs.

Unification allows us to represent dependencies and connections within one particular level of representation and also among different levels. Feature structures representing compatible types are unified in a new coherent structure by linking them to a single feature structure, which is shown with straight lines (not all such possible connections are here indicated): e.g., [VFORM PAST.ACTIVE] ∪ [SYN Vi]. Feature structures representing incompatible types cannot be unified: for example, active verbs cannot be projected into a passive clause. One advantage of this system is that it allows us to capture the observation that different types of information that characterize the use of a given word are dependent on each other so that accessing one type of information during sentence processing results in accessing others compatible with it. For example, if the sequence *The horse raced…* is understood as the main clause, the information associated with the verb *raced* will be a complex feature structure comprising the information that this verb shares with all active past tense verbs. If the same sequence is understood as the head noun modified by a reduced relative clause, *raced* will be associated with the information shared with all passive participles, and due to its passive argument structure it will also activate the information associated with the active transitive use of *race*.

5. Empirical study of effects at the main verb

We conducted a rating study in which we had six subjects complete questionnaires in which they made judgments about four of the dimensions that Dowty identified as being part of the Proto-Agent cluster: 'volition', 'sentience', 'causing an event or change of state', and 'movement'.[5] The questions concerned the subject argument of the main verb in the matrix clause. Thus to obtain ratings for *The horse raced past the barn died*, the subject would rate *The horse died*. Each simple sentence in the latter set of data was associated with four questions designed to illicit judgments about the four main Proto-Agent properties entailed by the verb for its subject argument. Each question was answered by our subjects using a scale from 1 to 5. For example, in the case of 'volition', 1 would indicate a completely non-volitional participation of the individual denoted by the subject argument (e.g., *The horse died*) and 5 would a fully volitional participation (e.g., *The patients complained*). We then averaged these ratings to come up with a composite Proto-Agent rating, with 1.0 being the lowest and 5.0, the highest. Subsequently, we selected matched pairs of reduced relatives with different main verbs, e.g., *The victims rushed to the hospital complained/died*, in which participants assigned different Proto-Agent ratings for the two main verbs (e.g., *The victims died* vs. *The victims complained*. We were able to identify 21 matched pairs of reduced relatives that met this criterion. We then had another group of subjects rate reduced relatives using these main verbs, e.g., *The victims rushed to the hospital complained/died shortly after arrival.*

Table 3. Rated difficulty for reduced relatives with main verbs differing in the Proto-Agent properties assigned to their subject argument. Numbers in parentheses represent the mean Proto-Agent rating

passive participle derived from	**Proto-Agent properties**			
	low		**high**	
unaccusative verbs	2.32	(1.37)	2.50	(2.35)
unergative verbs	2.81	(2.04)	3.31	(3.83)

The data are presented in Table 3. The numbers in brackets indicate the mean ratings for verbs with low Proto-Agent entailments in their subject argument, and the mean ratings for verbs with high Proto-Agent entailments in their subject argument. An ANOVA conducted on the difficulty ratings revealed a main effect of verb class, F(1,20)=9.50, p<.01, a main effect of the Proto-Agency of the main verb, F(1,20)=5.02, p<.05 and no interaction, F(1,20)=1.10. Overall, then, reduced relatives with main verbs with higher Proto-Agent properties were more difficult than reduced relatives with lower Proto-Agent properties.

To summarize this section, we showed that the unaccusative-unergative distinction that Stevenson and Merlo characterize as a syntactic distinction correlated with difficulty or ease of processing in reduced relative clauses can be re-cast as a distinction that concerns the assignment of thematic roles. One advantage of this novel way of looking at the garden-path phenomenon is that it allows us to understand something that has never been systematically commented on before: namely, the influence of the main predicate in a sentence on the magnitude of the garden-path effect. The analysis in terms of Dowty's thematic roles, formulated in (13), also makes the correct predictions here. These results also support the claim made by Carlson and Tanenhaus (1988), Tanenhaus and Carlson (1989), and in a number of later studies by Tanenhaus and his collaborators, that thematic roles play a central role in language comprehension. We also showed that our thematic analysis is consistent with an independently motivated linguistic model.

Taken together, the current work confirms Stevenson and Merlo's finding that sentences with reduced relatives headed by passive participles derived from unergative verbs pose more processing difficulty than sentences with reduced relatives based on unaccusative verbs. Contrary to Stevenson and

Merlo's claims though, this result is completely consistent with currant constraint-based lexicalist models. We also presented an analysis of the unergative/unaccusative distinction using thematic role properties along with some preliminary supporting evidence. In future research it will be important to combine more sophisticated thematic role representations into a constraint-based processing model.

Acknowledgments

The research presented in this paper was supported NIH grant HD 27206 and by NSF grant SBR-9511340. Many thanks to Michael Spivey for his comments on a previous draft of this paper and to Sarah Scheidel for her help with the questionnaires. We are especially grateful to Paola Merlo and Suzanne Stevenson for helpful discussions and for providing us with the corpus counts we used in our model.

Notes

[1] '+' Department of Linguistics, '*' Department of Brain and Cognitive Sciences.

[2] The unaccusative/unergative distinction (e.g., *melt* vs. *race*) was introduced by Perlmutter (1978), and also noticed by (Hall, 1965).

[3] According to semantic characterizations given by Van Valin (1990) and Dowty (1991), for example, unergative verbs tend to entail agentivity in their single argument and to be aspectually atelic. Unaccusative verbs take a patient-like argument and are mostly telic.

[4] It might be objected that our examples in (3) are easy to process, because they involve complex unaccusative predicates, rather than unergative verbs. However, for English at least, there seem to be no convincing grammatical tests for the unaccusative status of the combination 'unergative verb + directional PP'. (See Levin and Rappapport-Hovav, 1995: 188 and elsewhere, for a discussion of possible candidate tests, such as the occurrence of unaccusatives in the causative alternation.)

[5] One of Dowty's Proto-Agent properties was not included: namely, 'referent exists independent of action of verb'. It does not matter for our analysis, given that the constructions under consideration have the same value for this feature.

References

Bever, T.G. (1970). "The Cognitive Basis for Linguistic Structure." In J.R. Hayes (ed.) *Cognition and the development of language*. New York: Wiley.

Carlson, G.N. and Tanenhaus, M.K. (1988). "Thematic Roles and Language Comprehension." *Syntax and Semantics, 21*, 263–300.

Cruse, D.A. (1972). "A Note on English Causatives." *Linguistic Inquiry, 3*, 520–528.

Dowty, D.R. (1979). *Word Meaning and Montague Grammar. The Semantics of Verbs and Times in Generative Semantics and in Montague's PTQ*. Dordrecht: Reidel.

Dowty, D.R. (1989). "On the Semantic Content of the Notion of 'Thematic Role'." In Chierchia, G., Partee, B. and R. Turner (eds.), *Property Theory, Type Theory and Natural Language Semantics. Volume II: Semantic Issues*. Kluwer Academic Publishers. Dordrecht, Boston, London.

Dowty, D.R. (1991). "Thematic Proto-Roles and Argument Selection." *Language, 67*, 547–619.

Fillmore, Ch.J. (1986). *On Grammatical Constructions*. Department of Linguistics, The University of California at Berkeley. ms.

Fillmore, Ch.J. and P. Kay. In press. *Construction Grammar*. Stanford: CSLI Publications.

Hale, K. and J. Keyser. (1993). "On argument structure and the lexical representation of syntactic relations." In K. Hale and J. Keyser (eds.), *The view from building, 20*, 53–110. Cambridge, MA: MIT Press.

Hall, B. (1965). *Subject and Object in Modern English*. Cambridge, MA: MIT Ph.D. Thesis.

Jackendoff, R. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge, MA: The MIT Press.

Jackendoff, R. (1990). *Semantic Structures*. Cambridge, Mass.: The MIT Press.

Langacker, R. (1986). "Transitivity, Case, and Grammatical Relations: A Cognitive Grammar Prospectus." Ms.

Levin, B. and M. Rappaport Hovav. (1995). *Unaccusativity: At the Syntax-Lexical Semantics Interface*. Cambridge, Mass.: The MIT Press.

MacDonald, M.C. (1997). "Introduction." In *Lexical Representations and Sentence Processing*. A special issue of *Language and Cognitive Processes*. Hove, Great Britain: Psychology Press Ltd.

MacDonald, M.C., Pearlmutter, N.J. and Seidenberg, M.S. (1994). "The lexical nature of syntactic ambiguity resolution." *Psychological Review, 101*, 676–703.

McRae, K., Spivey-Knowlton, M.J. and Tanenhaus, M.K. (1998). "Modeling the Influence of Thematic Fit (and Other Constraints) in On-Line Sentence Comprehension." *Journal of Memory and Language, 38*, 283–312.

Merlo, P. and S. Stevenson. (1998). "What grammars tell us about corpora: the case of reduced relative clauses." Proceedings of the *Workshop on Very Large Corpora, 6*, 134–142.

Perlmutter, D.M. (1978). "Impersonal Passives and the Unaccusative Hypothesis." *Berkeley Linguistics Society, 4*, 15–189.

Pollard, C. and I. Sag. (1994). *Head-Driven Phrase Structure Grammar*. Stanford: CLSI/Chicago: The University of Chicago Press.

Pollard, C. and I. Sag. (1987). *An Information-Based Syntax and Semantics. Volume 1. Fundamentals*. Stanford: CLSI/Chicago: The University of Chicago Press.

Sag, I. (1998). "English Relative Clause Constructions." *Journal of Linguistics*.

Sag, I. and T. Wasow. (1997). *Syntactic Theory: A Formal Introduction*, ms. Stanford University.

Spivey-Knowlton, M. J. (1996). *Integration of Visual and Linguistic Information: Human Data and Model Simulations*. Ph.D. Thesis. University of Rochester, Rochester, NY.

Spivey, M.J. and M.K. Tanenhaus. (1998). "Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1521–1543.

Stevenson, S. (1994a). "Competition and Recency in a Hybrid Network Model of Syntactic Disambiguation." *Journal of Psycholinguistic Research, 23*, 295–322.

Stevenson, S. (1994b). *A Competitive Attachment Model for Resolving Syntactic Ambiguities in Natural Language Processing*. Ph.D. Thesis, Department of Computer Science, University of Maryland.

Stevenson, S. and P. Merlo. (1997). "Lexical Structure and Parsing Complexity." In M.C. MacDonald (ed.) *Lexical Representations and Sentence Processing*. A special issue of *Language and Cognitive Processes*.

Tanenhaus, M.K. and G.N. Carlson. (1989). "Lexical Structure and Language Comprehension." In W. Marslen-Wilson (ed.), *Lexical Representation and Processs*, pp. 529–561. Cambridge: MA: MIT Press.

Tanenhaus, M.K. and J.C. Trueswell. (1995). "Sentence comprehension." In J. Miller and P. Eimas (eds.), *Speech Language and Communication: Handbook of Perception and Cognition*. New York: Academic Press.

Van Valin, R.D.Jr. (1990). "Semantic Parameters of Split Intransitivity." *Language, 22*, 221–260.

Predicting thematic role assignments in context

Gerry T.M. Altmann
Department of Psychology, University of York, Heslington, UK

One of the major goals of sentence processing is to establish *who did what to whom*; that is, to assign thematic roles to the appropriate entities introduced into the discourse by that sentence and any others that preceded it. The original formulations of the constraint-based approach to sentence processing (e.g. MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994) did not specify how thematic role assignment might correspond to a particular interaction between activated representations corresponding to verb argument structures and activated representations corresponding to a verb's arguments. In this chapter, I sketch out this correspondence, drawing on data from a variety of studies which establish that thematic roles can be assigned during the processing of a sentence to discourse entities *before* those entities have been referred to within that sentence. These studies examined the processing of sentences in which the main verb conveyed selectional restrictions which ruled out all but one discourse entity as a potential direct object. The basic finding was that in such cases, the appropriate thematic role was assigned to that one entity before the direct object was itself encountered, and thus before the point at which obligatory syntactic dependencies dictated whether or not that particular role assignment was licensed.

Introduction

'*Paola asked me to submit this chapter precisely three months ago*'. Some readers will, following Frazier (1979), assume that the submission of this chapter is what should have occurred three months ago. Others may assume instead that the asking is what took place three months ago. Either way, this chapter will take as its starting point the finding that the initial interpretation of such ambiguous sentences can, apparently, be influenced by extra-sentential context

(Altmann, Garnham, van Nice, & Henstra, 1998). The purpose of this chapter is not to defend the claim that such influences, on the initial interpretation of syntactically ambiguous sentences, are possible, but rather to question the mechanism by which such influences could, at least in principle, come about. A further purpose will be to explore, given one particular mechanism, some empirical predictions which then follow in respect of contextual influences on the processing of syntactically unambiguous sentences. Specifically, the chapter will explore how thematic information associated with a verb interacts with information contained within the context, and how this interaction can lead to the assignment of thematic roles before any post-verbal arguments, ordinarily associated with such roles, are encountered.

Altmann et al. (1998) explored the processing of sentences such as '*He submitted the chapter he wrote for the edited volume yesterday*'. In the absence of any prior context, people tend to associate the adverb in sentences such as this with the more recent verb – they thus assume that the writing was done yesterday rather than the submission. When we preceded each target sentence by a context of the form '*Paola is wondering when he submitted the chapter he wrote for the edited volume*' we found evidence, using an eye-tracking methodology which allowed us to monitor fixation times on the critical adverb, that people did initially associate the adverb with the earlier verb complex.

To explain this effect of context we proposed that the context set up readers' expectations regarding what kind of information may appear where in the subsequent target sentence. Specifically, we suggested that information contained within the context caused the predictive activation, at different positions within the target, of some generalized representation corresponding to the sought-after adverbial. This is illustrated in the example below (each such position is marked with '△'):

(1) When did he submit the chapter he wrote for the edited volume?
$_{\triangle}$He submitted the chapter $_{\triangle}$ he wrote $_{\triangle}$ for the edited volume $_{\triangle}$ . . .

We argued that the question (whether direct, as in (1) above, or indirect as in the example given in the text) set up expectations regarding the possible locations of the relevant temporal information in the target sentence, and that these expectations were manifested as the predictive activation of adjunct structures which supported the subsequent integration of that information. We equated predictive activation with the 'support of subsequent integration' in order to draw a parallel with the way in which encountering a verb causes the activation of that verb's argument structure, which can be thought of as the projection of structures which support the integration of subsequent information contained within the sentence with information concerned with the meaning of that verb.

This predictive activation is similar, in spirit at least, to the predictive activation exhibited by Jeff Elman's simple recurrent network (SRN) that learned to predict, given the input so far, what input to expect next (e.g. Elman, 1990). In Elman's simulations, an SRN was able to compute and encode the contingencies between an individual word and the local contexts in which that word was experienced during the training period ('context' refers here to the other words in the same sentence). The encoding of these contingencies resulted in the network being able to 'guess' what could come next given the input thus far; on encountering particular verbs, only certain nouns would be predicted to come next (as a function of which nouns had followed those verbs in the training sequences – reflecting, in effect, selectional restrictions on what could appear in object position). Interestingly, and this is a point that will be revisited later, on encountering particular nouns, only certain verbs would be predicted to come next (as a function of restrictions on what could appear in subject position). The important principle underlying aspects of this work was that the unfolding of a sentence (in time) reflects the manifestation of predictive contingencies between elements of that sentence and the context (linguistic or otherwise) in which that sentence appears – with the relevant contingencies changing dynamically as the sentence unfolds. Thus, the encoding of a predictive contingency between aspects of the context and some subsequent input would create, if those aspects of the context recurred in a subsequent sentence, a predictive 'expectation' regarding subsequent input.

This last principle is reflected in (1) above, where it is assumed that a question (that is, an extra-sentential context) sets up an expectation regarding where, in the answer to that question, the information that actually answers the question could possibly be located. But the principle is also reflected in the notion of verb argument structure as introduced above. A verb's argument structure in fact represents the predictive contingencies that hold between a verb and other elements at particular positions relative to that verb within the sentence. An object that violates a verb's selectional restrictions, for example, will violate one such contingency. These contingencies are, by definition, probabilistic; a verb like *eat* will tend to be followed in object position by an expression referring to something edible, but will on occasion be followed (if at all) by something else (e.g. *words* or, when attempting to get into the record books, *car*).

Selectional restrictions – that is, the predictive contingencies that dictate what kinds of thing can take part in the event denoted by the verb, and thus what kinds of thing can receive a thematic role from the verb – are just one kind of restriction, or *constraint*, on what can follow a verb, and on what, consequently, can be predicted at the verb. In principle, *contextual* information, as embodied either within a linguistic or real-world context, could also restrict the range of things that could occur in object position. In (2) below, for example, there is a high likelihood that whatever occurs in object position in the final sentence will refer to the stepladders that were introduced in the first sentence:

(2) The librarian looked around for some stepladders that she needed.
She climbed...

If there is a tendency (perhaps only a small one) for the referring expressions in different argument positions to refer to entities that already exist either in the prior discourse or in the real world context (cf. Murphy, 1984, who demonstrated a processing cost associated with introducing new entities as opposed to referring to existing ones), there must exist predictive contingencies (perhaps only weak ones) between what can be referred to in the different argument positions and the extra-sentential contexts within which a verb may tend to occur. So long as the human sentence processing mechanism is sensitive to these contingencies, and uses them as the basis for predictive activation in the manner outlined above, it will activate, to some degree at least, representations at the verb which relate to whatever entities in the context of the sentence could subsequently be referred to in those different argument positions. In other words, so long as the sentence processor is sensitive to this probabilistic relationship between context and verb, it will, in (2) above, activate some representation corresponding to the stepladders when the verb *climbed* is encountered. In contrast, the equivalent is not true in (3):

(3) The librarian looked around for some papers that she needed.
She climbed...

In this case, what will be climbed will not, presumably, be the papers – there is thus a mismatch between the expectation that the verb will apply with some probability to something already mentioned, and what has actually been mentioned. The fragment in (3) could continue quite plausibly, though, as in '*she climbed the stairs to her office*', and it is thus unclear whether any behavioural consequence should be expected given that mismatch. The following, empirical, sections of this chapter will explore the processing consequences of the contrast between (2) and (3). In particular, they will establish that the mismatch in (3) does have quite profound behavioural, and theoretical, consequences.

Monitoring for contextual mismatch

A useful starting point for considering how to explore empirically the relationship between examples (2) and (3) above is a series of studies by Julie Boland and colleagues (Boland, Tanenhaus, Garnsey, & Carlson, 1995). Boland et al. contrasted filler-gap sentences such as:

(4) Which stepladders did the librarian climb_whilst tidying up?

(5) Which papers did the librarian climb_whilst tidying up?

(These are modified versions of the Boland stimuli; the gap is marked with an underscore.) Using a word-by-word 'stop making sense' judgement task, Boland et al. found that there were increased 'no' (it makes no sense) judgements in response to the verb *climb* in (5) relative to (4). They interpreted this result as implying that the papers were interpreted, at the verb *climb*, as the thing that was (improbably) climbed, even though, grammatically speaking, no such commitment is required, as shown by the alternative continuation in (6):

(6) Which papers did the librarian climb the stepladders to reach_whilst tidying up?

In many respects, the contrast between (4) and (5) is similar to that between (2) and (3). In both cases, the theoretical question is the same: Does the processor, at the verb, anticipate which previously mentioned entity should fill the thematic role associated with whatever will occur in object position? In the case of examples (4) and (5), this translates into whether or not the processor, at the verb, attempts to associate the wh-filler with the role that should be assigned to whatever will be co-indexed with the subsequent gap. For examples (2) and (3), this translates into whether or not the processor attempts, at the verb, to associate the role that should be assigned to whatever will appear in the subsequent object position with whatever had previously been mentioned. And in both cases, there is just one 'antecedent': In (4) and (5) it is limited to the wh-phrase, and in (2) and (3) to the one other entity that has been entered into the discourse model. Because of these similarities between the two sets of contrasts, a series of studies was conducted using the same methodology as used by Boland et al. (1995).

Study 1

The first study contrasted the following conditions:

(7) Antecedent condition
A car was driving downhill when it suddenly veered out of control.
In its path were some pigeons and a row of bollards.
It injured several bollards that came close to being destroyed.

(8) No-antecedent condition
A car was driving downhill when it suddenly veered out of control.
In its path were some dustbins and a row of bollards.
It injured several bollards that came close to being destroyed.

('Bollard' is the British English for a post blocking off access to a road). The labels '*antecedent*' and '*no-antecedent*' refer to whether or not something is introduced in the context (e.g. *pigeons*) which could subsequently fill the role associated with the grammatical object of the verb (*injured*) in the final sentence. In each case the target sentence referred in object position to an implausible entity (but see below for a replication that employed only plausible objects). If the processor is sensitive to the thematic fit between the main verb in each target sentence and the entities introduced in the prior context, we should see evidence of this sensitivity at the main verb – in the no-antecedent condition we should find evidence of an anomaly on *injured* when compared against the antecedent condition; in the antecedent condition the context introduces something that can fill the patient role associated with *injured*, whereas in the no-antecedent condition, no such injurable thing is introduced.

Thirty-two pairs of passages such as (7) and (8) were presented, intermingled with 68 filler passages (50 of which were plausible throughout, and which resembled the experimental items to varying degrees), to 42 participants. The first two sentences of each passage were presented one at a time, with participants having to press a button for each sentence. The third sentence was presented using a moving window paradigm, with all letters replaced initially by hyphens, and each successive button press revealing the next word (and causing the previous one to revert to hyphens). Participants were instructed to press a 'no' button as soon as they thought that the sentence ceased to make sense. Over the course of the instruction and practice phases, participants were shown just four examples of sentences that did not make sense, and in each case the direct object violated the selectional restrictions of the verb (e.g. '*he ate the books on golf*' and '*he read the fish on golf*').

At issue in this study was whether ‘no’ responses would be confined to the implausible object, or whether a difference between the two conditions might emerge at the verb. Because a ‘no’ response would terminate the trial, the number of ‘no’ responses that could be made at the post-verbal noun, for instance, would be contingent on how many ‘yes’ responses had been made at the verb. Thus, for any particular position within the sentence, the percentage of ‘no’ responses was calculated as a percentage of the responses still available at that point.

In the antecedent condition, 0.8% of responses at the verb were ‘no’ responses. This climbed to 78.9 at the post-verbal noun (it was always implausible as the direct object of the verb). For the no-antecedent condition, these figures were 6.3% and 76.1% respectively. The difference at the verb (0.8 vs. 6.3) was statistically significant; the difference at the post-verbal noun was not. The figures for words one to four of the target sentences are shown in Table 1. This table shows also the time it took participants to respond ‘yes’ at each position. The proportion of ‘no’ responses in the no-antecedent condition is small, and consequently it is useful to consider whether, when subjects responded ‘yes’, there is still some evidence of a processing cost associated with the no-antecedent condition. At the verb, the difference in reaction time (607 ms for the no-antecedent condition vs. 572 ms for the antecedent condition) just failed to reach significance (but see below).

The experiment also included two further conditions: In (7) and (8), the verb *injured* can apply to just one of the discourse entities previously introduced (the pigeons in the antecedent condition) – in the no-antecedent condition, there was nothing in the context to which the verb *injured* could apply, and hence the increased proportion of ‘no’ responses. In the two further conditions, the critical verb was replaced by one which could apply to any of

Table 1. Percentage remaining ‘no’ responses, for each of the first 4 word positions in Study 1. Numbers in parentheses indicate the mean judgement times (msec.) when subjects responded ‘yes’

		Word position			
Target	**Context**	**It**	**injured /missed**	**several**	**bollards…**
Selecting	Antecedent	0.5 (412)	0.8 (572)	1.6 (435)	78.9 (621)
(*injured*)	No Antecedent	0.5 (402)	6.3 (607)	2.1 (470)	76.1 (688)
Non-selecting	Antecedent	0.5 (408)	0.8 (504)	0 (442)	4.8 (624)
(*missed*)	No Antecedent	0 (404)	0.5 (511)	0 (440)	4.6 (637)

the discourse entities – *injured* in (7) and (8) was replaced by *missed*. It was predicted now that there would be no difference in the proportion of 'no' responses across the two contexts. And indeed, there was none, with 0.8% 'no' responses in the antecedent condition, and 0.5% responses in the no-antecedent condition (and 4.8% and 4.6% 'no' responses respectively at the noun). The interaction at the verb, between context (antecedent vs. no-antecedent) and verb (e.g. *injured* vs. *missed*) was statistically significant.

This entire pattern was replicated in a further experiment which was identical in all respects except that the post-verbal nouns were plausible, by virtue of introducing new discourse entities in object position which did not violate the selectional restrictions of the verb (e.g. '*it injured several tourists that were standing close by*'). The fillers were all plausible also, again defined in terms of non-violation of selectional restrictions. Participants were, nonetheless, given examples of such violations during the instruction and practice phases of the experiment. Again, there were significantly more 'no' responses at the verb following the no-antecedent context than following the antecedent context; see Table 2. And again, this difference occurred only for verbs that selected amongst the antecedents (e.g. *injured*) – there was no such difference for non-selecting verbs (e.g. *missed*). Once again, however, the difference for selecting verbs was small – on only 7.3% of trials did participants respond 'no' in the no-antecedent condition. However, although this proportion was still relatively small, the time it took participants to respond 'yes' was now significantly longer in the no-antecedent condition than in the antecedent condition (524 ms vs. 465 ms). This difference interacted with verb type – there was no difference for the non-selecting verbs (verbs which could apply to any of the discourse entities). The effect of context at the verb was not, therefore, restricted to just those 7.3% of trials on which participants responded 'no'.

Table 2. Percentage remaining 'no' responses, for each of the first 4 word positions in the replication of Study 1. Numbers in parentheses indicate the mean judgement times (msec.) when subjects responded 'yes'

		Word position			
Target	**Context**	**It**	**injured /missed**	**several**	**tourists…**
Selecting	Antecedent	0 (396)	2.3 (465)	0 (411)	38.8 (626)
(*injured*)	No Antecedent	0 (391)	7.3 (524)	4.5 (422)	30.2 (623)
Non-selecting	Antecedent	0.3 (403)	0 (470)	0 (411)	33.3 (685)
(*missed*)	No Antecedent	0 (390)	0.3 (463)	0.5 (392)	28.4 (659)

One final point of interest in these last data was the relatively high proportion of 'no' responses to the postverbal noun. Recall that in this replication the postverbal nouns were all plausible (at least insofar as they did not violate the selectional restrictions of the verb). However, they also introduced novel discourse entities, and participants appear to have judged the introduction of such entities as relatively implausible. The one significant difference in the data at the postverbal noun was between the antecedent and no-antecedent conditions for the selecting verbs (39% vs. 30%). One interpretation of this difference would be that when there is a single antecedent to which the verb can apply, a novel entity introduced in object position, as opposed to an expression referring back to that antecedent, is deemed particularly implausible. It is less clear why in the no-antecedent condition there is such a high proportion of 'no' responses to the post-verbal noun (not significantly different from the proportions for the non-selecting verbs) – after all, in this case, the processor should not expect the postverbal noun to refer to any of the antecedents in the context. A further replication of this study produced the same pattern. One possibility, which is necessarily speculative, is that participants were responding on the basis of a general bias against the introduction of novel entities (cf. Murphy, 1984), and that in the no-antecedent condition, the felicity of the post-verbal noun with respect to the verb's selectional restrictions was outweighed, at least in terms of its effect on the dependent measure, by the infelicity of introducing a new discourse entity.

What should we conclude from these data? Evidently, context has an effect – if there is nothing in the context that fits the thematic specifications of the verb, an anomaly is experienced which either takes the form of increased 'no' judgements at the verb, or increased latencies to respond 'yes'. The anomaly thus reflects the mismatch between the expectation that the action denoted by the verb will apply to something in the context and the finding that there is nothing suitable in the context to which it can apply. Of course, the action need *not* apply to anything in the context – both in this study and its replication there were at least as many trials in which it did not apply to anything in the context as there were trials in which it did.

If the human sentence processor operates under an expectation that the action denoted by a verb will apply to an entity already introduced in the context, how might this expectation be implemented in computational terms? One possibility, described in more detail in Altmann (1999), is that when the processor encounters the main verb, it projects structure corresponding to the upcoming referring expression, and then attempts to establish, given thematic criteria (selectional restrictions) associated with the verb, whether there are any discourse antecedents with which that referring expression could be coreferential. In other words, the processor projects structure which it then attempts to interpret anaphorically with respect to the context (an alternative to this account is developed in the penultimate section of this chapter).

The following study explores these effects further, but instead of looking for an anomaly when there is nothing in the context to which the verb can apply, it looks for an anomaly when there *is* something to which that verb can apply.

Study 2

This second study borrows from the filled-gap logic introduced by Crain & Fodor (1985) and by Stowe (1986), and used also by Boland et al. (1995). If the previous data are due to the processor actively predicting, at the verb, that the subsequent object will refer to something previously introduced in the context, then we should see evidence of an anomaly at that object if it is not consistent with the processor's prediction. The study contrasted the following two conditions:

(9) The twins listened to their father talking about their mother.
He asked them to be especially nice to her.

(10) The twins listened to their father talking to their mother.
He asked them to be especially nice to her.

The logic here was that in (9), the processor might anticipate at *asked* that the people being asked would be the twins – there is no one else explicitly mentioned in the context whom the father could be asking (although in principle the verb *talk* might introduce an implicit argument, cf. Mauner, Tanenhaus, & Carlson, 1995, that could be taken to refer to a third party – see below). The subsequent pronoun *them* is entirely compatible with this prediction. In (10), however, the mother is more plausibly the object of the asking (the father was talking to the mother, of whom he can more plausibly ask something than the twins, to whom he was not talking). If the processor predicts, at *asked* that the mother will be referred to next, an anomaly should arise when the postverbal pronoun *them* is encountered instead.

Sixteen pairs such as (9) and (10) were embedded amongst 96 filler items, all of whose final sentences were plausible (in respect of non-violation of selectional restrictions, or number/gender mismatches between any pronoun and its antecedents). Twenty-four participants took part in this study. The procedure was identical to that employed in the previous study, with participants being given five examples, across the instruction and practice phases, of anomalous pairs (e.g. '*Mary was talking to her mother. She told him she'd been unhappy*', '*Jake parked his car in the supermarket car park. He locked the banana and went in*').

Using the same 'stop making sense' judgement task as before, an anomaly was indeed observed on the pronoun *them* in (10) relative to (9); 25.4% 'no' responses to the pronoun in (10) compared to 8.5% in (9). Latencies to respond 'yes' were also significantly longer (on both subject and item analyses) at the pronoun in (10) than in (9). See Table 3 for the judgement data and 'yes' latencies for the first four words of each target sentence.

One interpretation of these data is that by the time the post-verbal pronoun was encountered in the target sentence, the processor had 'assumed' that the mother in the talking-to-their-mother context was the person being asked, and that when an alternative role assignment was signalled by the pronoun, an anomaly arose. In the talking-about-their-mother case, there was some evidence of a slight anomaly (8.5% 'no' responses on the pronoun), reflecting, possibly, the expectation that the pronoun refer to some third party that was implied by the verb but was not explicitly expressed (cf. Mauner et al.'s (1995) evidence that the 'implicit arguments' of a verb can be represented semantically, and can be available for subsequent anaphoric reference even when not expressed).

It is possible, however, that the effect observed on the pronoun is not indicative of the prior assignment at the verb of the Experiencer role (whoever was being talked to), but rather is indicative of a difficulty in resolving the pronoun *when it is encountered* – that is, it is indicative instead of the sensitivities of the anaphoric resolution process to prior discourse structure (cf. Garrod &

Table 3. Percentage remaining 'no' responses, for the first four word positions in Study 2. Numbers in parentheses indicate the mean judgement times (msec.) when subjects responded 'yes'

	Word position			
Context	**He**	**asked**	**them**	**to…**
Plausible Antecedent	1.0 (503)	0.5 (586)	25.4 (860)	14.4 (520)
Implausible Antecedent	0 (505)	0.5 (579)	8.5 (581)	3.1 (516)

Sanford, 1994). Thus, it may be harder to establish the appropriate anaphoric dependency if there exists some other discourse entity which could plausibly, given the situation described in the context, receive the same role as the actual anaphor. However, for this to be the case, either this alternative role assignment must *already* be encoded by the time the pronoun is encountered, or else it has to be 'discovered' during the pronominal resolution process. If the latter is true, it would appear that the plausibility of the alternative assignment can interfere with the resolution process even when number mismatch, for example, between the pronoun (*them*) and the alternative ('*their mother*') might be expected to rule out such interference. Nonetheless, the data are open to alternative interpretations.

In Study 1, an anomaly was observed on the main verb, before its grammatical object was encountered, when there was nothing in the context which could subsequently be referred to in object position – that is, when there was nothing in the context which could receive a thematic role from the verb. Whether or not the processor, when there was a suitable entity in the context, actually *assigned* that role to that entity, was not assessed directly by that experiment. In principle, the processor could have used thematic information associated with the verb to narrow down, or restrict, the range of entities such that, when the referring expression in grammatical object position was subsequently encountered, the search for a referent would already be greatly restricted. An anomaly might then occur if the restrictive constraints provided by the verb have nothing to restrict. In the second study, which found evidence of a 'filled-role' effect (cf. the 'filled-gap' effect), the anomaly was now found when there *was* something in the context to which the verb could apply. In this case, it occurred when the pronoun in grammatical object position did not refer to the one entity in the context which could have been anticipated to fill the object role. One interpretation of these data is that the processor anticipated which entity would be referred to in object position (although as outlined above, there are alternative interpretations). Once again, however, it is unclear whether such an interpretation necessitates that the processor actually *assigned* the role associated with object position to that discourse entity when the verb was first encountered. Instead, thematic information associated with the verb could have been used to restrict the range of entities which might subsequently be referred to, and the anomaly may have arisen when the referring expression in object position failed to refer to anything within that restricted set.

Although the effects reported thus far are compatible with the idea that the processor can use contextual information to anticipate at the verb the thematic role assignments that will subsequently be made when the verb's grammatical object is encountered, they are also compatible with an account in which thematic fit between verb and context is used referentially – thematic fit in this case might narrow down the reference set in anticipation of subsequent reference to that set, but might not necessitate that anything within that set be assigned a thematic role until that subsequent reference is made. Study 3, reported below, attempts to establish whether role assignments *are* actually made prior to that subsequent reference.

Study 3

To address this last issue, dative verbs such as *deliver* were used, which take two post-verbal objects. In (11) below, the two objects are '*machine guns*' in direct object position (which are assigned the theme role), and *them* (referring to the military base) in indirect object position (which is assigned the recipient role).

(11) Hank parked his van outside the local military base.
He delivered some machine guns to them and left.

In this case, the two objects refer to entities that are plausible recipients of their respective roles. If the processor can use contextual information to assign the roles associated with object position at the verb, it could in principle assign the recipient role to the military base as soon as *delivered* is encountered (it would be unlikely for the military base to fill the theme role). Similarly in the fragment shown in (12):

(12) Hank parked his van outside the local preschool nursery.
He delivered…

However, if such early assignments are made, we should find evidence of an anomaly on the post-verbal object in (12′) below:

(12′) Hank parked his van outside the local preschool nursery.
He delivered some machine guns…

We could expect such an anomaly because machine guns would be implausible given the preschool nursery as the recipient. Of course, the fragment could continue plausibly (at least with respect to the intra-sentential role assignments) as in (13):

(13) Hank parked his van outside the local preschool nursery.
He delivered some machine guns to the military base next door.

An anomaly on '*machine guns*' could only arise if it was implausible given the preschool nursery as a recipient of the machine guns. Thus, evidence of an anomaly at this point would strongly suggest that the processor had assumed that the preschool nursery was indeed the intended recipient (this logic is again borrowed from Boland et al., 1995, who used it in connection with filler-gap dependencies as in '*which preschool nursery did Hank deliver the machine guns to_last week?*').

Thirty-two pairs of dative constructions such as (11) and (13) were presented to 24 participants. They were intermingled with 80 filler sentences (all of which were plausible with respect to the intra-sentential role assignments). The procedure, instructions, and practice, were identical to those used for Study 2.

The data for the first six word positions are given in Table 4. There were significantly more 'no' responses at both *machine* (12.7%) and *guns* (30.1%) when the context introduced the implausible recipient ('*preschool nursery*') than when it introduced the plausible recipient ('*military base*'). In the latter case, the figures were 2.4% and 5.1% respectively. Latencies to respond 'yes' mirrored these patterns (longer latencies when the context introduced an implausible recipient), but the difference was significant on the by-subjects analysis only. These data suggest that, in the examples above, the preschool nursery is assumed to be the recipient of the delivery by the time '*machine guns*' is encountered.

Of course, an alternative interpretation would be that the observed increase in 'no' judgements when the van was parked outside the preschool nursery reflects the implausibility of the scenario as a whole, and not the implausibility of any specific role assignment. The same anomaly might, for example, be found on '*machine guns*' in (14) below:

(14) Hank parked his van outside the preschool nursery. He <u>saw</u> some machine guns...

Although the verb *saw* does not assign a recipient role to *nursery* in the same way that *delivered* would, it could in principle assign a locative role to *nursery*. The locative would not be specified as part of the verb *saw*'s argument structure, and so cannot be considered an implicit argument (cf. Mauner et al., 1995) in the same way that, for example, the recipient would be for the verb *deliver*. Nonetheless, the grammar does specify an optional adjunct, and if the processor can project structures at the verb corresponding to forthcoming arguments, it may also be able to project structures corresponding to forthcoming adjuncts (cf. Altmann et al., 1998 – see above), in which case an anomaly in (14) would be explained in much the same terms as it was explained for the

Table 4. Percentage remaining 'no' responses, for the first six word positions in Study 3. Numbers in parentheses indicate the mean judgement times (msec.) when subjects responded 'yes'

Word position						
Context	**He**	**delivered**	**some**	**machine**	**guns**	**to…**
Plausible Antecedent	0 (441)	0.5 (525)	0.8 (497)	2.4 (609)	5.1 (669)	0.6 (462)
Implausible Antecedent	0.3 (434)	0.5 (532)	2.9 (503)	12.7 (647)	30.1 (763)	13.6 (528)

actual items used in this study (namely that a 'no' response on '*machine guns*' would be based on the implausibility of machine guns given a prior assumption about the location). As Mauner et al. (1995) observed, the encoding of implicit arguments may support important aspects of text coherence – the predictive activation of representations corresponding to forthcoming arguments, as well as forthcoming adjuncts, may play a similar role.

To summarise the data thus far: In the first study, simple transitive verbs were presented in contexts which either did or did not introduce an entity which could take part in the action denoted by the verb – that is, an entity which could receive a thematic role from the verb. Participants in that study showed signs of experiencing a perceptual anomaly at the verb when there was nothing in the context that could receive the patient role (the role that would ordinarily be assigned to whatever would appear after the verb in object position). In the second study, a pronoun was present in grammatical object position following verbs such as *tell* and *remind*. In these cases an anomaly was observed when the context introduced someone who could plausibly be the person who was being told or reminded. When there was no such person introduced in the context, there was no equivalent anomaly. In both conditions, the actual pronoun always referred to the first-mentioned people in the context. Thus, whereas the first study found an anomaly (on the verb) when there was nothing in the context that could plausibly be referred to in subsequent object position, the second study found an anomaly (on the post-verbal object) when there was something in the context that could more plausibly be referred to in that position. The third study, which employed dative verbs that take both an direct and indirect object (receiving theme and recipient roles respectively), also monitored for an anomaly effect in grammatical (direct) object position. In this case, the effect was mediated by whether or not the context introduced a plausible recipient for whatever was referred to in direct object position. Thus, even before the indirect object was reached, the direct object was considered anomalous if the context had introduced an implausible recipient for the entity referred to in direct object position.

The data all converge on the view that the thematic fit between a verb and its context is evaluated as soon as the verb is encountered, and that if, in a sufficiently constrained context, there is just one entity in the context that can receive a hitherto unassigned role from the verb, that entity will indeed be assigned that role.

Much rests, however, on the status of these data; do they reflect unconscious sentential processing as it happens, or do they reflect a perhaps conscious or 'late' integrative stage in the process that bears little relation to 'normal' processing? The same methodology and associated interpretive criteria were adopted in the first three studies as in previous studies of thematic role assignment which have used the 'stop making sense' judgement task (e.g. Boland et al., 1995). There, anomalies that were directly comparable with those reported here were interpreted, like here, as indicating which thematic roles had been assigned when, and to what. Recently, Pickering & Traxler (1998) found the same patterns of data as were observed by Boland et al. (1995) with similar sentence structures. They monitored participants' eye movements as they read these sentences, and found increased first-pass reading times where Boland et al. had found increased 'no' judgements. At least in respect of the Boland et al. study, eye-movement analyses converge on the same pattern, and associated interpretation, as was found with the judgement task. The parallels between the structures used here and those used in the Boland et al. study make it unlikely that the judgement task has tapped fundamentally different processes as a function of which study it was used in. Nonetheless, the data presented thus far should perhaps be best viewed as suggestive, and further research is currently underway to validate, using alternative methodologies, the interpretation that has been given here. The fourth study, reported below, illustrates one such alternative.

Study 4

The final study in this series borrows from insights developed most recently by Michael Tanenhaus and colleagues (Allopenna, Magnuson, & Tanenhaus, 1998; Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995; Sedivy, Tanenhaus, Chambers, & Carlson, 1999; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1996), and originally proposed by Roger Cooper (Cooper, 1974). For example, Eberhard et al. (1995) and Sedivy et al. (1999) demonstrated that when shown a visual scene containing just one red item, participants would initiate eye movements to that item on hearing the word *red* uttered in '*pick up the red*...' or '*is there a red*...?' These studies demonstrated that intersective adjectives such as *red* can be used to restrict the domain of reference – if there is only one red item in the visual context, the processor can anticipate which item is about to be referred to. An extension of this paradigm is to explore what would happen if there was just one edible entity in the visual scene, and participants heard either '*the boy will eat* ...' or '*the boy will move*...'. Following the logic of Study 1, we might expect *eat* to trigger eye movements to the one edible thing in the visual context, whereas for *move* we might expect no such anticipatory eye movements (unless, perhaps, the edible object was also the most plausibly moveable object given the alternatives – see below).

Sixteen pairs of sentences such as those shown in (15) were recorded.

(15) The boy will eat the cake.
The boy will move the cake.

For each pair, a semi-realistic visual scene was constructed using a commercially available library of ClipArt images. For example (15), the scene showed a young boy sitting on a floor in front of whom there were a toy train set, a toy car, a balloon, and a birthday cake. In all cases there was one target object in the visual scene (in this case the cake) and three distractors (all of which could plausibly be moved). A further sixteen sentence-scene pairs were created to serve as fillers. In these cases the direct object mentioned in the spoken sentence did not have a corresponding referent in the visual scene. Participants were instructed to judge whether the sentence they heard could in principle apply to the picture. They were given the example 'the person will light the fire' and were told to respond 'yes' if the picture showed a fireplace, and 'no' if it did not. No mention was made of the speed with which they should respond.

Two groups of 12 participants each took part in this study, with each participant hearing only one version of each sentence pair, but seeing the same visual scene. The onset of each sentence coincided with the onset of the visual stimulus. Eye movements were monitored using an SMI EyeLink head-mounted eye-tracker sampling at 250 Hz.

The dependent measure was the time, relative to the onset of the target noun (*cake*), at which eye movements to the corresponding visual entity were launched. Trials on which the eyes were fixating the target at verb onset, or on which the eyes were moving to the target at verb onset, were eliminated from the analysis (amounting to 10% of trials). The mean launch time in the *move* condition was approximately 130 ms after the onset of the target noun. The mean launch time in the *eat* condition was 85 ms *before* the onset of the target noun. This difference was significant on both the by-subjects and by-items analyses. The probability of fixating the target entity by the time the onset of the target noun was heard was greater in the *eat* condition than in the *move* condition. There was some evidence in the *move* condition that movements to particular entities on particular trials were consistently launched prior to noun onset, even though the verb could in principle apply to more than one entity in the scene. In all likelihood this reflects a simple plausibility effect (with certain things being more plausibly moved, for example, than others in any given sentence-scene pairing), and further research is underway to explore this.

The data from this study support the conclusions derived on the basis of the first three studies, despite coming from a methodology that is quite distinct from the less than natural task used in those studies. Once again, it appears that the thematic fit between a verb and its context is evaluated (at least with respect to whether there are entities within that context which could in principle receive a role from that verb) as soon as the verb is encountered.

Predictive processing and thematic role assignment

This chapter started by considering the relationship between questions such as '*When will he answer the question?*' and answers such as '*He'll answer the question that was posed tomorrow*'. But how does thematic role assignment relate to the kind of predictive activation that was proposed in order to explain those effects of context? In the computational framework developed by Elman (1990; see also Jordan, 1986), a predictive contingency between some aspect of the intra-sentential context and some subsequent input creates, if that aspect of the context recurs in a subsequent sentence, a predictive 'expectation' regarding subsequent input. This same logic (which can be captured by networks other than SRNs, and indeed, by any mechanism capable of encoding probabilistic contingencies) can be applied to contingencies between an individual sentence and its *extra*-sentential context – any contingency between, for example, edible things in the context, and the occurrence of a subsequent fragment such as '*he ate X*' where *X* was one of those edible things would lead, when '*he ate*' was encountered, to the activation of internal representations corresponding to the edible alternatives contained within the context. However, these representations would reflect more than just the identity of the subsequent input. They would reflect also the relationship between that predicted input, the fragment '*he ate…*', and the context within which that fragment and subsequent input occurred – a contingency between one thing, another thing, and the contexts in which they can co-occur, must presumably reflect the relationship between the two things; the relationship between the cheese and the eating in a sentence such as '*he ate the cheese*' can be defined in terms of the predictive contingencies that exist between eating, cheese, and the contexts in which they each occur (as defined experientially). The activation of representations encoding these contingencies thus corresponds to the 'recognition' of that relationship, and as such, to the assignment of a particular role associated with the verb to its object.

It could be argued that a system behaving in this way might activate representations corresponding to discourse entities in a purely associative way without any explicit representation of the relationships underlying the contingencies. However, it is unclear how the representation of the appropriate contingencies does not constitute the representation of the corresponding relationship when the consequences of those contingencies are manifest, for example, in supporting subsequent predictions. According to Dienes & Perner (in press), such representation would constitute explicit representation of the relationship facts (even if the system did not have 'causal understanding' of the processes underlying those relationships – although issues regarding the computational implementation of causal understanding are beyond the scope of this chapter).

Elman's SRN would, when it encountered a verb, predict the range of nouns which, in its experience, could follow that specific verb. The network thus acquired a range of contingencies which encoded, in effect, the selectional restrictions associated with the different verbs in the limited language on which the network was trained (as defined by the statistically observable restrictions on what could or could not follow what). But as mentioned earlier, the network also would predict, on the basis of the first noun in each sentence, the range of verbs which, again in its experience, could follow that specific noun. For instance, *plate* in subject position could not be followed by the verb *sleeps*, and thus *sleeps* would not be amongst the verbs which the network would anticipate after encountering a sentence-initial *plate*. This finding is quite unsurprising given that it merely reflects the selectional restrictions on which kinds of thing can appear in subject position for a given verb. However, although unsurprising, it does lead to an alternative interpretation of one of the main empirical findings presented here: In the first study, there were more 'no' judgements to a verb like *injure* following a context which introduced nothing animate than following a context which did introduce something that was animate. On the one hand, this result could be interpreted as reflecting the mismatch between expecting the verb to apply to something already introduced in the context and finding that there is in fact nothing in the context to which it could apply (or, as suggested earlier, as the projection at the main verb of the upcoming referring expression, and the subsequent failure to establish anaphoric relations between this projected expression and the context). But an alternative interpretation (or rather, a restatement of the same idea) is that the context created an expectation for only certain verbs and not others – that is, on encountering the sentential subject in the target sentence, the processor anticipated (predictively activated representations corresponding to) only certain verbs. Precisely which verbs were anticipated was a function of which entities were introduced in the prior context – in effect, the set of possible thematic roles that could be adopted by the discourse entities restricted the class of potential verbs that could be projected to follow the sentential subject. Thus, the anomaly on the verb in '*the car injured...*' may have reflected the mismatch between the verb *injured* and the range of verbs that had been anticipated to occur in this position (and which had been anticipated on the basis of a context which introduced nothing animate). Under this account, selectional restrictions work both ways – they restrict the range of nouns (or rather, the range of things those nouns can refer to) that can occur in the different argument positions of a given verb, and they restrict the range of verbs (or rather, the range of actions denoted by those verbs) that can take as argument a given noun (as referring to a specific participant in the potential action).

Conclusions

Contemporary theories of sentence processing are based on the notion that multiple sources of information interact during sentence processing, with each source of information providing probabilistic constraints which are applied in parallel (cf. MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell & Tanenhaus, 1994). The selectional restrictions associated with a verb constitute a constraint on what can be referred to next. Referential restrictions associated with which entities exist within the context of the sentence constitute another constraint on what can be referred to next. On one interpretation of the constraint satisfaction approach to sentence processing, probabilistic constraints are applied as soon as they each become available. If the selectional and referential constraints are applied as soon as they become available, the only issue then concerns the nature of the consequences of such application. I have argued here, and elsewhere (see Altmann, 1997; Altmann, 1999) that one consequence is, in effect, the assignment, at the verb, of the thematic roles normally associated with whatever will appear in subsequent object position to entities that have been previously introduced in the context. Crucially, this means that the assignment can take place before that subsequent object position is reached. The original formulations of the constraint-based approach to sentence processing did not specify the correspondence between thematic role assignment and interactive activation – that is, they did not specify how thematic role assignment might correspond to a particular interaction between activated representations corresponding to verb argument structures and activated representations corresponding to a verb's arguments. The data presented here are a first step towards exploring the consequences of taking one particular theoretical view of this interaction.

Note

The work reported in this chapter was supported by grants from the Nuffield Foundation, from the Medical Research Council (G9628472N), and from the Economic and Social Research Council (R000222798). Yuki Kamide constructed the stimuli for Study 4. Further details of the studies reported here, and other relevant data, appear in Altmann (1999) and Altmann & Kamide (1999). The author thanks the editors and two anonymous reviewers for comments on a previous version of this chapter.

References

Allopenna, P.D., Magnuson, J.S. & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419–439.

Altmann, G.T.M. (1997). *The ascent of Babel: An exploration of language, mind, and understanding*. Oxford: Oxford University Press.

Altmann, G.T.M. (1999). Thematic role assignment in context. *Journal of Memory and Language*.

Altmann, G.T.M., Garnham, A., van Nice, K., & Henstra, J.A. (1998). Late Closure in Context. *Journal of Memory and Language, 38*(4), 459–484.

Altmann, G.T.M. and Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*(3), 247–264.

Boland, J.E., Tanenhaus, M.K., Garnsey, S.M., & Carlson, G.N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language, 34*(6), 774–806.

Cooper, R.M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing, *6*(1), 84–107.

Crain, S., & Fodor, J.D. (1985). How can grammars help parsers? In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives*. Cambridge: CUP.

Dienes, Z., & Perner, J. In press. A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*.

Eberhard, K., Spivey-Knowlton, S., Sedivy, J., & Tanenhaus, M. (1995). Eye movements as a window into real-time spoken language processing in natural contexts. *Journal of Psycholinguistic Research, 24*, 409–436.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Frazier, L. (1979). *On Comprehending Sentences: Syntactic Parsing Strategies*. Bloomington, Ind.: Indiana University Linguistics Club.

Garrod, S.C., & Sanford, A.J. (1994). Resolving sentences in a discourse context: How discourse representation affects language understanding. In M.A. Gernsbacher (Eds.), *Handbook of Psycholinguistics*, pp. 675–698. San Diego: Ca.: Academic Press.

Jordan, M.I. (1986). *Serial order: A parallel distributed processing approach* No. 8604. Institute of Cognitive Science, University of California, San Diego.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*(4), 676–703.

Mauner, G., Tanenhaus, M.K., & Carlson, G.N. (1995). Implicit arguments in sentence processing. *Journal of Memory and Language, 34*(3), 357–382.

Murphy, G.L. (1984). Establishing and accessing referents in discourse. *Memory and Cognition, 12*(5), 489–497.

Pickering, M.J., & Traxler, M.J. (1998). Strategies for processing unbounded dependencies: Lexical information and verb-argument assignment. *Submitted for publication*.

Sedivy, J.C., Tanenhaus, M.K., Chambers, C.G., & Carlson, G.N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*.

Stowe, L. (1986). Evidence for on-line gap location. *Language and Cognitive Processes, 1*, 227–245.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*(5217), 1632–1634.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., & Sedivy, J.C. (1996). Using eye movements to study spoken language comprehension: Evidence for visually mediated incremental interpretation. In T. Inui & J.L. McClelland (Eds.), *Attention and Performance XVI: Information integration in perception and communication*. Cambridge, MA: MIT Press/Bradford Books.

Trueswell, J.C., & Tanenhaus, M.K. (1994). Towards a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (Eds.), *Perspectives on Sentence Processing*, pp. 155–179. Hillsdale, NJ: Lawrence Erlbaum.

Lexical semantics as a basis for argument structure frequency biases

Vered Argaman and Neal J. Pearlmutter
Northeastern University

In theories of sentence processing, frequency effects have been demonstrated for a variety of linguistic elements. In most cases, as in the case of argument structure frequency biases, there has been little research with regard to the source of frequency differences. Extending lexical semantic theories of argument structure (e.g., Pinker, 1989; Levin, 1993), we propose that differences in argument structure biases are a function of semantics. We compiled completion-survey and corpus-based data for a set of sentence complement (SC) taking verbs and their corresponding nouns (e.g., *proposed-proposal*). We found significant correlations across these verb-noun pairs in their SC-taking bias, and these correlations remained the same in magnitude when tested within narrower semantic classes. In addition, using Levin's (1993) verb categorization scheme, we found that verbs belonging to different semantic categories differed from each other in their SC-bias. A corresponding analysis using morphological category (i.e., nominalizing suffix) in place of semantic category indicated that morphological properties could not account for differences in SC-bias. The results implicate lexical semantic properties as a source for argument structure frequency biases. Further directions in the study of frequency effects in language processing are discussed.

In language comprehension, the influence of the frequency of various elements has long been recognized, with more frequent elements typically being processed more quickly or easily than less frequent ones. This has been most obvious for words (e.g., Just & Carpenter, 1980; Morton, 1969; Rayner & Duffy, 1986) , but the frequency of a range of elements which must be accessed and combined to interpret sentences has recently been argued to play an important role in comprehension (e.g., Jurafsky, 1996; MacDonald, Pearlmutter, & Seidenberg, 1994; Mitchell & Cuetos, 1991; Tabor, Juliano, & Tanenhaus, 1997;

Trueswell & Tanenhaus, 1994; cf. Traxler, Pickering, & Clifton, 1998). Even in models where the frequency of elements is not one of the primary sources of information used during sentence processing, it still comes into play during later stages (e.g., Ferreira & Henderson, 1990; Frazier, 1995; Mitchell, 1989).

These claims have been supported primarily by ambiguity resolution studies, where the relative frequencies of all of the following have been shown to influence preferences for different interpretations of temporarily ambiguous phrases: (1) a word's alternative grammatical categories (Juliano & Tanenhaus, 1994; MacDonald, 1993; Tabor et al., 1997), (2) a verb's alternative morphological tense markings (Trueswell, 1996), and (3) alternative argument structures for verbs and nouns (e.g., Boland, Tanenhaus, Garnsey, & Carlson, 1995; Ferreira & Henderson, 1990; Garnsey, Pearlmutter, Myers, & Lotocky, 1997; MacDonald, 1994; Pearlmutter & Mendelsohn, 1998; Spivey-Knowlton & Sedivy, 1995; Taraban & McClelland, 1988; Trueswell, Tanenhaus, & Kello, 1993). These results are analogous to earlier findings in lexical ambiguity resolution, where the relative frequency of the different possible meanings of a word (e.g., *bank* as a financial institution vs. as the edge of a river) in part determines which meaning is preferred during comprehension (e.g., Rayner & Duffy, 1986; Simpson, 1984; Tabossi, Colombo, & Job, 1987). While all these types of information and their associated frequencies have specifically been shown to influence ambiguity resolution, the usual underlying assumption is that their influence is present even in unambiguous cases. In the current work, we will focus on argument structure frequency, because it is probably the most broadly-applicable of the lexical frequencies shown to influence sentence comprehension.

Despite the importance of frequency effects in sentence processing, the source of such effects (i.e., why there are differences in biases) has received little attention. This is particularly an issue for theories which rely on frequency bias as an explanatory variable, as in the constraint-based lexicalist framework (e.g., MacDonald et al., 1994) and the linguistic tuning framework (e.g., Mitchell & Cuetos, 1991): It is not problematic to explain the processing performance of particular individuals as a function of the frequency information they have acquired during prior comprehension, but this can lead to a circular explanation (e.g., as noted by Stevenson & Merlo, 1997), in that some account is then needed for the particular frequency biases which these individuals experienced during prior comprehension.

There are two general kinds of solution for this problem: One possibility is that existing frequency biases were originally just small random variations, which were reinforced and magnified over time. This magnification could in principle take place over generations or over the course of an individual's development; Tabor (1995) discusses a similar case of historical change in English in the use of the construction *be going to*. The second possible solution is that some other underlying property or variable is responsible for frequency biases. This variable could operate by influencing the relative frequencies with which the elements in question are produced, thus influencing the comprehension system; or it could operate directly on the comprehension system, in which case it might actually serve as a replacement for frequency bias. For example, the most likely explanation for the frequency biases among the meanings of semantically-ambiguous words like *bank* is that they are largely determined by properties external to language, such as the physics of the world, and human social and cultural phenomena: We spend more time talking about financial institutions than we spend talking about rivers because the former happen to be more important to current everyday life. The result is a higher relative frequency for the financial institution meaning than for the river's edge meaning of *bank* in the comprehension system.

Either type of solution is available in the case of argument structure frequency: Biases might be the result of small, historically-early, random differences across words, magnified and reinforced over time; or they might be the result of some underlying property. This issue has not been investigated in the literature, to our knowledge, so the first question to consider is what property (or properties) might underlie argument structure frequency.

One possible underlying property is lexical semantics. In many lexical-semantic theories (e.g., Jackendoff, 1990; Levin & Rappaport Hovav, 1995; Pinker, 1989; cf. Dowty, 1991), words can be categorized on the basis of a relatively limited set of shared meaning components which are relevant for the words' linguistic behavior, including its permitted argument structures.

In these kinds of theories, word meanings are configurations of basic meaning components plus some idiosyncratic information. For Pinker (1989), the configuration determines the argument structure by specifying how associated phrases are interpreted and by licensing operations which alter the configuration to form alternatives for the word. For example, supposing that the core meaning for the verb *break* specifies that it describes a change-of-state event (e.g., *The vase broke.*), this configuration will specify that its argument (*the vase*) is interpreted as the entity undergoing the change. But this configuration can also license an operation, causitivization, which permits the core change-of-state event to become caused (e.g., *Jerry broke the vase.*). This operation effectively creates an additional meaning for *break*, in which *the vase* is still the entity undergoing the change, but now a second argument (*Jerry*) explicitly causes the change to take place. The idiosyncratic component specifies information specific to the particular referent for the word which is irrelevant to the word's argument structure properties. So, for example, *shatter* and *crack* are both change-of-state verbs like *break*, and they therefore share the same permitted argument structures; they differ from *break* and each other, however, in what they specify about the actual change of state. Thus on theories of this sort, semantic categories can be identified in which all members share the meaning components which are relevant to argument structure, but in which members differ in their idiosyncratic components.

What is most critical about such theories for the present view is that argument structures can effectively be reduced to partial semantic representations (the part excluding the idiosyncratic component). As a result, selecting an argument structure during comprehension is identical to selecting (activating) a meaning for a word, and selecting from among multiple possible argument structures is a matter of resolving a lexical-semantic ambiguity like that associated with the word *bank*. This is a slightly stronger view than has been taken within the lexicalist constraint-based framework, which has generally assumed that argument structure ambiguity is a kind of lexical ambiguity, and that it would therefore demonstrate analogous ambiguity resolution effects. The current claim is that argument structure ambiguity is instead identical to lexical-semantic ambiguity – argument structures are lexical-semantic representations, just as the different meanings for *bank* are.

This proposal has two important properties with respect to argument structure frequencies. First, it extends the approach of Pinker (1989; Levin, 1993, and others) to suggest that lexical semantics not only determine the permitted (and ruled-out) argument structures for a word, but that it also controls the relative frequencies of the various possibilities. Just as the different meanings for *bank* have associated relative frequencies, so do the different meanings corresponding to a word's different argument structures. The second important property of the proposal is that it provides a potential underlying explanation for argument structure frequencies. Because different argument structures have different semantics, they refer to different kinds of events (and states, entities, circumstances, etc.) in the world, and the relative frequency of their use will be determined by the relative frequencies of the things to which they can refer. These latter frequencies need have nothing to do with language, and they can therefore (in principle) be determined independently, by properties of the world (human cultural phenomena, physics, etc.).

The proposal to treat argument structures as word meanings makes an important general prediction: Words which are closely related in meaning will have similar argument structure frequency distributions. In the limit, of course, this is certainly true: A verb whose meaning cannot possibly involve communication or propositional content, for example, will not take a sentential complement (SC) as an argument (e.g., **I napped that Mary was happy*). However, the more critical cases for examining argument structure frequency biases involve finer-grained effects, and we therefore examined the above prediction within sets of verbs and nouns which can at least potentially take an SC as an argument.

We focused on the SC argument structure because verbs which allow it can participate in the direct object versus SC ambiguity, which has the most evidence in the sentence comprehension literature for the importance of argument structure frequency. In addition, most verbs which take SCs have corresponding nouns which can also do so, and these nouns can participate in a related ambiguity (SC vs. relative clause). Assuming that verbs and nouns derived from the same stem have many shared components of meaning, our basic approach is to compare the SC argument structure frequencies of corresponding verbs and nouns (e.g., *propose* and *proposal*) using correlational techniques, and to look at the pattern of such correlations across semantically-coherent categories of verbs. We then also consider the influence of morphological and surface form properties as an alternative to semantics.

Data collection

We compiled a database of completion-survey and corpus-based data for a set of 167 pairs of SC-taking verbs (e.g., *propose*) and their corresponding nouns (e.g., *proposal*). For the verbs and some of the nouns, some or all data had been collected and coded earlier for other purposes (Garnsey, Lotocky, Pearlmutter, & Myers, in preparation; Garnsey et al., 1997; Pearlmutter & Mendelsohn, 1998). As a result, completion and corpus data were not necessarily both available for the verb and the noun member of a given pair, and each of our analyses made use of only a subset of the possible data, as indicated below. Further details of the methodology are provided in Garnsey et al. (1997, in preparation) for the verb data, and in Argaman, Pearlmutter, Randall, and Mendelsohn (in preparation) for the noun data.

A sentence-initial fragment-completion task was used to collect completion-survey data at the University of Illinois (verbs) and Northeastern University (nouns). For the verbs, the fragment always began with a proper name followed immediately by the verb in its past tense form (e.g., *Bill proposed*). For the nouns, the fragment always began with a proper name followed by a verb, a determiner (typically *the*), and the noun in singular form (e.g., *Caroline ignored the proposal*). Participants wrote an ending for each fragment to form a complete sentence. At least 105 students provided a completion for each word of interest.

Corpus data were collected by extracting sentences containing the singular noun form or the past tense verb form from the Linguistic Data Consortium's 1987–1989 *Wall Street Journal* corpus. At least 100 usable tokens were obtained for most words, beginning with the 1987 portion of the corpus.

Each of the collected sentence tokens (corpus or completion) was coded for the complement(s) of the word of interest. Cases in which the word was a different grammatical category or clearly a different sense than the one intended, cases in which the sentence was ungrammatical or globally ambiguous, and cases in which the noun was a modifier (rather than the head) in a noun-noun compound were all excluded. From this coding, a %SC measure was obtained for each verb and noun from each of the completion and corpus sources. %SC was measured as the number of sentential complements (finite or infinitival, with or without the complementizer *that*) out of the total number of complements for the word from the particular source.[1] In most cases, we will discuss analyses based on both the completion-survey and corpus sources (see, e.g., Argaman et al., in preparation; Merlo, 1994; Roland & Jurafsky, 1997, 1998, for comparisons between these sources).

Verb-noun correlations

As described above, argument structures can be mapped, or even reduced, to semantic representations, so that words with multiple possible argument structures are assumed to have multiple semantic representations. Although these semantic representations may be closely related, the differences between them will correspond to differences among the events (in the case of verb semantics) that the word can describe. As a result, differences in the frequencies of events in the world could be responsible for differences in the frequencies of semantic representations, which would in turn correspond to differences in argument

[1] By counting each complement independently, we are effectively assuming that combinations of arguments have no special status. This may be an oversimplification, which future work will have to address.

structure frequencies. Thus, words with similar semantics will tend to be used in similar situations and will therefore tend to have similar frequency biases.

We can begin to examine this possibility, then, by comparing words with very similar semantics, to look for similarities in argument structure biases. Our verb-noun pairs provide a good candidate set of words, because the members of each pair share substantial semantic information without sharing a grammatical category. For example, suppose that *admit* occurs more often with an SC (e.g., *He admitted that he stole the money*) than with a direct object (e.g., *He admitted his guilt*). If this preference for one argument structure over another reflects an underlying difference in the frequency of the events in the world describable by the two constructions, then for *admission*, which shares most of its semantics with *admit*, the same argument structure frequency biases should appear: *Admission* should more frequently appear with an SC (e.g., *the admission that he stole the money*) than with an argument corresponding to the verb's direct object (e.g., *the admission of his guilt*). Thus, correlations in %SC should be present across verb-noun pairs. Alternatively, if argument structure frequency biases are a matter of random variation, then the biases for a verb and a semantically-related noun should not be reliably related, and across pairs, no correlations are expected.

To examine these alternatives, we therefore compared corresponding verbs and nouns on %SC, computing separate correlations for the corpus and completion data sources.

Results and discussion

In the completion data ($N = 79$), the verb and noun %SC measures correlated significantly ($r = .52, p < .001$), and the same was true in the corpus data ($N = 21$, $r = .67$, $p < .001$). These results suggest that semantics can account for substantial variability in argument structure frequency and seem to contradict the notion that frequency biases are a matter of random variation.

However, because in our data each noun is necessarily derivationally related to its corresponding verb, an alternative explanation is that verb frequency biases do vary randomly, independent of semantics, and the frequency biases for a noun are just mapped or copied from the corresponding verb. This is probably an unlikely account, given claims that derived nouns like those in our set have idiosyncratic properties which are not derivable from the corresponding verbs (e.g., Chomsky, 1970), but there is no direct evidence about this possibility. Thus, to investigate further the potential influence of semantics, we conducted two additional sets of analyses. First, we considered whether semantic category could predict frequency biases. Second, we examined the possibility that morphological categories (e.g., *-s/tion*, *-ing*, *-ment*), instead of semantic categories, might underlie argument structure bias – like many semantic properties, morphological category is shared by members of a verb-noun pair.

Verb semantic categories

A different way to look for semantic influences on argument structure frequencies is to compare groups of verbs that belong to coherent semantic categories. Assuming that semantically-similar verbs refer to similar types of events, we would expect effects of semantic category on argument structure frequency bias such that biases within a category should tend to be similar.

We tested this possibility using two different semantic categorization schemes: Levin's (1993) and Wierzbicka's (1987). Levin's categorization scheme is based on a study of English verb argument structure alternations and focuses on noun phrase and prepositional phrase complements. It is based on the assumption that coherent classes of verbs can be identified in terms of shared meaning components and corresponding syntactic behavior. Wierzbicka's classification is presented as a dictionary of speech act verbs and focuses on a smaller number and range of verbs, but she proposes a comprehensive system of explicating meaning, making use of citations, collocations, pragmatic properties, and syntactic properties. Thus there are some substantial differences between the schemes, most notably in coverage and in the information used to form categories: Levin considers more than 3000 verbs but includes SC-taking verbs only incidentally, whereas Wierzbicka largely focuses on SC-taking verbs; and Levin focuses on categorizing verbs with respect to their argument-structure-taking properties, whereas Wierzbicka's approach to categorization considers a broader class of information sources. Both schemes are concerned with argument structure alternatives, but neither of the schemes considers argument structure frequency.

For each categorization scheme, we analyzed the verb completion and corpus data for those verbs which (1) were explicitly listed in the categorization, (2) belonged to exactly one category, and (3) were in categories for which we had data from at least two verbs. To determine whether semantic category could predict argument structure frequency biases, we performed between-verb ANOVAs with %SC as the dependent variable and semantic category as the independent variable. The ANOVAs should reveal significant differences between categories only if semantic category captures frequency bias information.

Results and discussion

For the completion data, 43 verbs from 12 of Levin's (1993) semantic categories satisfied the selection criteria and were included. The number of verbs in each category ranged from 2 to 7. Figure 1 presents the %SC values for each of the verbs, organized by semantic category, and shows that verbs within a given category do tend to cluster together. This was supported by a reliable effect of semantic category ($F(11, 31) = 3.63, p < .01, \eta^2 = .56$). For the corpus data, 21 verbs from 6 of Levin's categories were usable, and this ANOVA also revealed a reliable effect of category ($F(5, 15) = 3.02, p < .05, \eta^2 = .50$).

Wierzbicka's (1987) scheme yielded 45 verbs with completion data satisfying the selection criteria, in 15 semantic categories. The number of verbs in each category ranged from 2 to 5. However, the ANOVA using Wierzbicka's categories revealed no effect of category ($F(14, 30) = 1.55, p > .15, \eta^2 = .42$);

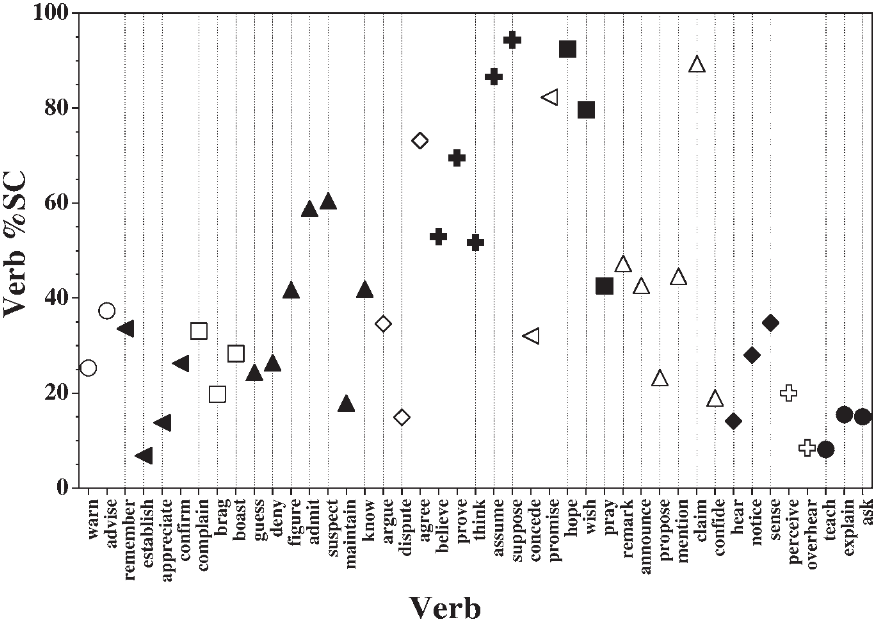


Figure 1. Verb %SC by Levin (1993) categorization (completion data). Different semantic categories are indicated by different open or shaded symbols.

and in the corresponding corpus ANOVA (with 23 verbs in 8 categories), the effect of semantic category was also non-significant ($F(7, 15) = 1.41$, $p > .25$, $\eta^2 = .39$).

The semantic categorization proposed by Levin (1993) yielded significant differences between categories in terms of %SC, suggesting that verb semantic class does play a role in accounting for argument structure biases. At least part of the reason that only Levin's scheme, and not Wierzbicka's (1987), captured significant variance in biases is probably that the former relies partially on argument alternations (as expressed in the syntax) to identify semantic verb categories. That is, in some cases, one of the decision criteria for placing a verb in a particular category is that it displays the same pattern of permitting or not permitting various syntactic alternatives as other members of the category. Wierzbicka notes that syntactic similarity is likely to reflect semantic similarity but does not rely on syntactic properties in categorizing the verbs. Instead, the categories (as opposed to the individual verb semantic descriptions) are semi-arbitrary, in that they only reflect some of the semantic relationships between the verbs, rather than all of them.

Levin's (1993) partial reliance on syntactic properties for categorization does present a potential confound, in that at least to some extent, it may be the syntactic properties that allow the ANOVAs to capture differences in argument structure bias. While this is a concern, her reliance on syntactic properties is limited to noting permitted argument structures; it does not give any consideration to the relative frequencies of the alternatives. The great majority of our verbs in her categories permitted approximately the same set of argument structures, and so most of the differences captured in the ANOVAs were unlikely to be just a matter of differences between which argument structures were allowed or disallowed.

If semantic category can at least partially account for argument structure frequency bias, it is then important to see if finer-grained semantics (i.e., differences within a semantic category) play a role in determining biases as well. One interpretation of Stevenson and Merlo's (1997) proposal, for example, would predict that this would not be the case: For at least some semantic categories (analogous to unergatives on their proposal), being a member of the category is sufficient to determine argument structure preferences, and thus frequency variations between members within a category should not reflect anything other than random variation. More generally, if semantic category captures frequency bias variation, but within-category semantic differences do not, then there would be reason to question either (1) the reliance on argument structure frequency as a primary variable in explaining comprehension (because it could be replaced by semantic category without loss of coverage), or (2) the use of fine-grained frequency information in particular, as opposed to gross differences in frequency biases as reflected in semantic categories. Obviously, the latter of these would be less serious for most theories, particularly given that current methods for determining argument structure biases can only approximate underlying biases anyway.

Correlations within semantic category

In order to see whether finer-grained semantics play a role in accounting for argument structure frequency biases, we selected two of Levin's (1993) semantic categories, *Conjecture* verbs (p. 183) and *Say* verbs (pp. 209–210), for which we had the largest number of verb-noun pairs with data. We used verbs explicitly listed by Levin, along with additional verbs judged by the authors to belong to these categories. We excluded verbs that were in both categories. Table 1 lists the verbs included in the analyses, as well as examples of some of the argument-taking properties of the two categories.

To examine the influence of fine-grained semantics, we computed correlations across verb-noun pairs for the %SC measure, within each category, as in the above analyses using the full dataset. Although verbs within each category share many properties, they also differ to some degree in their semantics. For example, while verbs in the *Say* category are all "verbs of communication of propositions and propositional attitudes" (Gropen, Pinker, Hollander, Goldberg, & Wilson, 1989), they vary in the propositional attitude they specify (cf. *claim*, *declare*, *suggest*). In a theory like Pinker's (1989; also, e.g., Jackendoff, 1990; Levin & Rappaport Hovav, 1995), fine-grained differences among different attitudes would not be relevant to determining whether a particular verb allows a particular argument structure; only semantic category would be. This might be the case for argument structure frequency as well, in which case there should not be a reliable correlation across verb-noun pairs within semantic categories. On the other hand, significant correlations within the categories would indicate that beyond the role of semantic category in accounting for frequency biases, differences within the categories also exist and can predict additional variation in biases.

Table 1. Analyzed verbs and syntactic properties of Levin's (1993) *Conjecture* and *Say* categories

Conjecture Verbs

assume[a,b]
believe[a,b]
conclude[a,b]
decide[a,b]
discover[b]
estimate[a]
expect[a]
feel
guarantee
guess
infer[a]
know[b]
realize[a]
recognize
sense[a]
suspect
understand[a]

*George assumed Jane the murderer.
*George assumed Jane as the murderer.
George assumed Jane to be the murderer.
George assumed *(to Susan) that Jane was the murderer.

Say Verbs

acknowledge[a,b]
announce[b]
claim[b]
concede[a]
confirm[a,b]
declare[b]
emphasize[a]
hint[a]
imply[a,b]
indicate[a,b]
insist[a]
mention
propose
prove[a,b]
remark
repeat
report
reveal
suggest[b]

*George mentioned Jane the problem.
George mentioned the problem to Jane.
George mentioned (to Jane) that Susan was happy.
*George mentioned to Jane.
*George mentioned about Jane.

[a] Added to Levin's categories based on the authors' intuitions.
[b] Included in corpus data correlations.

Results and discussion

Figure 2 shows the correlation between the verb and noun %SC measures for the *Conjecture* and *Say* categories, based on the completion-survey data. For both categories, the correlation was marginal (*Conjecture*: $N = 17$, $r = .48$; *Say*: $N = 19$, $r = .41$). For the corpus analyses, only 6 verb-noun pairs in the *Conjecture* category had data, and although a numerically large correlation was present, it was not reliable ($r = .66$, $p > .15$). In the *Say* category, 9 verb-noun pairs had corpus data, and a substantial correlation was present ($r = .86$, $p < .01$).

These results provide some evidence for effects of finer-grained semantics within Levin's (1993) semantic categories. In both the *Conjecture* and *Say* categories, the verb-noun correlations were numerically comparable to those computed for the overall set of verb-noun pairs. A z-test comparison of the *Conjec-*

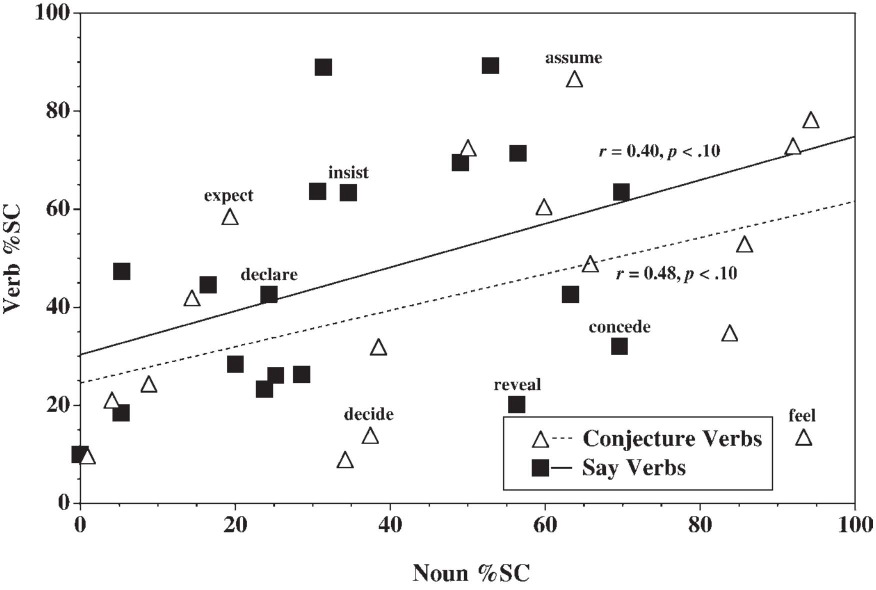


Figure 2. Verb %SC versus noun %SC for Levin's (1993) *Conjecture* (N = 17) and *Say* (N = 19) semantic categories (completion data).

ture completion-data correlation to the corresponding correlation computed for the full set of verb-noun pairs other than those in the *Conjecture* category (r = .64) revealed that the two did not differ reliably (z = –.78, p > .40). The same was true for the *Say* category (overall correlation excluding the *Say* verbs: r = .61; z = –.95, p > .30). These results thus suggest that while fine-grained semantic differences may not be relevant for determining what the permitted argument structures are for a word, they do play a role in determining how often those argument structures are used.

Morphological categories

The analyses relating verbs to corresponding nouns have so far assumed that the relationship is a matter of semantic properties. However, semantics is not the only possible connection: The pairs are explicitly related by a derivational morpheme (e.g., *-ment* in *argue-argument*, or often (zero), as in *report-report*), and the members of each pair share a stem (e.g., *argu-*, *report*), which yields phonological and orthographic overlap as well. We can therefore investigate whether these other properties might be able to account for argument structure frequency biases.

Derivational morphemes in particular can be responsible for a variety of semantic and/or syntactic effects (e.g., Aronoff, 1976). Randall (1984, 1988), for example, argues that nominalization systematically affects argument structure in a variety of ways depending on the morpheme involved. In many cases, one or more arguments permitted by the verb are no longer available to the corresponding derived noun. On this view, verbs that nominalize in the same way should share argument structure properties, as should the nouns derived from them. To the extent that different morphemes have differing effects on argument structure, ANOVAs comparing the different morphemes would be expected to yield reliable differences in argument structure frequency biases.

Each verb-noun pair was categorized according to the morpheme used to derive the noun. The majority of pairs fell into one of six categories: *-ation* (e.g., *accuse-accusation*), *-s/tion* (e.g., *anticipate-anticipation*), *-a/ence* (e.g., *accept-acceptance*), *-ing* (e.g., *feel-feeling*), *-ment* (e.g., *agree-agreement*), and zero-derived (e.g., *report-report*). These categories were used as the independent variable in completion-data ANOVAs like those for the semantic categorization schemes above, to determine if morphological category could account for any of the variation in argument structure frequency bias.

In addition, we selected two categories with a large number of pairs, in order to look for correlations within morphological category (as for the *Conjecture* and *Say* semantic categories). The first of these was the zero-derived category; the second (hereafter *-ion*) was formed by combining the *-ation* and *-s/tion* categories. These correlations will permit an additional examination of potential differences between morphological categories.

Finally, we also compared the verb-noun %SC correlation in the zero-derived category to that in the rest of the pairs, the latter of which were all related by some overt phonological and orthographic alteration. This comparison allowed us to examine the possibility that the actual surface form overlap between the verbs and nouns is what causes their argument structure correlation. If so, the zero-derived category should show stronger verb-noun correlations than the other morphological types, because only in the zero-derived cases do the verb and noun share the maximum amount of surface form.

Results and discussion

Sixty-five verbs in the six morphological categories had completion data and were included in the between-verb ANOVA. Figure 3 shows the %SC values

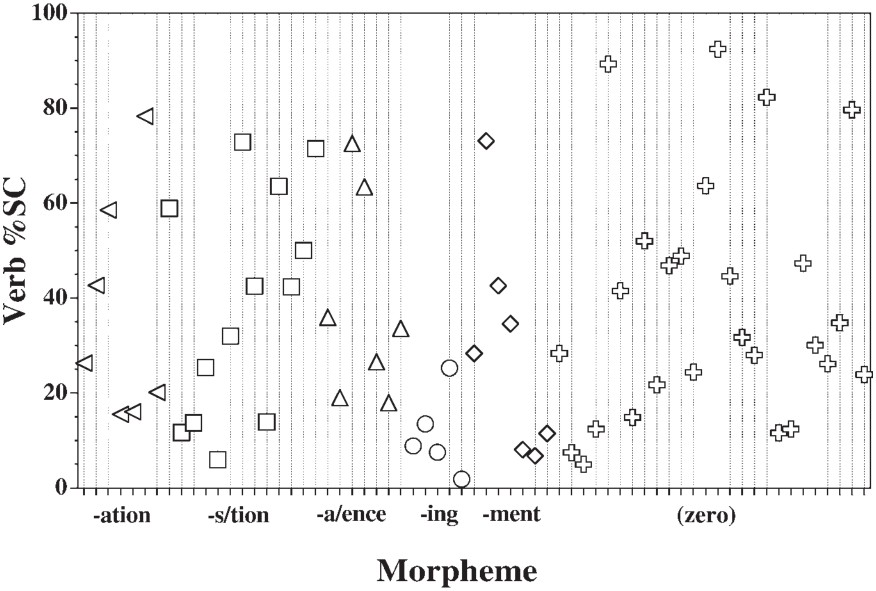


Figure 3. Verb %SC by morphological category (completion data). Different morphological categories are indicated by different symbols.

for the verbs in each category. In contrast to Figure 1 showing significant differences in Levin's (1993) semantic categorization, it is evident that within the morphological categories in Figure 3, the %SC values generally span the full range, and the categories do not appear to differ. The ANOVA confirmed this ($F(5, 59) = 1.30$, $p > .20$, $\eta^2 = .10$). A corresponding ANOVA on the noun completion data, for the 88 nouns with data in the same six morphological categories, also revealed no significant differences ($F(5, 82) = 1.45$, $p > .20$, $\eta^2 = .08$). Thus morphological category did not account for variability in either the verb or the noun argument structure frequency measure.

Within the morphological categories, on the other hand, there was a clear relation between verb and noun %SC: In the zero-derived category ($N = 25$), the verb and noun %SC measures were reliably correlated ($r = .65$, $p < .001$), and the same was true in the *-ion* category ($N = 19$, $r = .48$, $p < .05$). In addition, the comparison of the verb-noun %SC correlation in the zero-derived pairs to the corresponding correlation in the overall set excluding the zero-derived cases ($N = 54$, $r = .49$, $p < .001$) revealed that the two correlations did not differ ($z = .94$, $p > .30$).

These analyses yielded no evidence that morphological category can account for argument structure frequency biases. Categorizing the verb-noun pairs according to their nominalizing morpheme revealed no effects of category for the verbs or for the nouns. These results are particularly interesting when contrasted with those above for Levin's (1993) semantic categories, which did reveal differences in argument structure biases between categories. Furthermore, the lack of a difference between the verb-noun correlation for the zero-derived category and the corresponding correlation for the rest of the pairs indicated that the amount of surface overlap in the forms was also not responsible for argument structure bias differences.

The fact that we found no effects of morphological category is somewhat surprising for the proposals that specific morphemes affect meanings and thematic roles in predictable ways (e.g., Randall, 1984, 1988), but it is possible that the work done by morphological categories in affecting argument structure biases was masked by the different semantic categories included in our sample. Such effects might be revealed if examined within a single semantic category. Morpheme-category ANOVAs conducted on the verbs in the *Conjecture* and *Say* semantic categories did not support this possibility, but the amount of data available may have been too limited to detect effects. Another possibility is that all of the morpheme categories for our verb-noun pairs happened to be ones that would be expected to behave identically. This might be the case for some of our categories: Randall (1988) argues that *-a/ence*, *-ment*, *-s/tion*, and *-ation* do behave equivalently. However, zero-derived forms are generally assumed to have a range of properties different from overtly-derived forms (e.g., Marantz, 1984), and *-ing*, too, on either its process interpretation (e.g., *The warning of the children by the teacher took forever.*) or its result interpretation (e.g., *The warning was posted on the wall.*), should differ from the other overt forms (Randall, 1988).

General discussion

The central goal of this work was to explore possible underlying sources of argument structure frequency biases. The results of these analyses implicate lexical semantic properties as the candidate for such a source: First, the overall correlations between verbs and corresponding nouns, which share substantial components of meaning, were reliable for both completion and corpus frequencies. This argues against the possibility that argument structure biases are the result of random variation reinforced over time. Second, between-verb

ANOVAs using semantic categories identified by Levin (1993) revealed that semantic category could account for substantial variance in argument structure frequency.

Third, these results contrast sharply with non-significant ANOVAs on morphologically-defined categories, which indicated that morphological properties could not account for argument structure frequency biases, at least for the cases we considered. Furthermore, the noun-verb correlation within the zero-derived category did not differ from the corresponding correlation for the rest of the verb-noun pairs. This provides evidence against the possibility that surface form properties could account for argument structure variation, and it also supports the idea that verbs and corresponding nouns are handled as independent lexical entries: If this were not the case, then we would expect verb-noun pairs which are more similar in surface form to be more closely connected and thus to be more similar in argument structure frequency.

Finally, verb-noun correlations conducted within Levin's *Conjecture* and *Say* semantic categories for both completion and corpus data revealed effects similar in magnitude to those in the complete set of data, and correlations within the zero-derived and *-ion* morphological categories revealed similar results. These patterns indicate that fine-grained semantics plays a role in determining argument structure frequency biases. Semantic category alone appears not to be sufficient, and in fact the effect of the category itself in the ANOVAs may just be an epiphenomenon of the finer-grained semantic differences. In order to determine this, it will be necessary to examine how well semantic category and finer-grained semantic properties each predict variation in another variable, such as comprehension difficulty in an ambiguity resolution experiment (e.g., Schütze & Gibson, 1998; Stevenson & Merlo, 1997).

Taken together, these results suggest that the property relevant to predicting argument structure frequency variation across our verb-noun pairs is lexical semantics. The mechanism for these effects is a combination of the idea adapted from Pinker (1989; Jackendoff, 1990) and others that argument structures are essentially partial semantic representations, and the claim from the lexical access and constraint-based lexicalist literatures that elements in the lexicon have associated frequencies (e.g., MacDonald et al., 1994; Morton, 1969; Rayner & Duffy, 1986; Trueswell & Tanenhaus, 1994). As a result of these properties, the structure of the world and its relative frequencies can determine which meanings (and therefore which argument structures) are more or less commonly used for a word, and this can in turn determine the frequencies which are maintained in the comprehension system.

Despite the potential value of these results, however, they leave open a wide variety of issues which will eventually have to be resolved in developing a full account of the influence of frequency in language processing. For example, we used a fairly narrow set of verbs (SC-taking, and only a subset of them). This allowed us to focus on a particular subset of argument structure differences and to examine a relatively direct connection to corresponding nouns, but similar investigations will have to look at other classes of verbs and other argument structure biases. It is possible that the effect of morphology is quite different for other verb classes, or that the influence of argument structure frequency is more limited than we have assumed (e.g., Stevenson & Merlo, 1997).

In addition, the analyses presented here obviously rely solely on observational data; experiments designed to address these issues will be an important further step. Relatedly, although we now have evidence that semantics is related to frequencies, these results do not address the question of what information is actually used in processing. Even if semantics is the underlying source, the frequency biases might themselves be stored as part of the lexical entry, they might be computed when necessary during processing, or they might just be an observed by-product of semantic properties. One possible way to examine both of these issues is to make use of a word-learning paradigm (e.g., Gropen et al., 1989) with adults, in combination with a processing task. This would allow factorial manipulation of novel words' semantic properties and argument structure frequency distributions, as well as measurement of the influence of frequency versus semantics, and so forth.

Another concern about the current results, which may eventually become critical, is that we have not yet provided any actual semantic representations, even though such representations are assumed to be the domain through which properties of the world have their influence on the language system. In particular, for the strongest version of our argument, it must be the case that the meaning of a given verb is different when it takes an SC argument than when it takes a direct object. Pinker's (1989) and Jackendoff's (1990) theories of lexical-semantic representation provide detailed discussions of this for some classes of verbs, but they do not provide coverage of the SC argument structures relevant for our verb-noun pairs, although it may be possible to extend them. Levin's (1993) semantic categories worked well in predicting argument structures, but she focuses on subjects, direct objects, and prepositional phrases, and on covering a wide range of verbs, rather than on providing detailed semantics for each category. Wierzbicka's (1987) approach also does not provide detailed semantic representations.

Perhaps the broadest caveat is that it is important to realize that we have attempted to justify only one kind of lexical frequency information (argument structure bias), and that a variety of other sources of information are also important in many theories. Our approach of pushing the source of frequency information out of the language system and into the world will certainly also apply to basic word frequency (e.g., how often the string *bank* is encountered) and to other meaning and sense ambiguities (e.g., the relative frequencies of the different meanings for *bank*, or the different senses of *paper*, as in a substance vs a single sheet vs a journal article). However, it is less clear whether such an approach will account for grammatical category biases (e.g., Juliano & Tanenhaus, 1994; MacDonald, 1993; Tabor et al., 1997), or head versus modifier biases (e.g., MacDonald, 1993). The alternatives in these ambiguities do not appear to map as clearly onto differential circumstances in the world. An account of these kinds of frequencies may therefore rest on historical random variation, or on a more complex interaction between the language processing system and the properties of the world.

With respect to other sources of constraint, as well, it is worth noting that these results provide a tight link between plausibility and argument structure frequency, which have been mostly considered as independent sources of constraint in sentence comprehension (e.g., Garnsey et al., 1997; MacDonald et al., 1994; cf. McClelland, St. John, & Taraban, 1989; Pearlmutter & MacDonald, 1995). On the current view, the two might both be a reflection of the a priori probability of a particular circumstance occurring in the world, and thus at least in the limit, they might be interchangeable with respect to comprehension (and/or production). Of course, this still leaves open the question of whether the comprehension system actually makes use of them as separate sources of information or not, a question which might be addressed, as described above, in a verb-learning paradigm.

Author note

We are deeply indebted to Susan Garnsey for allowing us to use her verb data, and we greatly appreciate the advice and comments of Vic Ferreira, Ted Gibson, Karen Humphries, Dan Jurafsky, Aurora Mendelsohn, Paola Merlo, and Janet Randall. We also thank Collin Chang, Elana Chodin, Sahoko Hama, Melanie Lotocky, Kimberly Mitchell, Elizabeth Myers, Patricia Pimentel, and Jesse Stiles for their help in constructing materials, running participants, and coding lots and lots of data. Support for this work was provided by NSF grant BCS-9729007.

References

Argaman, V., Pearlmutter, N.J., Randall, J.H., & Mendelsohn, A.A. (in preparation). Argument structure data for 141 sentence-complement-taking nouns. Manuscript in preparation.

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Boland, J.E., Tanenhaus, M.K., Garnsey, S.M., & Carlson, G.N. (1995). Verb argument structure in parsing and interpretation: Evidence from wh-questions. *Journal of Memory and Language, 34*, 774–806.

Chomsky, N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (eds), *Readings in English transformational grammar*, (pp. 184–221). Waltham, MA: Ginn.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language, 67*, 547–619.

Ferreira, F., & Henderson, J.M. (1990). Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 555–568.

Frazier, L. (1995). Constraint satisfaction as a theory of sentence processing. *Journal of Psycholinguistic Research, 24*, 437–468.

Garnsey, S.M., Lotocky, M.A., Pearlmutter, N.J., & Myers, E. (in preparation). Argument structure frequency biases for 100 sentence-complement-taking verbs. Manuscript in preparation.

Garnsey, S.M., Pearlmutter, N.J., Myers, E., & Lotocky, M.A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., & Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language, 65*, 203–257.

Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: MIT Press.

Juliano, C., & Tanenhaus, M.K. (1994). A constraint-based lexicalist account of the subject/object attachment preference. *Journal of Psycholinguistic Research, 23*, 459–471.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: Univ. of Chicago Press.

Levin, B., & Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, MA: MIT Press.

MacDonald, M.C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*, 692–715.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*, 157–201.

MacDonald, M.C., Pearlmutter, N.J., & Seidenberg, M.S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

Marantz, A. (1984). *On the nature of grammatical relations*. Cambridge, MA: MIT Press.

McClelland, J.L., St. John, M., & Taraban, R. (1989). Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4*, SI287–335.

Merlo, P. (1994). A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research, 23*, 435–457.

Mitchell, D.C. (1989). Verb guidance and other lexical effects in parsing. *Language and Cognitive Processes, 4*, SI123–154.

Mitchell, D.C., & Cuetos, F. (1991). The origins of parsing strategies. *Proceedings of the Current Issues in Natural Language Processing Conference*, Univ. of Texas at Austin, 1–12.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165–178.

Pearlmutter, N.J., & MacDonald, M.C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language, 34*, 521–542.

Pearlmutter, N.J., & Mendelsohn, A.A. (1998). Serial versus parallel sentence comprehension. Paper presented at the eleventh annual CUNY Sentence Processing Conference, New Brunswick, NJ.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Randall, J.H. (1984). Thematic structure and inheritance. *Quaderni di Semantica, 5*, 92–110.

Randall, J.H. (1988). Inheritance. In W. Wilkins (ed.), *Syntax and semantics, volume 21: Thematic relations* (pp. 129–146). San Diego, CA: Academic Press.

Rayner, K., & Duffy, S.A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*, 191–201.

Roland, D., & Jurafsky, D. (1997). Computing verbal valence frequencies: Corpora versus norming studies. Poster presented at the tenth annual CUNY Conference on Human Sentence Processing, Santa Monica, CA.

Roland, D., & Jurafsky, D. (1998). How verb subcategorization frequencies are affected by the way you measure them. Poster presented at the eleventh annual CUNY Conference on Human Sentence Processing, New Brunswick, NJ.

Schütze, C.T., & Gibson, E. (1999). Argumenthood and English prepositional phrase attachment. *Journal of Memory and Language, 40*, 409–431.

Simpson, G.B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin, 96*, 316–340.

Spivey-Knowlton, M.J., & Sedivy, J.C. (1995). Resolving attachment ambiguities with multiple constraints. *Cognition, 55*, 227–267.

Stevenson, S., & Merlo, P. (1997). Lexical structure and parsing complexity. *Language and Cognitive Processes, 12*, 349–399.

Tabor, W. (1995). Lexical change as nonlinear interpolation. In J.D. Moore & J.F. Lehman (eds), *Proceedings of the 17th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Tabor, W., Juliano, C., & Tanenhaus, M.K. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes, 12*, 211–271.

Tabossi, P., Colombo, L., & Job, R. (1987). Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research, 49*, 161–167.

Taraban, R., & McClelland, J.L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language, 27*, 597–632.

Traxler, M.J., Pickering, M.J., & Clifton, C., Jr. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language, 39*, 558–592.

Trueswell, J.C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language, 35*, 566–585.

Trueswell, J.C., & Tanenhaus, M.K. (1994). Towards a lexicalist framework of constraint-based syntactic ambiguity resolution. In C. Clifton, L. Frazier, & K. Rayner (eds), *Perspectives on sentence processing*, pp. 155–179. Hillsdale, NJ: Erlbaum.

Trueswell, J.C., Tanenhaus, M.K., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition, 19*, 528–553.

Wierzbicka, A. (1987). *English speech act verbs*. San Diego, CA: Academic Press.

Verb sense and verb subcategorization probabilities

Douglas Roland and Daniel Jurafsky
University of Colorado

The probabilistic relation between verbs and their arguments plays an important role in psychological theories of human language processing. Unfortunately, different methods of calculating verb subcategorization probabilities yield different results. We argue for the *Lemma Argument Probability* hypothesis; a proposal that a separate set of subcategorization probabilities are associated with each sense of a word in the mental lexicon. Our results suggest that the differences in observed subcategorization probabilities found between various corpora and psycholinguistic experiments can be explained by a probabilistic combination of these lemma probabilities with other probabilistic factors. These factors include the discourse cohesion effects of natural corpora, the default referent effects of isolated-sentence experiments, the prompt given in sentence production experiments, the effects of different genres on verb sense, and the effect of verb sense on subcategorization. While we have only explored verbal lemmas, we assume this claim also holds of other predicates such as adjectives and nouns.

1. Introduction

The probabilistic relation between verbs and their arguments plays an important role in psychological theories of human language processing. For example, Ford, Bresnan and Kaplan (1982) proposed that verbs like *position* have two lexical forms: a more preferred form that subcategorizes for three arguments (SUBJ, OBJ, PCOMP) and a less preferred form that subcategorizes for two arguments (SUBJ, OBJ). Many recent psychological experiments suggest that humans use these kinds of verb-argument preferences as an essential part of the process of sentence interpretation. (Clifton et al. 1984; Ferreira & McClure 1997; Garnsey et al. 1997; MacDonald 1994; Mitchell & Holmes 1985;

Boland et al. 1990; Trueswell et al. 1993). It is not completely understood how these preferences are realized, but one possible model proposes that each lexical entry for a verb expresses a conditional probability for each potential subcategorization frame (Jurafsky 1996; Narayanan and Jurafsky 1998).

Unfortunately, different methods of calculating verb subcategorization probabilities yield different results. Recent studies (Merlo 1994; Gibson et al. 1996; Roland & Jurafsky 1997) have found differences between syntactic and subcategorization frequencies computed from corpora and those computed from psychological experiments. Merlo (1994) showed that the subcategorization frequencies derived from corpus data were different from the subcategorization data derived from a variety of psychological protocols. Gibson et al. showed that experimental PP attachment preferences did not correspond with corpus frequencies for the same attachments. In addition, different genres of corpora have been found to have different properties (Biber 1988, 1993).

In an attempt to understand this variation in subcategorization frequencies, we studied five different corpora and found two broad classes of differences.

1. **Context-based Variation**: We found that much of the subcategorization frequency variation could be accounted for by differing contexts. For example the production of sentences in isolation differs from the production of sentences in connected discourse. We show how these contextual differences (particularly differences in the use of anaphora and other syntactic devices for cohesion) directly affect the observed subcategorization frequencies.
2. **Word-sense Variation**: Even after controlling for the above context effects, we found variation in subcategorization frequencies. We show that much of this remaining variation is due to the use of different senses of the same verb. Different verb senses (i.e. different *lemmas*) tend to have different subcategorization probabilities. Furthermore, when context-based variation is controlled for, each verb sense tends towards having unified subcategorization probabilities across sources.

These two sources of variation have important implications. One important class of implications is for cognitive models of human language processing. Our results suggest that the verb sense or *lemma* is the proper locus of probabilistic expectations. The *lemma* (our definition follows Levelt (1989) and others) is the locus of semantic information in the lexical entry. Thus we assume that the verb *hear* meaning 'to try a legal case' and *hear* meaning 'to perceive auditorily' are distinct lemmas. Also following Levelt, we assume that a lemma expresses expectations for syntactic and semantic arguments. Unlike Levelt and many others, our *Lemma Argument Probability* hypothesis assumes that each verb lemma contains a vector of probabilistic expectations for its possible argument frames. For simplicity, in the experiments reported in this paper we measure these probabilities only for syntactic argument frames, but the *Lemma Argument Probability* hypothesis bears equally on the semantic/thematic expectations shown by studies such as Ferreira and Clifton (1986) and Trueswell et al. (1994).

Our results also suggest that the subcategorization frequencies that are observed in a corpus result from the probabilistic combination of the lemma's expectations and the probabilistic effects of context.

The other important implication of these two sources of variation is methodological. Our results suggest that, because of the inherent differences between isolated sentence production and connected discourse, probabilities from one genre should not be used to normalize experiments from the other. In other words, 'test-tube' sentences are not the same as 'wild' sentences. We also show that seemingly innocuous methodological devices, such as beginning sentences-to-be-completed with proper nouns (*Debbie remembered...*) can have a strong effect on resulting probabilities. Finally, we show that such frequency norms need to be based on the lemma or semantics, and not merely on shared orthographic form.

2. Methodology

We compared five different sources of subcategorization information. Two of these are psychological sources; corpora derived from psychological experiments in which subjects are asked to produce single isolated sentences. We chose two widely-cited studies, Connine et al. (1984) (CFJCF) and Garnsey et al. (1997) (Garnsey). The three non-experimental corpora we used are all online corpora which have been tagged and parsed as part of the Penn Treebank project (Marcus et al. 1993): the Brown corpus (BC), the Wall Street Journal corpus (WSJ), and the Switchboard corpus (SWBD). These three all consist of connected discourse and are available from the Linguistic Data Consortium (http://www.ldc.upenn.edu).

Although both sets of psychological data consist of single sentence productions, there are differences. In the study by Connine et al. (1984), subjects were given a list of words (e.g. *charge*) and asked to write sentences using them, based on a given topic or setting (e.g. *downtown*). We used the frequencies published

Table 1. Approximate size of each corpus

Corpus	Token/Type	Examples per verb
CFJCF	5,400 (127 CFJCF verbs)	n ≅ either 29, 39, or 68
Garnsey	5,200 (48 Garnsey verbs)	n ≅ 108
BC	21,000 (127 CFJCF verbs) 6,600 (48 Garnsey verbs)	$0 \leq n \leq 2{,}644$
WSJ	25,000 (127 CFJCF verbs) 5,700 (48 Garnsey verbs)	$0 \leq n \leq 11{,}411$
SWBD	10,000 (127 CFJCF verbs) 4,400 (48 Garnsey verbs)	$0 \leq n \leq 3{,}169$

in Connine et al. (1984) as well as the sentences from the subject response sheets, provided by Charles Clifton. In the sentence completion methodology used by Garnsey et al. (1997), subjects are given a sentence fragment and asked to complete it. These fragments consisted of a proper name followed by the verb in the preterite form (i.e. *Debbie remembered* ________). We used the frequency data published for 48 verbs as well as the sentences from the subject response sheets, provided by Susan Garnsey.

We used three different sets of connected discourse data. The Brown corpus is a 1-million-word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc). The texts had all been published in 1961, and the corpus was assembled at Brown University in 1963–1964 (Francis and Kucera 1982). Because the Brown corpus is the only one of our five corpora which was explicitly balanced, and because it has become a standard for on-line corpora, we often use it as a benchmark to compare with the other corpora. The Wall Street Journal corpus is a 1-million word collection of Dow Jones Newswire stories. Switchboard is a corpus of telephone conversations between strangers, collected in the early 1990's (Godfrey et al. 1992). We used only the half of the corpus that was processed by the Penn Treebank project; this half consists of 1155 conversations averaging 6 minutes each, for a total of 1.4 million words in 205,000 utterances.

We studied the 127 verbs used in the Connine et al. study and the 48 verbs published from the Garnsey et al. study. The Connine et al. and Garnsey et al. data sets have nine verbs in common. Table 1 shows the number of tokens of the relevant verbs that were available in each corpus. It also shows whether the sample size for each verb was fixed or frequency dependent. We controlled for verb frequency in all cross-corpus comparisons.

Deriving subcategorization probabilities from the five corpora involved both automatic scripts and some hand re-coding. Our set of complementa-

Table 2. Raw subcategorization vectors for *hear* from BC and WSJ

hear	0	PP	Swh	Sfin	VPbrst	NP	NP PP	passive
BC	4	12	3	1	15	47	4	14
WSJ	0	17	3	5	13	56	10	10

tion patterns is based in part on our collaboration with the FrameNet project (Baker et al. 1998; Lowe et al. 1997). Our 17 major categories were 0, PP, VPto, Sforto, Swh, Sfin, VPing, VPbrst, NP, [NP NP], [NP PP], [NP Vpto], [NP Swh], [NP Sfin], Quo, Passives, and Other. These categories include only true syntactic arguments and exclude adjuncts, following the distinction made in Treebank (Marcus et al. 1993). We used a series of regular expression searches and *tgrep* scripts[1] to compute probabilities for these subcategorization frames from the three syntactically parsed Treebank corpora (BC, WSJ, SWBD). Some categories (in particular the quotation category Quo) were difficult to code automatically and so were re-coded by hand. Since the Garnsey et al. data used a more limited set of subcategorizations, we re-coded portions of this data into the 17 categories. The Connine et al. data had an additional confound; 4 of the 17 categories did not distinguish arguments from adjuncts. Thus we re-coded portions of the Connine et al. data to include only true syntactic arguments and not adjuncts.

We also hand tagged the data from seven verbs for semantic sense. We used the semantic senses provided in Wordnet (Miller et al. 1993). We collapsed across senses in the few cases where we could not reliably distinguish between the Wordnet senses. When there were more than 100 tokens of a verb in a single corpus, we coded the first 100 randomly selected examples. This sample size was chosen to match the maximum sample size in the psychological corpora.

The subcategorization frequencies for a verb can be treated as a vector in multidimensional space. This allowed us to use the cosine of the angle between the vectors (Salton & McGill 1983) as a measure of the agreement between the subcategorization frequencies of verbs in different corpora. Table 2 shows the vectors for the verb *hear* in the Brown corpus and in the Wall Street Journal corpus. Using Formula 1, the cosine of the two vectors shown in Table 2 is 0.98. For non-negative vectors, the cosine ranges from 0 (complementary distribution) to 1 (complete agreement).

To measure whether the differences shown in the cosine were significant, we performed a chi-squared test on the same vectors, collapsing low frequency categories into an *other* category.

$$\text{Cosine} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Formula 1. Cosine of two vectors, x and y.

3. Isolated sentence versus connected-discourse corpora

A portion of the subcategorization frequency differences are the result of the inherently different nature of single sentence production and connected discourse sentence production. This section will show that the single sentence/connected discourse opposition affects subcategorization through two general mechanisms: the use of discourse cohesion in connected discourse and the use of default referents in null context (isolated sentence production).

Discourse cohesion

The first difference between single sentence production and connected discourse involves discourse cohesion. Unlike isolated sentences, a sentence in connected discourse must cohere with rest of the discourse. Halliday and Hasan (1976) use the notion of cohesion to show why sentences such as "So we pushed him under the other one" sound odd as the start of a conversation. Because a large number of syntactic phenomena such as pronominalization, fronting, deixis, and passivization play a role in discourse coherence, we would expect these syntactic devices to be used differently in connected discourse than in single sentence production. In addition, to the extent that these syntactic phenomena affect subcategorization, we would expect sentences produced in isolation (such as in the Connine et al. and Garnsey et al. experiments) to have different subcategorization probabilities than sentences found in connected discourse, such as in the Brown corpus, the Wall Street Journal corpus, and the Switchboard corpus. Because we counted dislocated arguments and pronominalized arguments in the same categories as their non-dislocated and full NP counterparts, pronominalization and most kinds of movement do not affect our subcategorization frequencies. Two syntactic devices that do affect our subcategorization frequencies are passivization and zero anaphora.

The passive in English is generally described as having one of two broad functions: (1) de-emphasizing the identity of the agent and (2) keeping an

Table 3. Use of passives in each corpus

Data source	% passive sentences
Garnsey	—
CFJCF	0.6%
Switchboard	2.2%
Wall Street Journal	6.7%
Brown corpus	7.8%

undergoer topic in subject position (Thompson 1987). Because both of these functions are more relevant for multi-sentence discourse, one would expect that sentences produced in isolation would make less use of passivization. As shown in Table 3, we found a much greater use of the passive in all of the connected discourse corpora than in the isolated sentences from Connine et al.[2]

Zero anaphora also plays a role in discourse cohesion. Whether an argument of a verb may be omitted depends on factors such as the semantics of the verb, what kind of omission the verb lexically licenses, the definiteness of the argument, and the nature of the context (Fillmore 1969, 1986; Fraser and Ross 1970; Resnik 1996 *inter alia*). In one common case of zero anaphora, Definite Null Complementation (DNC), "the speaker's authority to omit a complement exists only within an ongoing discourse in which the missing information can be immediately retrieved from the context" (Fillmore, 1986). For example the verb *follow* licenses DNC only if the 'thing followed' can be recovered from the context, as shown in example (1). Because the referent must be recoverable from the context, this type of zero anaphora is unlikely to occur in single sentence production, where the context is limited at best.

(1) The shot reverberated in diminishing whiplashes of sound. Hush **followed**. (Brown corpus)

The lack of Definite Null Complementation in single sentence production results in single sentence corpora having a lower occurrence of the [0] subcategorization frame. For example the direct object of the verb *follow* is often omitted in the connected discourse corpora, but never omitted in the Connine et al. data set. By hand-counting every instance of *follow* in all four corpora, we found that every case of omission was caused by definite null complementation. The referent is usually in a preceding sentence or a preceding clause of the same sentence.

Table 4. The object of *follow* is only omitted in connected-discourse corpora (numbers are hand-counted, and indicate % of omitted objects out of all instances of *follow*)

Data source	**% [0] subcat frame**
Garnsey	—
CFJCF	0%
Wall Street Journal	5%
Switchboard	11%
Brown	22%

Table 5. Greater use of first person subject in isolated-sentences

Data source	**% first person subject**
Garnsey	—
CFJCF	40%
Switchboard	39%
Brown corpus	18%
Wall Street Journal	7%

Default referents

In connected discourse, the context controls which referents are used as arguments of the verb. In single sentence production tasks, there is no larger context to provide this influence. In the absence of such demands, one might expect the subjects to use a wider variety of arguments with the verbs. On the contrary, we observe that the subjects favor a set of default referents – those which are accessible in the experimental context, or which are prototypical arguments of the verb. We found three kinds of biases toward these default referents.

First, non-zero subjects of single sentence productions were more likely to be *I* or *we* than subjects in connected discourse. Presumably the participants tended to use themselves as the topic of the sentence since in a null context there was no topic under discussion. Table 5 shows that the single sentence production data has a higher use of first person subjects than the written connected discourse data. Note that the Switchboard corpus also has a higher use of first person subjects. This could reflect a tendency for the participants, who are talking to strangers, to use themselves as a topic, given the absence of shared background.

Second, VP internal NPs (e.g. NPs which are c-commanded by the subject of the verb) are more likely to be anaphorically related to the subject of the verb. This includes cases such as (2) where the embedded NP is co-referential with the subject, and cases such as (3) where the embedded NP and the subject are related by a possession or part-whole relationship. To simplify judgement of relatedness, we only counted co-referential pronouns and traces. We did not count inferentially related NPs.

(2) Tom$_i$ noticed that he$_i$ was getting taller. (Garnsey et al. data)

(3) Alice$_i$ prayed that her$_i$ daughter wouldn't die. (Garnsey et al. data)

By contrast, VP-internal NPs in the natural corpora were more likely to refer to referents other than the subject of the verb. This additional sentence-internal anaphora in the isolated sentences is presumably a strategy for avoiding sentences like (4) which require the creation of an additional referent that is not already present in the context.

(4) Alice prayed that Bob's daughter wouldn't die. (made up example)

Table 6 shows how often the subject was anaphorically related to a VP internal NP in a hand-counted random sample of 100 examples from each corpus.

Third, the objects in the single sentence production data were more likely to be *prototypical* objects. That is, subjects tended to use default, relatively predictable head nouns for the direct objects of verbs. For example, of the 107 Garnsey sentences with the verb *accept,* 12 (11%) had a direct object whose head nouns was *award*. In fact 33% of the 107 sentences had a direct object whose head was one of the most common four words *award*, *fact*, *job*, or *invitation.* By contrast, the 112 Brown corpus sentences used a far greater variety of objects; it would take 12 different object nouns to account for 33% of the 112 sentences. Furthermore, the most common Brown corpus objects were pronouns (*it*, *them*); no common noun occurred more than 3 times in the 112 sentences. A formal metric of argument prototypicality is the token/type ratio. The ratio of the number of object noun tokens to object noun types will be high when a small number of types account for a greater percentage of the tokens. Table 7 shows that the token/type ratio is much higher for Garnsey data set than for the Brown corpus.

Table 6. Use of VP-internal NPs which are anaphorically related to the subject

Data source	% related subject/NP
Garnsey	41%
CFJCF	26%
Wall Street Journal	15%
Brown corpus	12%
Switchboard	8%

Table 7. Token/type ratio for arguments of *accept*

Data source	Token count	Type count	Argument token/type ratio
Garnsey	107	54	2.0
CFJCF	—	—	—
Wall Street Journal	138	105	1.3
Brown corpus	112	86	1.3
Switchboard	15	14	1.1

These uses of default references can all be seen as a device that experimental participants use to avoid introducing multiple new referential expressions into the single sentences. Natural sentences are known to generally contain only one new (inactive) piece of information per intonation contour (Chafe 1987) or clause (Givon 1979, 1984, 1987).

This section has shown several different ways in which discourse context affects observed subcategorization frequencies. These effects suggest that a psychological model of subcategorization probabilities will need to control for such discourse context effects. These contextual effects also have a methodological implication. Because of the biases inherent in isolated sentence production, we should not expect results from such psychological experiments to directly match natural language use.

4. Other experimental factors

The previous section discussed context effects that distinguish isolated sentence corpora from connected discourse corpora. This section discusses a further experimental bias that is specific to the sentence completion task. In sentence completion, the participants are given a prompt consisting of a syntactic subject as well as a verb. The nature of this syntactic subject can influence the verb subcategorization of the resulting sentence. Indeed this fact explains the single largest mismatch between the Garnsey data set and Brown corpus data. The verb *worry* was the only verb in these two corpora with an opposite preference between direct object and sentential complement; in Brown *worry* was more likely to take a direct object, while in the Garnsey data set *worry* was more likely to take a sentential complement.

This reversal in preference was caused by the properties of two of the subcategorization frames of *worry*. In frame 1 below, *worry* takes an experiencer as a subject, and subcategories for a finite sentence [Sfin]. In frame 2 below, *worry* takes a stimulus as a subject, and subcategorizes for an [NP].

Table 8. Subcategorization of *worry* affected by sentence-completion paradigm

Subcategorizations of worry	% Direct object	% Sentential complement
Garnsey	1%	24%
BC	14%	4%

Table 9. Uses of *worry*

#	Frame	Example
1	[experiencer] worries [stimulus]	Samantha worried that trouble was coming in waves. (Garnsey)
2	[stimulus] worries [experiencer]	Her words remained with him, worrying him for hours. (BC)

In the Garnsey protocol, proper names (highly animate) were provided. This provides a bias towards the first use, since animate subjects are more likely to be experiencers than stimuli. All of the sentential complement uses in the Brown corpus data had a human/animate subject. In the direct object uses, only 30% of the subjects were animate. It is uncontroversial that the nature of the prompt in a sentence completion experiment affects factors such as whether the sentence will be active or passive. This analysis shows that the nature of the prompt has more subtle but equally important effect on how subjects will use a verb.

5. Different verb senses have different subcategorization frequencies

Much work on subcategorization frequencies assumes implicitly that these frequencies were indexed by the orthographic word. Presumably this is because in many cases (e.g. Connine et al. (1984) and Garnsey et al. (1997)) these frequencies were collected to use in norming reading studies. Since we are making a psychological claim about the locus of frequency effects in the mental lexicon, the orthographic word assumption may not be a good one. Indeed, linguists have long suggested that the *lemma* or *sense* of a word is the locus of subcategorization; for example Green (1974) showed that two different senses of the verb *run* had different subcategorizations. Indeed, since Gruber (1965) and Fillmore (1968), linguists have been trying to show that the syntactic subcategorization of a verb is related to the semantics of its arguments. Thus one might expect a verb meaning *accuse* to have a different set of syntactic properties than a verb meaning *bill*. Similarly, if two senses of a single verb mean *accuse* and *bill*, these two senses should have different syntactic properties. The notion of a semantic base for subcategorization probabilities is consistent with work such as Argaman et al. (1998), which shows that verbs and their nominalizations have similar subcategorization preferences.

We propose that this fact about possible subcategorizations is also a fact about subcategorization *probabilities*, as the *Lemma Argument Probability* hypothesis:

Lemma Argument Probability hypothesis: The lemma or word sense is the locus of argument expectations. Each lemma contains a vector of probabilistic expectations for its possible syntactic/semantic argument frames.

We give a four-step argument for the *Lemma Argument Probability* hypothesis. In this section we start by showing that different corpora can yield different subcategorization probabilities. We show that different corpora contain different senses of verbs. We then show that it is this different distribution of lemmas or senses that accounts for much of the inter-corpus variability in subcategorization frequencies. Finally, in Section 6, we show a specific example of how when context-based variation is controlled for, each verb sense has a unified subcategorization probability vector across sources.

In order to investigate the relationship between verb sense and verb subcategorization, we hand coded the data for six verbs for sense/lemma. We primarily compare the data from the Brown corpus and the Wall Street Journal corpus since these two corpora had the largest amount of data. Although the data from the other corpora was less plentiful, it still provided useful insights.

First, we analyze three verbs, *pass*, *charge*, and *jump*, which were chosen because they had large differences in subcategorization frequencies between the Wall Street Journal corpus and the Brown corpus. Table 10 shows that all three verbs have significant differences in subcategorization frequencies between the Brown corpus and the Wall Street Journal corpus.

Next, we measured how often each sense occurred in each corpus. We found that each of the verbs showed a significant difference in the distribution of senses between the Brown corpus and the Wall Street Journal corpus,

Table 10. Agreement between WSJ and BC data

Verb	Cosine (all senses combined)	Do BC and WSJ have different subcategorization probabilities?
pass	0.75	Yes ($X^2 = 22.2$, $p < .001$)
charge	0.65	Yes ($X^2 = 46.8$, $p < .001$)
jump	0.50	Yes ($X^2 = 49.6$, $p < .001$)

as shown in Table 11. This is consistent with Biber et al. (1998), who note that different genres have different distributions of word senses.

Table 12 uses the verb *charge* to show how the sense distributions are different for a particular verb. The types of topics contained in a corpus influence which senses of a verb are used. Since the Brown corpus contains a balanced variety of topics, while the Wall Street Journal corpus is strongly biased towards business related discussion, we expect to see more of the business-related senses in the Wall Street Journal corpus. Indeed we found that the two business related senses of *charge* (*accuse* and *bill*) are used more frequently in the Wall Street Journal corpus, although they also occur commonly in the Brown corpus, while the *attack* sense of charge is used only in the Brown corpus. The *credit card* sense is probably more common in corpora that are more recent than the Brown corpus.

We also found this effect of corpus topic on verb sense in the isolated sentence corpora. When topics such as *home*, *school*, and *downtown* were provided to the subjects in the Connine et al. sentence production study, subjects used

Table 11. Differences in distribution of verb senses between BC and WSJ

Verb	Do BC and WSJ have different distributions of verb sense?
pass	Yes (X^2 = 59.4, p < .001)
charge	Yes (X^2 = 35.1, p < .001)
jump	Yes (X^2 = 103, p < .001)

Table 12. Examples of common senses of *charge*

Senses of charge	BC %	WSJ %	Example of the senses of charge
attack	23%	0%	His followers shouted the old battle cry after him and **charged** the hill, firing as they ran. (BC)
run	8%	0%	She **charged** off to the bedrooms. (BC)
appoint	6%	4%	The commission is **charged** with designing a ten year recovery program. (WSJ)
accuse	39%	58%	Separately, a Campeau shareholder filed suit, **charging** Campeau, Chairman Robert Campeau and other officers with violating securities law. (WSJ)
bill	24%	36%	Currently the government **charges** nothing for such filings. (WSJ)
credit card	0%	2%	Many auto dealers now let buyers **charge** part or all of their purchase on the American Express card. . ..(WSJ)
TOTAL	100%	100%	

different senses of the verbs. For example, the school setting caused 5 out of 9 subjects to use the *test* sense of the verb pass. By contrast, the *test* sense was used only 2 times in 230 examples in the Brown corpus.

For each of these three verbs, we then examined the subcategorization frequencies for each sense. In each case, the relative frequency of the verb senses in each corpus resulted in a difference in the overall subcategorization frequency for that verb. This is due to each of the senses having separate subcategorization probabilities. Table 14 illustrates that different senses of the verb *charge* have different subcategorizations (examples of each sense are given in Table 12).

Further evidence that subcategorization probabilities are based on verb sense is provided by the fact that for two of the verbs, *pass* and *charge*, the agreement for the most common sense was better than the agreement for all senses combined. The third verb, jump, also shows improvement, but the single sense value is not significant. This is because the nearly complementary distribution of senses between the corpora results in low sample sizes for one of the corpora whenever only a single sense is taken into consideration. Table 15 shows that the agreement for the most common sense is better than the agreement for all senses combined. We attribute the remaining disagreement between the corpora to context and discourse based subcategorization differences.

We also examined three verbs with good agreement (*kill*, *stay*, and *jump* – Table 16) in overall subcategorization between the Wall Street Journal corpus and the Brown corpus data as a preliminary effort to see what factors might prevent subcategorization frequencies from changing between corpora.

Table 13. Uses of *pass* in different settings in the CFJCF sentence production study

	Movement	Test	Pass the buck
home	6	1	1
downtown	5	1	0
school	4	5	0

Table 14. Different senses of *charge* in WSJ have different subcategorization probabilities. Dominant prepositions are listed in parentheses after the frequency

Senses of charge	that-S	NP	NP PP[3]	Passive	Other
appoint	0%	0%	0%	4%	0%
accuse	18%	0%	12% (with)	24%	2%
bill	0%	9%	24% (for)	1%	1%
credit card	0%	0%	2% (on)	0%	0%

We would expect no changes in subcategorization (beyond context/discourse changes) in cases where 1) the verb only had one common sense, or 2) the multiple senses of a verb had similar subcategorizations. We found that all three verbs with high agreement did in fact have different distributions of sense between the corpora, as shown in Table 17. These verbs showed equally high agreement for their most frequent senses.

Why do certain sense differences not cause subcategorization differences? One factor is that senses that are very closely (polysemously or metaphorically) related, like the senses of *kill* and *stay*, tend to have similar subcategorization probabilities across corpora. However, contextual factors may combine with the subcategorization probabilities for the similar senses, resulting in different observed probabilities. For example, the verb *jump* has two senses related by metonymy, *leap* and *rise in price*. While these have similar possible subcategorizations, the actual distribution of these subcategorizations was very different in the Brown corpus and the Wall Street Journal corpus data, due to the discourse circumstances under which each of the senses was used. The information demands in the Wall Street Journal resulted in stock price jumps being given with a distance and stopping point (jumped five eighths to five dollars a share).

Table 15. Improvement in agreement when after controlling for verb sense

Verb	Cosine (all senses combined)	Cosine (most common sense)
pass	0.75	0.95
charge	0.65	0.80
jump	0.50	0.59

Table 16. Agreement between BC and WSJ data

Verb	Cosine (all senses combined)	Do BC and WSJ have different subcategorization probabilities? (X^2)
kill	1.00	No
stay	1.00	No
try	1.00	No

Table 17. Differences in distribution of verb sense between BC and WSJ

Verb	Do BC and WSJ have different distributions of verb sense?
kill	Yes ($X^2 = 26.9$, $p < .001$)
stay	Yes ($X^2 = 26.1$, $p < .001$)
try	Yes ($X^2 = 8.74$, $p < .025$)

This section has shown that different verb senses can have different subcategorization probabilities. It also showed that different corpora tend to have a different distribution of verb senses, and that this different distribution can result in overall subcategorization differences between the corpora. Showing that different senses have different subcategorizations is only part of the argument for the Lemma Argument Probability hypothesis. Section 6 will complete the argument by investigating one verb in detail and showing that a given sense/lemma has the same subcategorization probability vector across sources when we control for context-based variation.

This relationship between verb sense and subcategorization leads to an important methodological caveat as well: our psychological models and experimental protocols which rely on verb subcategorization frequencies must also take verb sense into account.

6. Evidence for the Lemma Argument Probability hypothesis

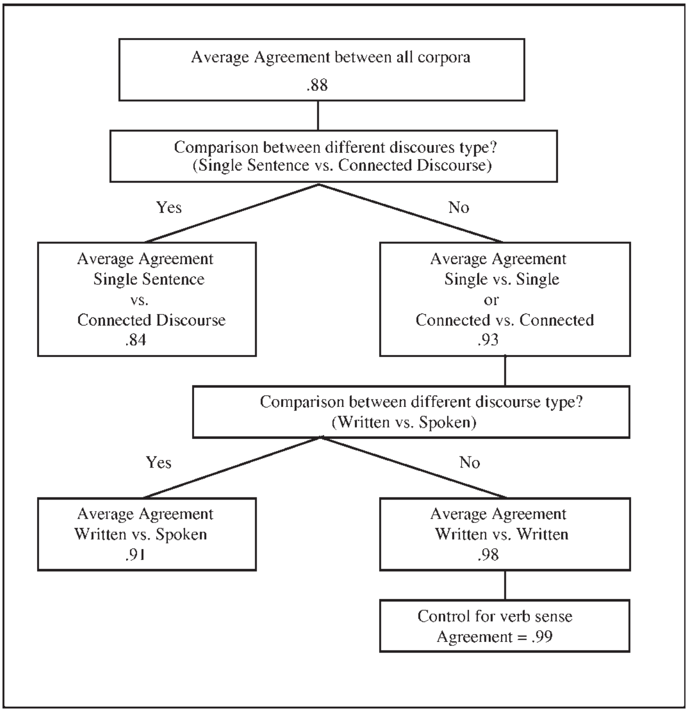
The previous section showed that different senses of a verb could have different subcategorizations. In this section we show preliminary evidence that a single sense tends to have a single subcategorization probability vector, when we control for other factors. We use data for the verb *hear*, which is one of the few verbs that appeared on all five corpora.

Our procedure is to show that the agreement between subcategorization vectors iteratively improves as we control for more factors, from .88 for agreement between uncontrolled vectors, to .99 for agreement between vectors controlled for verb sense as well as discourse context effects.

We began by calculating the average agreement between each of the 10 possible pairs of corpora. For example we compared the Brown corpus and the Wall Street Journal corpus, the Brown corpus and the Connine data set, the Brown corpus and the Garnsey data set, the Brown corpus and the Switchboard corpus, the Wall Street Journal corpus and the Switchboard corpus, and so on. The average agreement was .88.

We then controlled for the 'isolated-sentence' effect by *only* comparing pairs of corpora if they were *both* isolated-sentences or *both* connected sentences. Thus we compared the Garnsey data set to the Connine data set, the Brown corpus to the Wall Street Journal corpus, the Wall Street Journal corpus to the Switchboard corpus, and the Brown corpus to the Switchboard corpus. The average agreement improved to .93. We then controlled for spoken versus written effects by comparing only the Brown corpus and the Wall Street Journal

Table 18. Improvements in agreement for 'hear'



corpus. The average agreement improved to .98. Finally, instead of comparing all sentences with *hear* in the Brown corpus to all sentences with *hear* in the Wall Street Journal corpus, we compared only sentences which used the single most frequent sense of *hear*. The average agreement improved to .99. Table 18 shows a schematic of our comparisons. Note that although verb sense is controlled for only in the final step, controlling for sense results in improvement at any point in the chart. For example, the average agreement for all corpora also improves to .89 when we control for sense.

Unfortunately, this methodology does not allow us to assign factor weights to the relative contributions of verb sense and discourse context. While we had hoped to establish such weights, it now seems to us that such factor weights would be extremely dependent on the verb and the idiosyncrasies of the context.

7. Conclusion

We have shown that subcategorization frequency variation is caused by factors including the discourse cohesion effects of natural corpora, the default referent effects of isolated-sentence experiments, the prompt given in sentence production experiment, the effects of different genres on verb sense, and the effect of verb sense on subcategorization. Our evidence shows clearly that in clear cases of polysemy, such as the *accuse* and *bill* senses of *charge*, each sense has a different set of subcategorization probabilities. We have not investigated subtler differences in meaning, such as in *load the wagon with hay* and *load hay into the wagon*. Such alternations are usually modeled by one of two theories. Our data is currently unable to distinguish between them. For example, a Lexical Rule account (Levin and Rappaport Hovav 1995) might consider each valence possibility as a distinct lemma; our results merely show that these lemmas would have to be associated with lemma probabilities. An alternative constructional account (Goldberg 1995) would include both valence possibilities as part of a single lemma for load, with separate valence probabilities. In the constructional account, the shadings in sense are determined by the combination of lexical meaning and constructional meaning.

Our experiments do have a number of implications both for cognitive modeling and for psycholinguistic methodology. The *Lemma Argument Probability* hypothesis makes a psychological claim about mental representation: that each lemma contains a vector of probabilistic expectations for its arguments. While we have only explored verbal lemmas, we assume this claim also holds of other predicates such as adjectives and nouns. Furthermore, our results suggest that the observed subcategorization probabilities can be explained by a probabilistic combination of these lemma probabilities with other probabilistic factors. That is, the probability of linguistic events occurring "in the world" can be accounted for by probabilistic combinations of mentally represented linguistic knowledge. If this is true, it supports models of human language interpretation such as Narayanan and Jurafsky (1998) which similarly rely on the Bayesian combination of different probabilistic sources of lexical and non-lexical knowledge.

Acknowledgments

This project was supported by the generosity of the NSF via NSF IIS-9733067, NSF IRI-9704046, NSF IRI-9618838 and the Committee on Research and Creative Work at the graduate school of the University of Colorado, Boulder. Many thanks to Giulia Bencini, Charles Clifton, Charles Fillmore, Susanne Gahl, Susan Garnsey, Adele Goldberg, Michelle Gregory, Uli Heid, Paola Merlo, Laura Michaelis, Neal Pearlmutter, Bill Raymond, Philip Resnik, and two anonymous reviewers.

Notes

[1] We evaluated the error rate of our search strings by hand-checking a random sample of our data. The error rate in our data is between 3% and 7%. The error rate is given as a range due to the subjectivity of some types of errors. 2–6% of the error rate was due to mis-parsed sentences in Treebank, including PP attachment errors, argument/adjunct errors, etc. 1% of the error rate was due to inadequacies in our search strings, primarily in locating displaced arguments via the Treebank 1 style notation used in the Brown Corpus data.

[2] We also found that there were more passives in the written than in the spoken corpora, supporting Chafe (1992).

[3] The set of subcategorization frames that we use does not take the identity of the preposition into account.

References

Argaman, Pearlmutter, and Garnsey (1998). *Lexical Semantics as a Basis for Argument Structure Frequency Biases*. Poster presented at CUNY Sentence Processing Conference.

Baker, C.F., Fillmore, C.J., and Lowe, J.B. (1998). *The Berkeley FrameNet Project*. Proceedings of the 1998 COLING-ACL Conference, Montreal, Canada. 86–90.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D. (1993). Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics, 19*(2), 219–241.

Biber, D, Conrad, S., & Reppen, R. (1998). *Corpus Linguistics*. Cambridge University Press, Cambridge.

Boland, J.E., Tanenhaus, M.K., Garnsey, S.M. (1990). Evidence for the immediate use of verb control information in sentence processing. *Journal of Memory & Language, 29*(4), 413–432.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In Tannen, D. (ed.), *Spoken and Written Language*. Norwood, New Jersey: Ablex.

Chafe, W. (1987). Cognitive constraints on information flow. In Tomlin, R.S. (ed.), *Coherence and grounding in discourse* (1–16). Amsterdam: Benjamins.

Clifton, C., Frazier, L., & Connine, C. (1984). Lexical expectations in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 23*, 696–708.

Connine, C., Ferreira, F., Jones, C., Clifton, C., and Frazier, L. (1984). Verb Frame Preference: Descriptive Norms. *Journal of Psycholinguistic Research, 13*, 307–319.

Dowty, D. (1979). *Word meaning and Montague grammar*. Dordrecht: Reidel.

Ferreira, F. and Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348–368.

Ferreira, F., and McClure, K.K. (1997). Parsing of Garden-path Sentences with Reciprocal Verbs. *Language and Cognitive Processes, 12*, 273–306.

Fillmore, C.J. (1968). The Case for Case. In Bach, E.W. and Harms, R.T. (eds), *Universals in Linguistic Theory* (1–88). Holt, Rinehart & Winston, New York.

Fillmore, C. J. (1969). Types of lexical information. In Ferenc Kiefer (ed.), *Studies in Syntax and Semantics* (109–137). Dordrecht: Reidel.

Fillmore, C. J. (1986). *Pragmatically Controlled Zero Anaphora*. Proceedings of the 12th Annual Meeting of the Berkeley Linguistics Society (95–107). Berkeley, CA.

Ford, M. Bresnan, J., & Kaplan, R.M. (1982). A Competence-Based Theory of Syntactic Closure. In Bresnan, Joan (ed.), *The Mental Representation of Grammatical Relations* (727–796). Cambridge: MIT Press, 1982.

Francis, W. and Kucera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.

Fraser, B., and Ross, J.R. (1970). Idioms and unspecified NP deletion. *Linguistic Inquiry, 1*, 264–265.

Gahl, S. (1998). *Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus*. Proceedings of ACL-98, Montreal.

Garnsey, S.M., Pearlmutter, N.J., Myers, E. & Lotocky, M.A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language, 37*, 58–93.

Gibson, E., Schutze, C., & Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research, 25*(1), 59–92.

Givon, T. (1979). *On understanding grammar*. NY: Academic Press.

Givon, T. (1984). *Syntax: a functional/typological introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Givon, T. (1987). Beyond foreground and background. In Tomlin, R.S. (ed.), *Coherence and grounding in discourse*. Amsterdam: Benjamins.

Godfrey, J., E. Holliman, J. McDaniel. (1992). *SWITCHBOARD: Telephone speech corpus for research and development*. Proceedings of ICASSP-92, 517–520, San Francisco.

Goldberg, A.E. (1995). *Constructions*. Chicago: University of Chicago Press.

Green, G. (1974). *Semantics and Syntactic Regularity*. Bloomington: Indiana University Press.

Gruber, J. (1965). *Studies in lexical relations*. Bloomington: Indiana University Linguistics Club. [MIT Dissertation, 1965]

Halliday, M.A.K., and Hasan, R. (1976). *Cohesion in English*. London/New York Longman.

Juliano, C., and Tanenhaus, M.K. *Contingent frequency effects in syntactic ambiguity resolution*. In proceedings of the 15th annual conference of the cognitive science society, LEA: Hillsdale, NJ.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Levelt, W. (1989). *Speaking: from intention to articulation*. Cambridge: MIT Press.

Levin, B. and M. Rappaport Hovav (1995). *Unaccusativity at the syntax-lexical semantics interface*. Cambridge: MIT Press.

Lowe, J.B., Baker, C.F., and Fillmore, C.J. (1997). *A frame-semantic approach to semantic annotation*. Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics: Why, What, and How?" in conjunction with ANLP-97. Washington, D.C., USA.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9*, 157–201.

Marcus, M.P., Santorini, B. & Marcinkiewicz, M.A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Marcus, M.P., Kim, G. Marcinkiewicz, M.A., MacIntyre, R., Ann Bies, Ferguson, M., Katz, K., and Schasberger, B. (1994). *The Penn Treebank: Annotating predicate argument structure*. ARPA Human Language Technology Workshop (114–119). Plainsboro, NJ.

Merlo, P. (1994). A Corpus-Based Analysis of Verb Continuation Frequencies for Syntactic Processing. *Journal of Pyscholinguistic Research, 23*(6), 435–457.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). *Introduction to WordNet: an on-line lexical database*.

Mitchell, D.C. and V.M. Holmes. (1985). The role of specific information about the verb in parsing sentences with local structural ambiguity. *Journal of Memory and Language, 24*, 542–559.

Narayanan, S. and Jurafsky, D. (1998). Bayesian models of human sentencing processing. Procedings of 20th annual conference of the Cognitive Science Society. 752–757.

Resnik, Philip. (1996). Selectional constraints: an information-theoretic model and its computational realization. *Cognition, 61*(1–2), 127–159.

Roland, D. and Jurafsky, D. (1997). *Computing verbal valence frequencies: corpora versus norming studies*. Poster session presented at the CUNY sentence processing conference, Santa Monica, CA.

Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*, New York: McGraw-Hill.

Thompson, S.A. (1987). The Passive in English: A Discourse Perspective. In Channon, Robert & Shockey, Linda (eds), *In Honor of Ilse Lehiste/Ilse Lehiste Puhendusteos* (497–511). Dordrecht: Foris.

Trueswell, J.C., Tanenhaus, M.K., and Garnsey, S.M. (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.K., 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 Cognition, 19*(3), 528–553.

Author index

Item index