Volume 57, Number 3, October 2007

ISSN 0749-596X

ELSEVIER

Journal of Memory and Language

Available online at www.sciencedirect.com

ScienceDirect

Journal of Memory and Language 57 (2007) 348–379

Journal of Memory and Language

www.elsevier.com/locate/jml

Frequency of basic English grammatical structures: A corpus analysis ☆

Douglas Roland [a,*], Frederic Dick [b,c], Jeffrey L. Elman [c]

[a] Department of Linguistics, University at Buffalo, The State University of New York, USA
[b] Birkbeck College, University of London, UK
[c] Center for Research in Language, University of California, San Diego, USA

Received 3 February 2006; revision received 5 March 2007
Available online 24 May 2007

Abstract

Many recent models of language comprehension have stressed the role of distributional frequencies in determining the relative accessibility or ease of processing associated with a particular lexical item or sentence structure. However, there exist relatively few comprehensive analyses of structural frequencies, and little consideration has been given to the appropriateness of using any particular set of corpus frequencies in modeling human language. We provide a comprehensive set of structural frequencies for a variety of written and spoken corpora, focusing on structures that have played a critical role in debates on normal psycholinguistics, aphasia, and child language acquisition, and compare our results with those from several recent papers to illustrate the implications and limitations of using corpus data in psycholinguistic research.

Keywords: Corpus; Verb subcategorization; Frequency; Word order; Sentence processing

Introduction

Many recent models of language comprehension have stressed the role of distributional frequencies in determining the relative accessibility or ease of processing associated with a particular lexical item or sentence (Bybee, 1995; Kempe & MacWhinney, 1999; MacDonald, 1997, 1999; MacDonald, Pearlmutter, & Seidenberg, 1994; McRae, Jared, & Seidenberg, 1990; Mitchell, Cuetos, Corley, & Brysbaert, 1995; Plunkett & Marchman, 1993; Plunkett & Marchman, 1996; St. John & Gernsbacher, 1998). These approaches are known by a number of names—constraint-based, competition, expectation-driven or probabilistic models—but all have in common the assumption that language processing is closely tied to a user's experience, and that distributional frequencies of words and structures play an important (though not exclusive) role in learning.

This interest in the statistical profile of language usage coincides with two parallel developments in theoretical and computational approaches to language. An increasing

☆ This research was funded by MRC NIA G0400341, NIH/NIDCD5T32DC00041 and NIH/NIMHR01-MH60517. We wish to thank Ted Gibson, Dan Jurafsky, and several anonymous reviewers for extensive comments on earlier versions of this paper. We also thank Vic Ferreira and Florian Jaeger for helpful discussion, and David Temperley for sharing his data. Earlier versions of many of the tgrep search patterns were developed in the first author's dissertation.

* Corresponding author. Address: Department of Linguistics, University at Buffalo, The State University of New York, 605 Baldy Hall, Buffalo, NY 14260-1030, USA.
E-mail address: droland@buffalo.edu (D. Roland).

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.03.002

number of linguistic theories have shifted the locus of linguistic knowledge into the lexicon, partly in recognition of the lexical-specificity of many grammatical phenomena (e.g., Goldberg, 1995; Sag & Wasow, 1999). This emphasis has focused greater attention on actual patterns of lexical and grammatical usage, including distributional frequency. Secondly, there have appeared over the past two decades a number of statistically based natural language processing approaches to knowledge representation, processing, and learning. These include probabilistic/Bayesian models (e.g., Narayanan & Jurafsky, 1998, 2002), information theoretic approaches, and connectionist models (e.g., Christiansen & Chater, 1999; Sharkey, 1992). Here again, the actual statistical patterns of language play an important role.

As noted above, frequency-based and/or distributional analyses of some psycholinguistic phenomena are well established in the literature. The relationship between frequency and lexical access, for example, has been fairly extensively characterized (Bates et al., 2003; Snodgrass & Vanderwart, 1980). There is also a lively debate regarding the role played by construction frequency in online processing of sentences with temporary syntactic ambiguities (Cuetos & Mitchell, 1988; Desmet, De Baecke, Drieghe, Brysbaert, & Vonk, 2006; Desmet & Gibson, 2003; Gibson & Schütze, 1999; Gibson, Schütze, & Salomon, 1996; Gilboy, Sopena, Clifton, & Frazier, 1995; MacDonald et al., 1994; Mitchell et al., 1995).

If one is to test claims about the extent to which structural frequency plays an explanatory role in various psycholinguistic phenomena, it is crucial to have an accurate and reliable estimate of the frequency of the structures in question. It thus seems natural to attempt to derive such estimates from large-scale corpora of actual language usage. But with the exception of the above and a few other studies (e.g., Bybee, 1995; Kempe & MacWhinney, 1999), all of which focus on a narrow and very specific set of structures, corpus analyses have played a relatively small role in psycholinguistic research on higher-level language processing.

There are several obstacles to using corpus data to test psycholinguistic claims, including problems faced in extracting data of adequate quantity and quality, problems in deciding which types of corpora the data should be extracted from, and problems in deciding which levels of corpus data are most relevant for modeling the psycholinguistic phenomenon in question.

The quantity/quality problem occurs primarily because the relative frequencies of sentence structures are difficult to obtain (compared to lexical frequency, for example), since the corpora used to derive the distributions must be grammatically parsed in order to correctly identify syntactic structures. Parsing by hand is a labor-intensive task, so much so that it essentially precludes hand coding of large (e.g., >1,000,000 word) corpora. Automatic parsing makes the task tractable, but parser reliability has been a major issue. The difficulties posed by parser reliability are compounded by the fact that the error tends to be systematic rather than random. Additionally, the extraction procedures used by various authors are not necessarily transparent, and most psycholinguistics papers which rely on corpus data do not include sufficient detail about how the data was extracted (and indeed, the inclusion of such information is typically discouraged due to space limitations). As a result, reproducing corpus data can be difficult.

The problem of choosing which corpus or corpora to extract data from occurs because different corpora and genres of language use yield different probabilities for the same structures (e.g., Biber, 1988, 1993; Merlo, 1994; Roland & Jurafsky, 2002). In this paper, we find considerable variation in structural probabilities between corpora – with the largest differences typically occurring between the spoken and written corpora. While one solution might be to use the largest, most balanced corpus available, and hope that it most accurately reflects the overall language patterns of the typical participant in a psycholinguistic experiment, we suggest that the issue is more complicated. It is not the case that producers and comprehenders rely on different probabilistic grammars for written and spoken English. The differences in structural frequencies between corpora are the direct result of the way in which structures are used to accomplish specific linguistic goals in normal language use. Thus, we argue that it is not the structures themselves that vary in probability across corpora, but the contexts in which the structures are used. To the extent to which we can find discourse and contextual explanations for the observed frequencies of each structure, we provide an alternative to attributing the low frequency of a given structure purely to the complexity of that structure.

The third problem researchers face is in deciding which levels of corpus data are most relevant for modeling the psycholinguistic phenomenon in question—typically referred to as the issue of *grain size*. For example, in deciding whether *her goals* is a direct object or the subject of a sentential complement in *the athlete realized her goals*..., one could consider the probabilities of transitive vs. intransitive sentences, the probabilities of a direct object or sentential complement across all verbs where both are possible, the probabilities associated just with the verb *realize*, the probabilities associated with *realize* when the subject is animate, or the probabilities when the subject of *realize* is *athlete*.

Because of these problems, it has been quite difficult to test and/or falsify some of the predictions of sentence processing models whose proposed mechanisms are heavily influenced by distributional weighting. In this paper, we will provide a comprehensive set of structural frequencies for a variety of written and spoken corpora, focusing on structures that have played a critical role in debates on normal psycholinguistics, aphasia, and child language acquisition, and compare our results with those from several recent papers to illustrate the implications and limitations of using corpus data in psycholinguistic research. We will also address the issue of how and why structural probabilities vary across corpora. In addition, Appendix A provides an in-depth analysis of the error found in our data, and Appendix B provides descriptions of our verb subcategorizations and the issues faced in identifying the examples of each subcategorization. Finally, we provide a detailed description of the search patterns which we used in generating the data found in this paper as part of the online Supplementary materials accompanying this article, in hope that other researchers will both find them useful and contribute to the improvement of these search patterns.

Previous work

A variety of previous studies have attempted to determine structural frequencies. One method that has been used to estimate structural probabilities has been to run experiments in which participants generate language examples. This is done either using sentence production ("write a sentence using *answer*") or sentence completion ("complete the following—*Tom answered* ____"). Perhaps the best known of these studies is the sentence production study by Connine, Ferreira, Jones, Clifton, and Frazier (1984), which looked at a set of 127 verbs, and was coded for 16 possible subcategorizations. Much of the work using experimentally generated data has been done by researchers investigating the direct object/sentential complement ambiguity (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997; Holmes, Stowe, & Cupples, 1989; Kennison, 1999; Trueswell, Tanenhaus, & Kello, 1993), and thus, the data have tended to be limited to the 100 or so most frequent verbs that permit both direct object and sentential complement subcategorizations. Additionally, the data are usually coded in a direct object/sentential complement/other scheme, and thus provide little information about the other possible structures in which these verbs occur. For example, the verb *decide*, which is listed as a sentential complement biased verb in Garnsey et al. (1997), has more sentential complement uses than direct object uses in our British National Corpus data (21% of the examples are sentential complements, 6% are direct object examples, and 6% are passive), but 68% of the corpus examples would be classified as *other*. In fact, the most common subcategorization for the verb *decide* in the British National Corpus data is the *to* marked infinitival clause, such as in (1).

(1) But, by God, lady, you've **decided** [to survive]$_{\text{Infinitival Clause}}$. (British National Corpus)

One potential disadvantage of using such studies to generate structural probability data is that the exact task used in generating the data has a strong influence on the type of structures observed (see Roland & Jurafsky, 2002). Sentence completions of the form *Tom answered* ____ preclude the generation of passives, and the desire on the part of participants to minimize effort may artificially inflate the number of direct object completions, relative to other (longer) completions.

Other studies have attempted to derive structural frequencies from corpus data. Kempe and MacWhinney (1999) assembled sentence frequency counts from textbooks for learners of Russian and German, with samples sizes of 560 and 670 sentences respectively. Using sentence type ratios derived from these samples as parameter estimates in an instantiation of the Competition Model (Bates & MacWhinney, 1987), Kempe and MacWhinney were able to predict a number of behavioral outcomes. St. John and Gernsbacher (1998) reviewed several studies of the relative frequency of active and passive sentences as part of an analysis and simulation of aphasic patients' deficits in comprehending passive sentences. More recently, Thomas and Redington (2004) used frequency estimates of actives, subject clefts, object clefts, and passives in a neural network simulation of developmental and adult-onset language impairments, showing that the processing impact of sentence-type frequency is modulated by the network's developmental and learning trajectories, as well as by task demands.

With the advent of electronically distributed corpora, some researchers have also used automated tools for analyzing corpus frequency data. Biber and colleagues (Biber, 1988, 1993; Biber, Conrad, & Reppen, 1998) have extensively analyzed the differences between various registers of corpora such as fiction and academic writing and have found that many features of corpora differ between registers. The features they discuss range from syntactic (such as the use of passive, that-clauses, and to-clauses), to lexical (such as the distribution of similar-meaning words like *big*, *large*, and *great*) to discourse (such as the relative amount of new and given information). However, work on unparsed corpora has some limitations. For example, with the tools used by Biber, it was possible to compare frequencies of *that* and *to* clauses, since these clauses are marked by the presence of specific overt items, but it would be very difficult to accurately identify the number of *that*-less sentential complements, because there would be no easily identifiable item to search for. Other work, such as COMLEX (Grishman, Macleod, & Meyers, 1994), which provides detailed syntactic information about 38,000 words, relies on hand coding all data to avoid such limitations.

More recently, parsed corpora, such as those from the Penn Treebank (Marcus et al., 1994; Marcus, Santorini, & Marcinkiewicz, 1993), have made it possible to generate more reliable syntactic frequency information. A specific advantage of the Treebank data is that the parses were hand corrected to allow for a level of accuracy that is much higher than that which can presently be achieved by automatic methods. Much of the work with the Treebank data has looked at the frequencies of specific structures occurring with specific verbs. For example, Merlo (1994) examined the frequencies of five possible syntactic structures for 105 verbs. Her results were based on automatic counts, with overall frequencies adjusted using a hand corrected sub-set. Merlo also describes differences in verb subcategorization probabilities between her corpus-derived data and norming study data such as Connine et al. (1984).

Other work in this area (Roland, 2001; Roland & Jurafsky, 1998, 2002; Roland et al., 2000) has used Treebank data to generate subcategorization probabilities based on the set of 16 possible subcategorizations frames from Connine et al. (1984). These papers also compare data from a variety of corpora and norming studies, and attribute the differences in verb subcategorization probabilities to *context-based variation* and *word sense variation*. Hare, McRae, and Elman (2004) found that word sense variation plays an important role in accounting for the different outcomes of several recent experiments addressing the relationship between structural frequency and online sentence processing.

Previous work faced several limitations: the number of verbs covered, the number of structures covered, and limits on the amount of data available for low frequency items imposed by the size of the Treebank corpora. While some work (e.g., Lapata, Keller, & Schulte im Walde, 2001; Roland, Elman, & Ferreira, 2006) has used data from larger corpora, additional work is needed to investigate the reliability of the automatic extraction methods used in such papers.

Methodology

The core set of data used in this paper consists of a comprehensive set of structural frequency information which was automatically extracted from a set of the most commonly used corpora. The set of corpora used is described in section ‘Corpora used’, while the automatic extraction techniques are described in section ‘Structures included in analysis’. In addition to the large overall set of automatically extracted data, we also relied on several small-scale hand analyses of random samples of the data, either for error analysis or for providing estimates of distributional properties which could not easily be coded for using automatic methods.

Corpora used

In order to investigate the frequencies of different syntactic structures in the English language, a total of eight corpus data sets were used in this paper. These corpora were chosen to represent a variety of genres of written and spoken language. The syntactic structures in some of these corpora were determined through automatic means, while in others, the syntactic structures were hand corrected. These two methodologies involve a tradeoff between accuracy (highest for hand labeling, but still not perfect) and the ability to label a large sample of data (100 million words of text can be automatically parsed overnight using a cluster of 30 desktop PCs, circa 2003—in contrast with the estimate from Marcus et al. (1993) of 2.5 million words per year for a team of 5 part time annotators working 3 hours per day at hand correcting automatic parses).

Once the corpora are annotated for syntactic structure (either by us or by other researchers), exemplars of specific structures are extracted from the parsed corpora using search patterns (described in the online Supplementary materials) in conjunction with tree-structure-sensitive search tools such as *tgrep* (included in the Treebank corpus distribution) and *tgrep2* (http://tedlab.mit.edu/~dr/Tgrep2/). The search patterns and tools described in this paper can be used with either readily available syntactically annotated corpora such as those from the Penn Treebank Project (Marcus et al., 1994, 1993), or with text that has been parsed by any automatic parser which follows the Treebank format, such as the Charniak parser (Charniak, 1995, 1997) and the Collins parser (Collins, 1996, 1997, 1999).

Several of the data sets used in the paper came from the Penn Treebank Project (Marcus et al., 1994, 1993). These data were automatically parsed, and then hand corrected. These data are generally considered to be the most accurately labeled data available, and are regularly used as the gold standard in the development of automatic parsers. Specifically, the data in this paper rely on three different corpora: the Penn Treebank versions of the Brown corpus, the Wall Street Journal corpus, and the Switchboard corpus.

The Brown corpus (Francis & Kučera, 1982) is a one-million-word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc). The combination of texts included in the Brown corpus was intended to represent a wide cross-section of English usage, although it does consist entirely of written language samples. All included texts were published in or before 1961, and the corpus was assembled at Brown University in 1963–1964.

The Wall Street Journal corpus, which was parsed by the Penn Treebank Project, is a one-million-word subset of the DARPA WSJ-CSR1 collection of Dow Jones Newswire stories. As the name of the corpus implies, it consists entirely of written texts, primarily with a business focus. In this paper, we will rely primarily on the Treebank 2 version of this corpus, because it has the most detailed and accurate labeling (referred to in this paper as either Wall Street Journal or Wall Street Journal Treebank 2 data). In discussions on error analysis, we will also report on data taken from an additional version of the Wall Street Journal which was prepared by re-parsing the same 1 million words of text using the same parser we used for some of the other data sets described below (Charniak, 1995, 1997)—hereafter noted as Wall Street Journal-Charniak.

The final Treebank dataset reported on in this paper is the Switchboard corpus. Switchboard is a corpus of telephone conversations between strangers, collected in the early 1990's (Godfrey, Holliman, & McDaniel, 1992). Only the half of the corpus that was processed by the Penn Treebank project was used; this half consists of 1155 conversations averaging 6 min each, for a total of 1.4 million words in 205,000 utterances. Using the other portion of the Switchboard corpus would have required us to hand label the data.

Finally, we report on two sets of data taken from the British National Corpus (Burnard, 1995). The British National Corpus is a 100 million word sample of English taken from a variety of genres. It consists of 90% written text and 10% transcribed speech. This data set has the advantage of being much larger than any of the other sets we looked at. We automatically parsed this data using the Charniak parser (Charniak, 1995, 1997). We report separate numbers for the whole 100 million words and for the 10% subset of spoken language.

The three corpora (Brown, Wall Street Journal, Switchboard) chosen from Treebank were selected because they represent a variety of different genres and discourse typeswhile affording us the convenience and accuracy of the Treebank data. The British National Corpus data were used, because, though less accurate than the Treebank data, it allows us to look at items whose frequency of occurrence in the Treebank corpora would be too low to allow us to draw accurate generalizations. Because we expect the frequencies of different structures to vary across genre and discourse type (see Biber, 1988, 1993; Biber et al., 1998; Roland, 2001; Roland & Jurafsky, 1998, 2002; Roland et al., 2000), choosing different types of corpora allows us to have some indication of the degree to which the frequencies of the structures we are investigating vary across genre and discourse type. However, it is important to note that none of these corpora, either individually or in combination, can truly be considered to constitute an over-all representative sample of language. The choice of the British National Corpus, Brown, Wall Street Journal, and Switchboard represents a tradeoff between the quality and quantity of data and the psycholinguistic relevance of that data. Although it can be argued that (for example) Wall Street Journal text may not be the best model of a typical undergraduate subject in a psycholinguistic experiment, we still find the comparison between the Wall Street Journal data and the other data to be highly informative and relevant to the issue of the role of frequency in language processing.

Structures included in analysis

In order to examine the relative frequencies of different structures in English (i.e., the ordering of subject, object, and verb), we analyzed all instances of all verbs in each of the corpora. However, for ease of exposition, we primarily discuss three subsets of the data: cleft sentences, relative clauses, and the set of structures covered in the Connine et al. (1984) verb norming study. In addition, we report data for all 204 verbs used in the Connine et al. (1984) and Garnsey et al. (1997) studies as part of the online Supplementary materials accompanying this article. Appendix B contains more complete descriptions of each category, including discussion of issues related to defining each category and extracting examples for each category. The individual search patterns used (sometimes several per high-level target structure) and details about the commands used for determining the total number of sentences, words, noun phrases, and verb phrases in each corpus are provided as part of the online Supplementary materials accompanying this article.

Results and discussion

Cleft sentences

Although extremely infrequent, subject and object clefts (shown in Table 1) have figured prominently in a number of studies of language processing in normal adult populations, aphasic patients, and child language (Berndt, Mitchum, & Haendiges, 1996; Caplan & Waters, 1999; Dick et al., 2001; Dick, Wulfeck, Krupa-Kwiatkowski, & Bates, 2004; Ferreira, 2003; Gordon, Hendrick, & Johnson, 2001; Gordon, Hendrick, & Levine, 2002; Grodzinsky, 2000; Shapiro, Gordon, Hack, & Killackey, 1993; Warren & Gibson, 2005). For example, aphasic performance in processing subject clefts is often claimed to be superior to their processing of object clefts. It has been claimed that this difference cannot be due to differences frequency of usage, but rather to the deletion—in aphasic patients—of so-called trace elements that occur in object clefts but not subject clefts (Grodzinsky, 2000). Such claims assume there are no differences in frequency of the two structures, but this has not (until now) been tested empirically.

Table 1
Examples of cleft sentences from the Wall Street Journal

Structure	Example (from the Wall Street Journal corpus)
Subject cleft	It was Richard Nixon's first visit to China in 1972 that set in motion the historic rapprochement between Beijing and Washington
Object cleft	It's paper profits I'm losing

The frequencies of these structures were obtained for the Wall Street Journal and Switchboard corpora only. As Table 2 shows, the subject clefts in fact occur more frequently than do object clefts in both corpora. There are two ways of looking at the cleft structure frequencies. On one hand, the difference between subject and object cleft frequencies seems large. In the Wall Street Journal data, the subject cleft is more than 13 times more likely than the object cleft. This difference is well above the 5× difference in likelihood used in Jurafsky (1996) to explain garden path effects. On the other hand, when compared to the frequency of non-cleft sentences, both subject and object cleft structures are exceedingly rare, occurring in less than one tenth of a percent of all sentences. This means that the subject cleft, which is easier for aphasia patients to process, is much lower in frequency than the passive structure, which is more difficult. This suggests that in a frequency-based account, the absolute frequencies of individual structures alone cannot account for the patterns of difficulty observed.

In order to compare the data from the Wall Street Journal with that from Switchboard, we must first normalize the data, because the two corpora are different sizes. Traditionally, cross-corpus comparisons are made by normalizing the frequency of the item in question by the number of words in the corpus, and presenting the results in terms of the frequency of the target item per thousand or million words. This is an appropriate means for normalizing corpus sizes when looking at issues such as word frequency, but can result in misleading comparisons when the domain of the target item is at a higher level than that of the individual word. Because clefts are a sentence level phenomenon, we feel that it is more appropriate to normalize the number of cleft sentences by the number of sentences (rather than words) in each corpus. The impact of this decision can be seen in Tables 3 and 4, where we have normalized the number of subject and object clefts by corpus size in words and sentences, respectively. Note that the choice of normalization methods dramatically changes the relative frequencies of subject clefts in the two corpora. When normalizing by word count, subject clefts are more frequent in Switchboard, whereas they appear more frequent in Wall Street Journal when normalizing by sentence count. However, since the results of these two methods of normalization are based on fairly small corpus counts, the margin of error is correspondingly large.

Table 2
Raw counts of subject and object cleft counts

	Wall Street Journal	Switchboard
Subject cleft	40	12
Object cleft	3	0

Table 3
Subject and object cleft counts, normalized to corpus size of 1 million words

	Wall Street Journal	Switchboard
Subject cleft	32	38
Object cleft	2	0

Table 4
Subject and object cleft counts, normalized to corpus size of 1 million sentences

	Wall Street Journal	Switchboard
Subject cleft	813	577
Object cleft	61	0

Finally, although—*contra* previous claims—there is a large difference in the frequency of occurrence between subject clefts vs. object clefts, it is also true that when compared to the frequency of non-cleft sentences, both subject and object cleft structures are exceedingly rare, occurring in less than one tenth of a percent of all sentences. One possible interpretation of this fact is that it is relative frequency of occurrence, rather than absolute frequency of occurrence, that confers a processing advantage for a structure. Alternatively, it could be that the effects of frequency on processing are more complex than would be revealed by a simple tabulation of frequency of occurrence of a target structure. Note that although subject and object cleft structures are rare, these constructions share much in common with simple active declaratives and subject relative clauses, and with some interrogative and object relative clauses, respectively. These other constructions are much more frequent, and it is not unreasonable to suppose that exposure to those structures will affect the processing of cleft sentences. This latter possibility then implies that accounts of processing that appeal to frequency of usage will need to be sensitive to issues of granularity: What usage is relevant and how wide a net should be cast when one characterizes the frequency of a target structure?

Relative clauses

The processing of relative clauses has been one of the primary focuses of psycholinguistic research for almost half a century, from early work on generative grammars (e.g., Chomsky, 1956) to current debates on the time course of thematic role assignment (e.g., Dahan & Tanenhaus, 2004; Kamide, Altmann, & Haywood, 2003; Kutas & King, 1996; Rayner, Warren, Juhasz, & Liversedge, 2004). In particular, there has been increasing interest in the role of relative frequency of occurrence in establishing preferences or biases for relative clause disambiguation (Cuetos & Mitchell, 1988; Desmet et al., 2006; Gibson & Schütze, 1999; Gibson et al., 1996; Gilboy et al., 1995; MacDonald et al., 1994; Mitchell et al., 1995).

The distributional patterns of relative clauses are also relevant to models of language production. The factors governing the presence or absence of optional words, such as *that* in sentential complements and object relatives, has been used as evidence in the debate over the extent to which language production is driven by speaker-based production factors and the extent to which it is driven by comprehender based factors such as the desire to avoid potential ambiguity (e.g., Ferreira & Dell, 2000; Jaeger & Wasow, 2005; Temperley, 2003; Wasow, Jaeger, & Orr, 2005).

As will become clear in this section, the different corpora that we consider paint very different pictures of relative clauses. The corpora appear to vary greatly in the distribution of many of the factors that have been claimed to play a role in the processing of relative clauses—the comparative frequencies of subject and object relative clauses in each corpus, whether the modified noun is animate or inanimate (Traxler, Morris, & Seely, 2002; Traxler, Williams, Blozis, & Morris, 2005), whether the embedded noun phrase is a full noun phrase or pronominal (Reali & Christiansen, 2007; Warren & Gibson, 2002), whether the embedded noun phrase is a proper name (Gordon et al., 2001), or the quantified pronoun *everyone* (Gordon, Hendrick, & Johnson, 2004). Yet, as we hope this section will demonstrate, the observed distributions are the direct result of the discourse functions that the structures serve, and the discourse needs of each corpus genre or context. This section will address the distributional information found in the corpus data, the implications of this data for theories of relative clause processing and the implications for relative clause production.

The overall summary of the data relevant to these questions is presented in Tables 6 and 7, while the details are discussed in subsequent sections. Examples of each type of relative clause are shown in Table 5. As in the previous section, we must choose the appropriate unit for normalizing the data for corpus size. In this case, we feel that normalizing by the number of noun phrases in each corpus is the most appropriate method for comparing the frequencies of the relative clauses, because a relative clause can only occur given that a noun phrase exists. Table 6 shows the raw counts for each type of relative clause, and Table 7 shows the frequencies of each type of relative clause when normalized by the number of noun phrases in each corpus. The distribution of relative clauses across corpora is also depicted in Fig. 1.

As Fig. 1 shows, most types of relative clauses are considerably less frequent in spoken use than in written use. This suggests that the functions carried out by such relatives, such as distinguishing between multiple potential referents or providing additional information about a referent, are either carried out by other means—with the additional information potentially expressed in separate sentences (the spoken data tend to consist of shorter sentences with simpler structures than the written data)—or are less relevant—potentially due to the interactive nature of spoken discourse.

Subject relatives

Subject relatives are less common in spoken corpora than in the written corpora, as can be seen in Fig. 1. Beyond the overall difference in the frequency of subject relatives in the spoken and written corpora, the nature

Table 5
Examples of relative clauses from the Wall Street Journal

Structure	Example (from the Wall Street Journal corpus)
Subject relative	The researchers who studied the workers
Object relative (full)	The 25 countries that she placed under varying degrees of scrutiny
Object relative (reduced)	The foreign stocks they hold
Passive relative (full)	Delmont A. Davis, who was named president and chief operating officer in August
Passive relative (reduced)	A competing psyllium-fortified cereal called Heartwise
Infinitive subject relative	A computer system to map its waterworks
Infinitive object relative	A blind girl to cure
Infinitive passive relative	The last women to be executed in France

Table 6
Raw counts of various types of relatives in each corpus

	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
Subject relative	354,752	25,024	4622	760	5585
Object relative	73,614	9812	608	447	552
Object relative (reduced)	136,448	36,638	1460	423	1037
Passive relative	77,993	4392	882	24	375
Passive relative (reduced)	268,412	7332	3302	62	3918
Subject infinitive relative	118,211	4594	1306	193	2178
Object infinitive relative	80,763	5769	658	109	561
Passive infinitive relative	13,567	469	150	7	129

Table 7
Frequencies of each type of relative clause per 1 million noun phrases

Type of relative clause	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
Subject relative	14,182	9851	15,024	9548	18,229
Object relative	2943	3863	1976	5616	1802
Object relative (reduced)	5455	14,423	4746	5314	3385
Passive relative	3118	1729	2867	302	1224
Passive relative (reduced)	10,730	2886	10,733	779	12,788
Subject infinitive relative	4726	1808	4245	2425	7109
Object infinitive relative	3229	2271	2139	1369	1831
Passive infinitive relative	542	185	488	88	421
Total	44,924	37,015	42,218	25,440	46,788

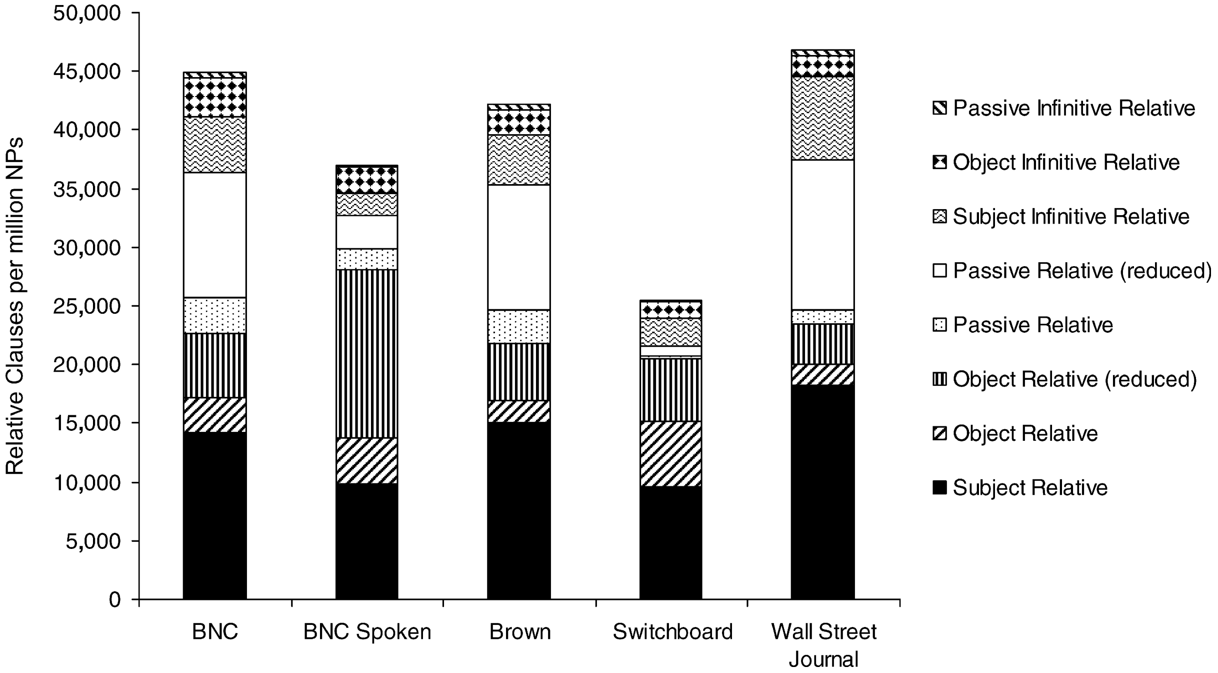

Fig. 1. Distributions of relative clauses in each corpus (per million noun phrases).

of the subject relatives found in these corpora are also different. The subject relative clauses in the spoken corpus data are more likely to modify noun phrases with a low content value, such as *someone*, *people*, or *something*, than are the subject relative clauses in written corpus data. In addition, the verb in the relative clause is more likely to be either the verb *be* or another light verb. In all of the corpora, the most common verb lemma appearing in subject relative clauses is *be*, such as in examples (2) and (3). Within our corpora, between 16.8% (Brown) and 28.7% of the subject relatives clauses have *be* as the verb. In these cases, the information content of the subject relative clause is carried by the noun phrase predicate of the verb *be*. In all corpora, the most frequent verbs to appear in subject relative clauses are low content verbs. In fact, in Switchboard, the four most frequent verbs in subject relative clauses, *be*, *have*, *go*, and *do* account for nearly 50% of the examples. Many of the subject relative clauses in the spoken data give the subjective impression of representing either some form of production difficulty or potentially deliberate avoidance of lexical items (e.g., using *the people that ran the prison* instead of *the wardens* in Switchboard) while the subject relative clauses in the written data appear to more likely to specify additional information about a noun phrase (e.g., *the neuronal changes which underlie the neuroses* in Brown)

(2) Uh, I have [a sister that **is** in nursing school]$_{\text{subject relative clause}}$. [Switchboard]

(3) And [the plans that **are** available to us]$_{\text{subject relative clause}}$, uh, range from kind of mediocre to really sweet. [Switchboard]

(4) Well [one thing that pops into my mind]$_{\text{subject relative clause}}$ real quick is, uh, about the, uh, funding of, the, the school system right now. [Switchboard]

(5) [A buyer who **chooses** to fly to his destination]$_{\text{subject relative clause}}$ must pay for his own ticket but gets a companion's ticket free if they fly on United Airlines. [Wall Street Journal]

(6) One engineer developed [a "crab car" that **moves** sideways]$_{\text{subject relative clause}}$. [Wall Street Journal]

Passive relatives

Passive relatives and the passive infinitive relative are less common in the spoken data, most dramatically in the Switchboard corpus (see Fig. 2). This is in part a result of the first person narrative nature of spoken discourse—particularly that in Switchboard. Given that the speaker and the speakers actions are typically the topic of discussion, there is less need to de-emphasize the 'doer' of an action—a typical function of passive structures (Thompson, 1987). This is consistent with other sets of corpus data. Biber (1993), for example, pointed out that passives are much more common in scientific texts than in fiction or conversation, and Chafe (1982) found that passives occurred about five times more often in the written corpora than in the spoken corpora that he investigated. With regard to the fourfold difference between the two spoken corpora, we suggest that the more formal nature of the spoken matter in the British National Corpus spoken corpus (which includes speech from contexts such as lectures, speeches, and news broadcasts, as well as conversational material such as that found in Switchboard) may contribute to this disparity. Again, this finding highlights the importance of

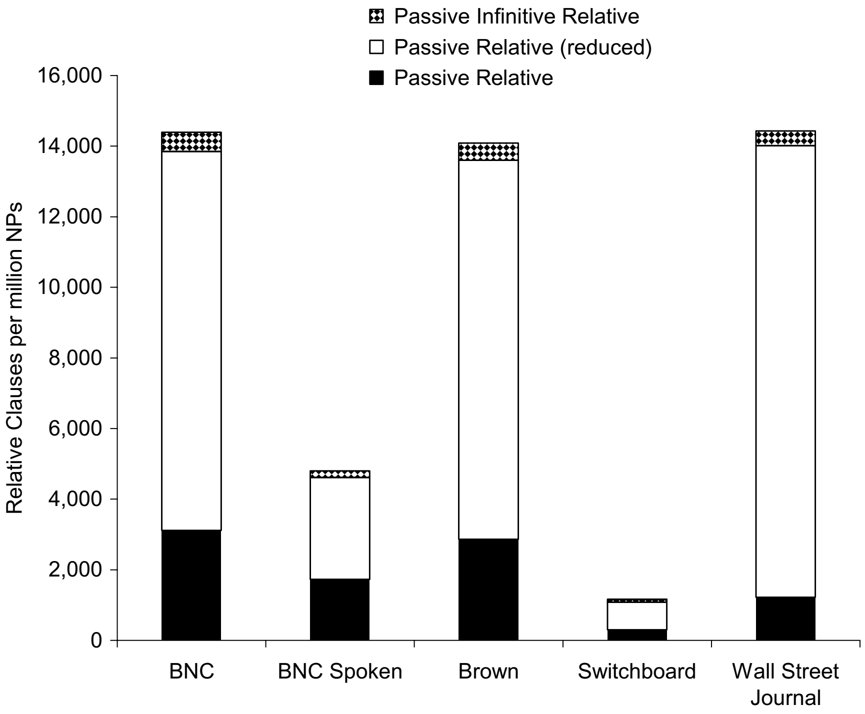

Fig. 2. Distribution of passive relative clauses across corpora (per million noun phrases).

discourse register for studies of relative structural frequency.

Object relatives

Object relative clauses in particular have been the focus of sustained attention by investigators in psycholinguistics (Blumstein et al., 1998; Cohen & Mehler, 1996; Dickey & Thompson, 2004; Gordon et al., 2001), cognitive development (Friedmann & Novogrodsky, 2004), and neuroimaging (Constable et al., 2004; Stromswold, Caplan, Alpert, & Rauch, 1996). Of particular interest is the difficulty in understanding object relatives compared to other clause types such as subject relatives (Gibson, 1998; but cf. Hsiao & Gibson, 2003), their processing time course (Kutas & King, 1996; Penolazzi, De Vincenzi, Angrilli, & Job, 2005; Vos & Friederici, 2003), and the reasons for their reduction through omission of the complementizer *that* (Jaeger & Wasow, 2005; Temperley, 2003; Wasow et al., 2005). Because these clauses have played such a central role in language research, we investigated the frequency and usage of object relatives in particular detail.

Unlike subject and passive relative clauses, object relative clauses are more common in spoken than written corpora (see Fig. 3). This is mainly the result of the relationship between the discourse needs of written and spoken English and the discourse functions served by object relative clauses. However, a portion of the object relative clauses observed in the spoken British National Corpus are actually the result of parser errors—an issue discussed below.

One result of the higher use of object relatives in spoken data, in conjunction with the lower rate of use of subject relatives is that spoken corpora actually have more object relatives than subject relatives (see Fig. 4). This extends the observations of Fox (1987) to a much larger set of data.

In the case of Switchboard, this preponderance of object relatives is in part the result of two different properties of the Switchboard data. One property is that low content filler words such as *things* and *something* are used more frequently in Switchboard than they are in the other corpora. The content intended for these words is often specified via a relative clause, such as in (7) and (8), resulting in a higher occurrence of object relatives and object infinitive relative clauses, as in (9). As a result, more than 27% of the object relatives in Switchboard contained the string *thing* (e.g., *things*, *anything*, *something*) in the main noun phrase of the object relative, while in all other corpora, this was true for less than 10% of the object relatives.

(7) It came on after **something** [we used to watch]$_{\text{Object Relative Clause}}$. (Switchboard)

(8) Several of the **things** [you mentioned]$_{\text{Object Relative Clause}}$ were the things that, uh, our son has talked a lot about Texas A and M. (Switchboard)

(9) We ended up watching it for a couple of hours—zooming out and grabbing **something** [to eat]$_{\text{Object Infinitive Relative Clause}}$ and then zooming back and watching it some more. (Switchboard)

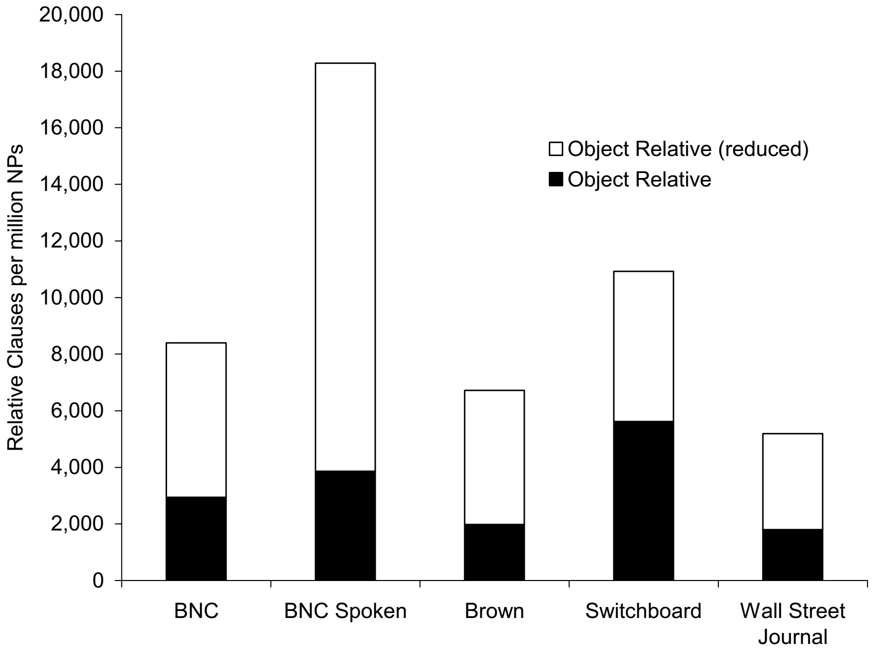

Fig. 3. Distribution of reduced and unreduced object relatives across corpora.

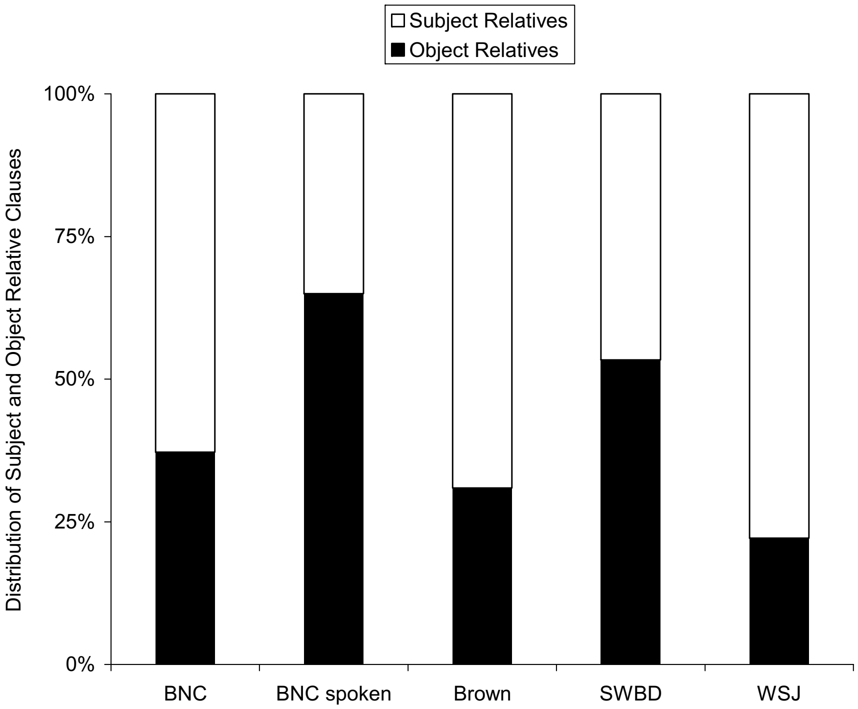

Fig. 4. Percentage of object relative clauses across corpora (out of subject and object relative clauses).

Switchboard has a higher rate of first person subjects (see Roland & Jurafsky, 2002) than the other corpora, due in part to the nature of the task involved in producing the data—conversations between strangers without a common background being asked to discuss a pre-assigned topic. The focus on a first person topic also affects the rate of object relatives, as various noun phrase referents are tied back to the first person topic, as in (10) and (11), again resulting in object relatives. This relationship between object relatives and information flow is discussed extensively in Fox (1987) and Fox and Thompson (1990).

(10) But I found it hard to deal with the, the **dealership** [I was going through]$_{\text{Object Relative Clause}}$. (Switchboard)

(11) And, you know, I want a **car** [that I can work on]$_{\text{Object Relative Clause}}$ because I think it just costs too much even to get the oil changed anymore. (Switchboard)

The presence of these object relatives that relate the modified noun phrase back to the first or second person topic/subject of the discourse in the spoken data appears to be largely responsible for the split of object relatives having pronominal embedded noun phrases and subject relatives having full embedded noun phrases observed by Reali and Christiansen (2007) in data from the American National Corpus. They do not break their data down into spoken versus written, but our data from Brown and Switchboard show that subject relatives with full noun phrases as embedded noun phrases predominate in the written Brown data, while object relatives with pronominal embedded noun phrases predominate in the spoken Switchboard data (see Table 8). In the Switchboard data, cases where the embedded noun phrase is the pronoun *I* account for 47% of all object relatives, and the combined set of first and second person pronouns accounts for 80% of the object relatives. For purposes of comparison with the Reali and Christiansen data, Table 8 includes only transitive subject relatives, and only the cases where the relativizer is *that*. However, both examples with other relativizers and reduced relatives follow the same general pattern.

We also observed differences between the subject and object relative clauses with respect to whether the modified noun phrase was animate or inanimate. Animacy has been argued to play a role in relative clause

Table 8
Noun phrase status and relative type in Brown and Switchboard (raw counts)

Corpus	Embedded noun phrase	Subject relative	Object relative
Brown	Pronominal	97	64
	Full	**521**	93
Switchboard	Pronominal	61	**338**
	Full	162	36

processing (e.g., Traxler et al., 2002, 2005). Gennari and MacDonald (2005) found in a gated sentence completion task that sentence fragments consisting of an animate noun phrase followed by *that* (e.g., *The director that*___) were completed with a subject relative clause 90% of the time, while fragments starting with an inanimate noun phrase (e.g., *The movie that*___) were completed with an object relative clause 65% of the time.

While we did not code our entire data set for animacy, since this can not be done with sufficient accuracy using automated methods, we did code a random sample of 100 relative clauses each from the Brown and Switchboard corpora for animacy. We found that when the modified noun phrase was animate, the relative clause was usually a subject relative clause, while when the noun phrase was inanimate, the relative clause was more likely to be an object relative clause (see Table 9).

In addition, the object relative clauses in both corpora typically had the form of inanimate noun phrase + that + pronoun. In Switchboard, the pronouns were typically *I* or *you*, while in Brown, *he* was the most common pronoun. In both cases, this suggest that the object relative clauses were serving the same discourse function (tying an inanimate subject noun phrase back to the topic of discourse), but that the exact pronoun varies due to the 1st/2nd person nature of spoken discourse and the 3rd person nature of written text.

The surprisingly large number of reduced object relatives in the spoken British National Corpus data are in part the result of a class of errors in the structures assigned by the automatic parser which we used. Specifically, utterances in the spoken British National Corpus data containing the tag phrase *you know* are frequently assigned an incorrect structure. Such utterances, shown in examples (12) and (13) occur in all of the corpora we examined. In the hand corrected Treebank corpora, these examples are given the correct structure, as in Fig. 5. However, in the automatically parsed British National Corpus data, these utterances are frequently assigned the incorrect structure shown in Fig. 6—giving them the same structure as the reduced object relatives in this corpus. We estimate that of the approximately 25,000 reduced object relatives in the spoken British National Corpus data, 4000 are actually examples of the *you know* structure. Because it is possible to form a reduced object relative with *you know*, all of the examples would need to be hand checked in order to completely correct this error. Please see Appendix A for further discussion of the errors in the automatically parsed data.

Table 9
Noun phrase animacy and relative type in Brown and Switchboard

Modified noun phrase	% subject relative	
	Brown (%)	Switchboard (%)
Animate	75	91
Inanimate	47	31

```
((TOP (S (S (NP (PRP It))
            (VP (VBZ 's)
                (NP (DT a)
                    (NN formality))))
         (, ,)
         (S (NP (PRP you))
            (VP (VBP know))))
     (. .))
```

Fig. 5. Example of “you know” with correctly assigned structure (Brown).

```
(S1 (S (NP (PRP I))
       (VP (AUX had)
           (NP (NP (CD three)
                   (NNS sisters))
               (SBAR (S (NP (PRP you))
                        (VP (VBP know))))))
       (. .)))
```

Fig. 6. Example of “you know” with incorrectly assigned structure (British National Corpus spoken).

(12) I had three sisters you know. (British National Corpus spoken)
(13) It’s a formality, you know. (Brown)

Object relative reduction. Object relatives have been used to investigate optionality in language production (e.g., Jaeger & Wasow, 2005; Temperley, 2003; Wasow et al., 2005). This is because the reduced and unreduced forms of object relatives appear to have equivalent meanings and conditions of use. However, there is also building evidence that the distribution of reduced and non-reduced object relatives is in fact governed by discourse factors (Fox & Thompson, 2007). In addition, Jaeger (2005) presents corpus data suggesting that the *that* in object relatives is used to buy time for speakers, and the presence of *that* correlates with disfluencies. We consider the issue of object relative reduction with respect to discourse, disfluencies, and ambiguity avoidance.

The frequency of object relative reduction can be found by comparing the frequency of reduced and unreduced object relatives in each corpus. As Table 10 shows, object relatives in Switchboard have a comparatively low likelihood of being reduced (or alternatively, a high presence of relativizers such as *that*).

While the high rate of *that* omission in the spoken British National Corpus data suggests that the low rate of *that* omission in the Switchboard data may not be due to a spoken vs. written difference, there are two

Table 10
Percent reduction for different types of relative clauses in each corpus

	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
% reduction of object relatives	65	79	71	49	65

potentially confounding factors; the rate of *that* omission in the British National Corpus spoken data is artificially inflated by the *you know* error discussed in the previous section, and Switchboard consists entirely of spontaneous telephone conversations, while the British National Corpus spoken data include material that is more likely to be pre-prepared, such as lectures, speeches, and news broadcasts, along with conversational material. Because *that* use is correlated with disfluencies and production difficulties, it is more likely to occur in spontaneous speech.

Disfluencies and object relative reduction. The relationship between *that* use and disfluencies is predicted by Ferreira and Dell's notion of the complementizer *that* (in sentential complements, rather than object relatives) being eliminated to allow early mention of readily available material (Ferreira & Dell, 2000). This relationship has also been observed in sentence production experiments (Ferreira & Firato, 2002). Further support for the relationship between disfluencies and *that* use comes from an independently performed corpus based analysis of relative clauses by Jaeger and colleagues (Jaeger, 2005; Jaeger & Wasow, 2005; Wasow et al., 2005). Indeed, as would be suggested by these previous results, there is a relationship between the degree of fluency within a Switchboard example and whether the object relative clause is likely to be reduced or not. We find that 52% of the sentences containing an un-reduced object relative contain a repair (as marked by "/") within the relative clause, while only 39% of the sentences containing a reduced object relative contain a repair. This suggests that either relativizers (primarily *that*) are more likely during disfluent speech, or that they are less likely during fluent speech. We find a similar relationship between disfluencies and *that* use in sentential complements (see below).

Although Switchboard carefully transcribes disfluencies and repairs, some disfluencies are also marked in the British National Corpus spoken data, typically by the presence of *erm*. But contrary to the findings with Switchboard, in the British National Corpus spoken data, there is no relationship between the presence of *erm* and whether an object relative is reduced or not. This may be because disfluencies may not be consistently marked in the British National Corpus transcriptions. Only 15% of all object relatives in the British National Corpus contain an *erm*, while 46% of those in the Switchboard data contain a "/". The lower rate of (transcribed) disfluency in the British National Corpus spoken data may also be a product of the differences in the types of spoken language included in Switchboard and the British National Corpus.

Ambiguity avoidance through use of relativizers. Two different views of object relative clauses suggest that *that* use may be related to ambiguity avoidance. Temperley (2003) suggests that the relativizer *that* is used to avoid a potential ambiguity. This ambiguity arises in object relatives such as *the car that companies buy most often* when the relativizer is not present, due to a misparsing the two separate noun phrases as a single noun phrase—in this case, *the car companies* (e.g., where *car* is interpreted as an adjectival modifier of *companies*). Experimental evidence has shown that such misparsings do occur when the two noun phrases form a plausible combination (Grodner, Gibson, & Tunstall, 2002). This ambiguity cannot arise when the embedded noun phrase is pronominal (14), or starts with a determiner (15), and is unlikely or impossible when it is a proper name (17). Thus, the ambiguity can only arise in a subset of the cases where the embedded noun phrase is a bare noun (16).

(14) the bonds [it]$_{\text{pronoun}}$ purchased in 1984 (Wall Street Journal)
(15) the test [the]$_{\text{determiner}}$ student was taking (Wall Street Journal)
(16) The more accounts [customers]$_{\text{noun}}$ have (Wall Street Journal)
(17) the last thing [Mr. Smith]$_{\text{proper noun}}$ needs right now (Wall Street Journal)

Alternatively, Wasow et al.'s 'predictability hypothesis' more generally claims that the more predictable the existence of a relative clause is, the less likely it is to have a *that*. In a sense, this has the end result of *that* being used to avoid ambiguity, but allows for ambiguity to be avoided without necessitating a direct chain of cause and effect between potential ambiguity and *that* use. Rather, the use of *that* could be due to discourse factors and the production needs of the speaker, but still result in an apparent relationship between ambiguous situations and *that* use.

Temperley (2003) uses labeled object relative data (summarized in Table 11) from the first 6 (out of 25) sections of the Wall Street Journal corpus to examine the occurrence of the relativizer *that* with different types of

Table 11
Temperley (2003) complementizer vs. part of speech data

	Pronoun	Determiner	Noun	Proper noun
Full	17	18	32	23
Reduced	139	54	6	40
% full	11	25	84	37

embedded noun phrases. He finds that the bare noun cases are very likely (84%) to have a *that*, while the other cases are more likely to be reduced, suggesting that ambiguity avoidance plays a role in *that* use.

In order to reproduce and expand upon Temperley's data, we modified our object relative search patterns to exclude non-restrictive relative clauses (by eliminating cases marked with commas) and to exclude examples that were contained in quotes (such examples were excluded from his data). We then used the Treebank part of speech tag for the first word following the relativizer to place each example into one of his four categories (pronoun, determiner, noun, proper noun). We identified the same 329 examples as Temperley—as verified by comparison with his (unlabeled) data, but our automatic labeling results in a slightly different distribution of the examples within the categories, as shown in Table 12. Our results also show that relative clauses are less likely to be reduced when the embedded subject noun phrase starts with a noun, and more likely to be reduced otherwise. However, this pattern is slightly weaker in our results than in Temperley's results.

We then expanded our data to include the entire set of Wall Street Journal data as well as the data from the other corpora (British National Corpus, British National Corpus spoken, Brown, Switchboard) that we have been using. These results, shown in Table 13, suggest a slightly different picture than the original data. When the remaining Wall Street Journal data are taken into account, there is still a pattern of an increased likelihood of a relativizer when the following word is a noun, but the difference between the noun and other cases is somewhat less pronounced than in the first 6 sections of the Wall Street Journal data. The British National Corpus and British National Corpus spoken data show a similar pattern. However, the Brown and Switchboard data have a clearly different pattern. As Fig. 7 shows, these two corpora have nearly the same rate of relativizer use whether the second noun phrase has a determiner or whether it is a bare noun phrase (contrary to the predictions of the ambiguity avoidance hypothesis). In these two corpora, the relevant distinction in determining how likely the object relative is to be reduced is whether the embedded subject is pronominal or not. In fact, in all of the corpora we looked at, reduction was most likely when the embedded subjects were pronominal. This suggests that the presence or absence of *that* in object relative clauses is more likely to be due to discourse factors than to ambiguity avoidance per se.

This relationship between pronominal embedded subject noun phrases and relative clause reduction has also been reported in other studies (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Fox & Thompson, 2007; Tottie, 1995). Fox and Thompson attribute this relationship to a tendency to omit the relativizer more often when the combination of the relative clause and the modified noun phrase can be analyzed as a monoclausal structure, as in the comparatively idiomatic example in (18), and a tendency to produce the relativizer more often when the two noun phrases function as two separate clauses, with the relative clause providing additional information about the modified noun phrase, as in (19).

(18) "That's the way [I am]$_{\text{reduced object relative}}$", he says. (Brown)

(19) Carrying it to the living room, she imagined the picture she made: tall and roundly slim, a bit sophisticated in her yellow sheath, with a graceful swingy walk [that she had learned as a twirler with the school band]$_{\text{object relative}}$. (Brown)

Table 12
Our Wall Street Journal (first 6 segments) relativizer vs. part of speech data

	Pronoun	Determiner	Noun	Proper noun
Full	22	25	19	26
Reduced	138	51	9	39
% full	14	33	68	40

Table 13
Percent relativizer present in various corpora

	Pronoun (%)	Determiner (%)	Noun (%)	Proper noun (%)
British National Corpus	25	46	62	59
British National Corpus Spoken	18	27	55	57
Brown	18	39	39	36
Switchboard	47	65	69	100 ($N = 5$)
Wall Street Journal 1st 6	14	33	68	40
Wall Street Journal all	17	37	52	42

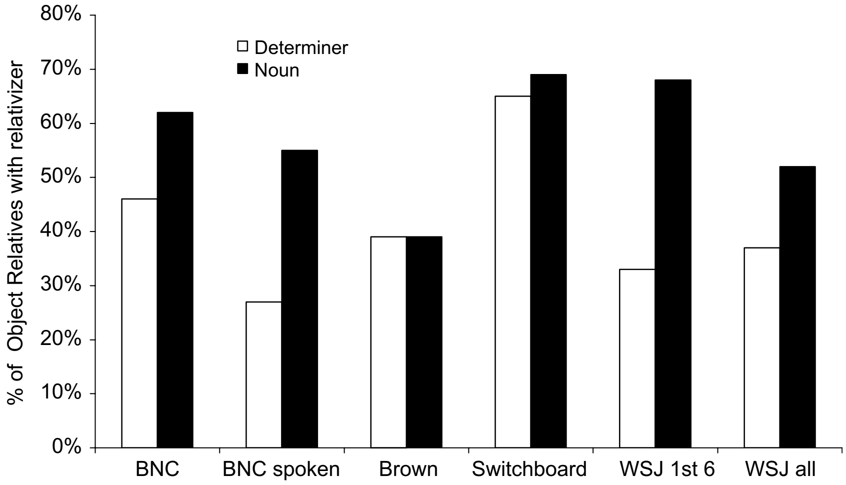

Fig. 7. Percentage of object relative clauses with relativizer when the subject noun phrase of the relative clause either starts with a determiner or is a bare noun.

While our data does not preclude the relativizer *that* being used directly to avoid ambiguity, our corpus data suggests that the use of the relativizer *that* is governed by a complex set of factors, of which ambiguity is only one.

Subcategorization frame probabilities

One of the goals of this study is to provide verb subcategorization data for individual verbs, based on the subcategorization frames shown in Table 14. We also summarize and discuss the patterns found when generalizing across all verbs.

These data are relevant to a number of important psycholinguistic issues. In virtually all linguistic theories, a verb's representation includes a specification of which syntactic frames a verb can occur in. In some processing accounts, the relative frequency of occurrence of the various frames is also claimed to play a role in processing. These theories would predict, for example, that when a comprehender hears the initial sentence fragment, *The man accepted his job…*, in which *job* could either be the direct object of the verb, of the subject of a sentential complement (e.g., *… would be at risk*), the direct object interpretation will be preferred because the verb *accept* occurs more frequently in this structure.

A second, related question is whether *that*-omission, which leads to temporary ambiguity regarding the structural interpretation of a sentence (as with object relative clauses), is truly optional and random, or whether it correlates with some other predictable factors. In the latter case, some potential examples of temporary syntactic ambiguity might in fact not be ambiguous. As in the case of object relatives, we shall see that the patterns of omission suggest that at least in some cases, omission of the complementizer is non-random.

Third, any assessment of word order patterns in English (more specifically, the relative ordering of subject, object, and verb) must be based on a detailed analysis of all various word orders that occur in the subcategorization frames that are possible in the language. In the earlier discussion of relative clauses, we saw that in fact, object-before-subject constructions are more frequent than subject-before-object constructions (i.e., in all corpora, the combination of passive and object relatives are more common than subject relative clauses). In this section, we consider the question of word order in the context of passives. In the final section, we provide a comprehensive analysis of word order over all grammatical structures.

Finally, we will note that there are significant cross-corpus differences in verb subcategorization usage. These have been noted previously (Hare, McRae, & Elman, 2003; Hare et al., 2004; Roland & Jurafsky, 1998; Roland et al., 2000), and are confirmed here. These differences often arise for interpretable and interesting reasons that reveal subtleties of language use that might not be obvious from mere introspection.

Table 15 shows the relative frequencies of each of the Connine verb subcategorizations, collapsing across verb/lemma (see Table 14 for examples of each subcategorization frame, and Appendix B for a more detailed description of each subcategorization frame). Most of the subcategorizations have similar relative frequencies across the corpora. However, there are three sets of subcategorizations where there are notable differences: The simple intransitive, the sentential complements (both with and without the complementizer *that*), and the passive subcategorization. The causes of these differences are illustrative of the issues faced in generating verb subcategorization data from corpus data.

Table 14
Examples of each subcategorization frame taken from the Brown Corpus

Subcategorization	Example from Brown Corpus (verb in **bold**, constituents in [square brackets])
Simple Intransitive	Instantly, he **chilled**
Prepositional Phrase	Guerrillas were **racing** [toward him]$_{\text{Prepositional Phrase}}$
To Infinitive Verb Phrase	Hank thanked them and **promised** [to observe the rules]$_{\textit{To}\text{ Infinitive Verb Phrase}}$
Prepositional Phrase + *To* Infinitive Verb Phrase	...Papa **agreed** [with Mama]$_{\text{Prepositional Phrase}}$ [to make a joint will ...]$_{\textit{To}\text{ Infinitive Verb Phrase}}$
WH Clause	I **know** now [why the students insisted that I go to Hiroshima even when I told them I didn't want to]$_{\textit{WH}\text{ Clause}}$
Sentential Complement with Complementizer	She **promised** [that she would soon take a few days' leave and visit the uncle she had never seen, on the island of Oyajima—which was not very far from Yokosuka]$_{\text{Sentential Complement with }\textit{that}}$
Sentential Complement (no Complementizer)	I **wish** [I was younger and less timid]$_{\text{Sentential Complement}}$
Gerundive Verb Phrase	But I couldn't **help** [thinking that Nadine and Wally were getting just what they deserved]$_{\text{Gerundive Verb Phrase}}$
Perception Complement	Far off, in the dusk, he **heard** [voices singing, muffled but strong]$_{\text{Perception Complement}}$
Simple Transitive	The turtle immediately withdrew into its private council room to **study** [the phenomenon]$_{\text{Noun Phrase}}$
Ditransitive	The mayor of the town **taught** [them]$_{\text{Noun Phrase}}$ [English and French]$_{\text{Noun Phrase}}$
Transitive with Prepositional Phrase	They **bought** [rustled cattle]$_{\text{Noun Phrase}}$ [from the outlaw]$_{\text{Prepositional Phrase}}$, kept him supplied with guns and ammunition, harbored his men in their houses
Transitive with *To* Infinitive Verb Phrase	She had assumed before then that one day he would **ask** [her]$_{\text{Noun Phrase}}$ [to marry him]$_{\textit{To}\text{ Infinitive Verb Phrase}}$
Transitive + *WH* clause	I **asked** [Wisman]$_{\text{Noun Phrase}}$ [what would happen if he broke out the go codes and tried to start transmitting one]$_{\textit{WH}\text{ clause}}$
Transitive + Sentential Complement with Complementizer	But, in departing, Lewis **begged** [Breasted]$_{\text{Noun Phrase}}$ [that there be no liquor in the apartment at the Grosvenor on his return]$_{\text{Sentential Complement with }\textit{that}}$, and he took with him the first thirty galleys of Elmer Gantry
Transitive + Sentential Complement (no complementizer)	But I **promised** [Joyce]$_{\text{Noun Phrase}}$ [I would mention her name]$_{\text{Sentential Complement}}$, if at all, only as a last resort
Passive	A cold supper was **ordered** and a bottle of port

Table 15
Relative frequencies of each subcategorization in each corpus (excludes data for PRT, nominal, and other/miscellaneous categories)

Mapped Connine Subcategorization	British National Corpus (%)	British National Corpus Spoken (%)	Brown (%)	Switchboard (%)	Wall Street Journal Treebank 2 (%)
Simple intransitive	11	14	18	32	11
Prepositional phrase	17	13	15	11	11
To Infinitive verb phrase	6	6	5	7	5
Prepositional phrase + *To* Infinitive Verb Phrase	0	0	0	0	0
WH clause	4	6	1	4	3
Sentential complement with Complementizer	3	3	3	2	4
Sentential complement (no Complementizer)	4	9	1	6	7
Gerundive verb phrase	1	0	1	1	1
Perception complement	3	5	2	2	2
Simple transitive	30	31	32	25	29
Ditransitive	1	1	1	2	2
Transitive + Prepositional Phrase	7	5	8	5	11
Transitive + *To* Infinitive Verb Phrase	1	1	1	1	2
Transitive + *WH* clause	2	2	0	2	2
Transitive + Sentential Complement with complementizer	0	0	0	0	0
Transitive + Sentential Complement (no complementizer)	0	1	0	0	0
Passive	9	3	11	2	9
Total	100	100	100	100	100

Simple intransitives

The reason for the comparatively high frequency of the simple intransitive subcategorization in the Switchboard data can be found by looking at the most common verbs having the simple intransitive subcategorization in Switchboard. The verb form *know* accounts for fully 25% of the simple intransitive examples in Switchboard, such as in (20) and (21) while the second most common verb form, *mean*, as in (22) accounts for another 6% of the data. Other discourse related structures, such as (23) account for lesser percentages of the data.

(20) I don't know. (Switchboard)
(21) You know, hobbies. (Switchboard)
(22) I mean I've never seen two cats so close. (Switchboard)
(23) I see. (Switchboard)

Passives

A comparison of the frequency of the passive structure across the three corpora shows that it is primarily a written structure, not a spoken structure. Table 15 shows that a verb phrase in a written corpus such as Brown or the Wall Street Journal is between four and five times more likely to be passive than one in a spoken corpus such as Switchboard. This is consistent with both the data in the previous section for passive relative clauses, and with previously described differences between spoken and written English (Biber, 1993; Chafe, 1982).

Again, as in the previous sections, there are several possible methods for normalizing the data. In this instance, it seems most appropriate to normalize the number of passive verb phrases by the number of verb phrases in each corpus, as is inherently done in Table 15. However, for the purposes of comparison with other studies, we also present the number of passives normalized by the number of words in Table 16. Our results for the number of passives per 1000 words measure are similar to previous results, such as Biber (1988), who reported an average of 0.8 *by* passives and 9.6 *agentless* passives per 1000 words of text. Depending on the genre and nature of the specific sample, the frequency of passive ranged between 0.0 and 8.0 for *by* passives and 0.0 and 38.0 for agentless passives.

One of the reasons for an interest in the relative frequency of passives has to do with the difficulties faced by agrammatic aphasics in interpreting sentences such as passives and object relative clauses where the object/patient precedes the verb (Caramazza & Zurif, 1976, inter alia). The more general question one might ask of corpus data vis-à-vis passives is 'how often does the

Table 16
Frequencies of passive in each corpus

	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
Passive count	755,890	28,250	10,533	566	8081
Passive per 1000 words	7.9	3.2	10.5	2.4	7.7
Passive per 100 verb phrases	9	3	11	2	9

agent (or patient) appear before the verb, and how often does it appear after the verb?'

Although there is not a strict correspondence between thematic roles such as agent and patient, and grammatical roles, such as subject and object, the data in Tables 15 and 16 hint at an answer. The Brown Corpus data suggest that the patient may be before the verb 11% of the time (passives), and after the verb 45% of the time (transitive active—the other subcategorizations in the lower half of Table 15), but this still leaves 44% of the data (the subject verb cases in the upper half of Table 15) unaccounted for. For many of the subject verb cases, such as when there are sentential complements and other additional post-verbal arguments, it is likely that the subject is an agent, but this 44% also includes the pure intransitive examples, which consist both of unaccusative cases and unergative cases. For unaccusative verbs, the subject undergoes the action, such as in (24), while for unergative verbs, the subject is the agent, such as in (25) (Perlmutter, 1978). Thus, *agent verb (patient)* seems to be the dominant order, but it would require hand coding of the data to verify this.

(24) During this study, septic conditions developed in the oxidation pond in the spring when **the ice melted**. (Brown)

(25) **He jumped**, and sank to his knees in muddy water. (Brown)

That Omission

Sentential complements can occur either with the complementizer *that*, as in (26), or without a complementizer, as in (27). The complementizer is frequently treated as being optional, although various researchers have described factors that correlate with complementizer use (e.g., Ferreira & Dell, 2000; Hawkins, 2002; Jaeger, 2006; Roland et al., 2006; Thompson & Mulac, 1991).

(26) Mr. Martinelli **explained** [that there should be more than enough signatures to assure the scheduling of a vote on the home rule charter and possible election of a nine member charter commission within 70 days]$_{\text{sentential complement with } \textit{that}}$. (Brown)

(27) The driver **admitted** [he was the Dresbachs' son]$_{\text{sentential complement without } \textit{that}}$ and all three were taken to the Edgewater Station, police said. (Brown)

Our data can also be used to investigate the conditions of complementizer use. Table 17 shows the rate of *that* omission for sentential complements in the various corpora. In all of the corpora except Brown, *that* omission is more common than *that* presence. In the Brown corpus, *that* omission is much less common. Table 15 suggests that this is due to a decrease in the number of complementizer-less sentential complement in the Brown data, rather than an increase in the number of sentential complements with the complementizer *that*.

The key to explaining the relative lack of complementizer-less sentential complements in the Brown corpus lies in the typical sources for complementizer-less examples in all of the corpora. Thompson and Mulac (1991) argue that there are really two different types of sentential complement uses, one of which tends to result in an complementizer omission, and the other, in complementizer use. They treat examples such as *I believe that the world is flat*, where the speaker is declaring the existence of a belief (and its contents), as having the traditional main verb (*believe*) subordinate clause (*the world is flat*) relationship, and thus likely to have a complementizer. However, they consider examples such as *I believe it's going to rain* as being different. In this case, they consider *it's going to rain* as the main information in the sentence, and the *I believe* portion as modifying the main proposition to express the speaker's lack of certainty. These epistemic examples are less likely to have a complementizer.

Table 17
Frequency of *that* omission (SC-0 vs. SC-that)

	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
% That omission	56	79	35	76	65

The verb *think* is very common in these epistemic examples. This suggests that high degree of complementizer omission in the Switchboard data is due to a high frequency of epistemic examples within the Switchboard data. In fact, in all of the corpora, the verb *think* is either the most common complementizer-less sentential complement verb or the second most common, and is much more likely to be used without a complementizer than with one. Furthermore, in all corpora, the leading complementizer-less sentential complement verbs tend to be verbs such as *think*, *say*, *guess*, *know*, and *believe*.

The lower rate of complementizer omission in the Brown corpus appears to be the result of two factors, illustrated in Table 18. First, the verbs that tend to have a high probability of complementizer omission are less common in the Brown corpus (as indicated by the smaller percentage of sentential complement examples accounted for by the top four complementizer-less sentential complement verbs in each corpus). Second, at least two of these verbs (*say* and *know*) have a lower rate of complementizer omission than the same verbs do in the other corpora.

Disfluency also plays a role in determining whether *that* is used in sentential complements. As in the case of *that* use with relative clauses (seen in our data and in Jaeger, 2005), the Switchboard sentential complement data (but not the British National Corpus spoken data) shows a relationship between *that* use and disfluencies. In the Switchboard data, 56% of the examples with a complementizer also contained a repair (as indicated by the presence of a "\"), while only 41% complementizer-less examples contained a repair.

Subcategorization probabilities for individual verbs

One of the goals of this paper was to expand upon the previously available verb subcategorization norming data by expanding on both the number of verbs and the set of subcategorizations, while also providing other researchers with the tools necessary to expand upon this data. The verb subcategorization probabilities from all corpora for the verbs included in Connine et al. (1984) and Garnsey et al. (1997) are provided as part of the online Supplementary materials accompanying this article. The counts in these tables represent the actual results from our search patterns. Caution should be used in interpreting these results, since many of the low valued cells represent errors where a verb is alleged to occur in an "impossible" subcategorization, rather than low frequencies of true "possible" subcategorizations. It is difficult to know which is which, without hand checking all of the examples, due to the possibility of novel uses such as *sneeze the foam off the cappuccino* (example from Goldberg, 1995).

In examining these data, we find differences in verb subcategorization probabilities among the different corpora (see Roland & Jurafsky, 2002 for further discus-

Table 18
Top 4 complementizer-less sentential complement verbs for each corpus

Corpus	Verb lemma	% Omission	% of total sentential complement examples accounted for by verb	% of total sentential complement examples for by 1st 4 verbs	Average % omission for first 4 verbs
British National Corpus	Say	69	13	34	74
	Think	86	11		
	Know	66	5		
	Mean	66	4		
British National Corpus Spoken	Think	90	22	55	87
	Say	81	15		
	Mean	94	11		
	Know	83	8		
Brown	Say	59	13	32	65
	Think	86	9		
	Know	50	7		
	Suppose	76	2		
Switchboard	Think	86	48	75	86
	Guess	100	14		
	Say	74	7		
	Know	65	7		
Wall Street Journal Treebank 2	Say	87	56	64	86
	Think	88	4		
	Believe	70	3		
	Mean	71	1		

sion). As an example, Table 19 summarizes the frequencies of each of the subcategorizations for the verb *charge* across the three corpora. The most obvious effect for *charge* is that this verb is most often used in the passive in the two written corpora, while there is only a single example of a passive use of *charge* in the spoken Switchboard corpus. This fits in with the consistent observations of passive being predominately a written structure. Thus, we would expect various subcategorization differences between the corpora to be the result of a spoken vs. written dichotomy.

There are also other factors that contribute to the cross-corpus differences in verb subcategorization frequencies. One of these is the fact that many verbs have multiple senses, and that these senses each have a different set of possible subcategorizations. To the extent that different corpora have biases towards different topics and contexts, and thus towards different senses of the same verb, we would expect there to be resulting differences in overall subcategorization frequencies between corpora. In the case of our data for *charge*, some of the differences between the Brown and Wall Street Journal are caused by (1) an intransitive use of *charge* meaning *run* or *attack* which occurs in the Brown data but is rare in the Wall Street Journal data, as in (28), (2) a noun phrase prepositional phrase use of *charge* meaning *accuse* which is more prevalent in the Wall Street Journal data, as in (29), and (3) a sentential complement use, also meaning *accuse*, which is more prevalent in the Wall Street Journal data, as in (30).

(28) They grew louder as the Indians **charged** again. (Brown)

(29) Sony promptly countersued, **charging** [Warner]$_{\text{noun phrase}}$ [with trying to sabotage its acquisitions and hurt its efforts to enter the U.S. movie business]$_{\text{prepositional phrase}}$. (Wall Street Journal)

(30) Some researchers have **charged** [that the administration is imposing new ideological tests for top scientific posts]$_{\text{Sentential Complement}}$. (Wall Street Journal)

Although we consistently find differences in verb subcategorization probabilities between corpora, these differences are typically not so large as to affect common psycholinguistic uses of verb subcategorization data such as selecting groups of verbs with a bias towards one subcategorization or another. However, within the corpus data presented in this paper, there are some cases where verbs have different transitivity biases in different corpora. One example is the verb *run*, which has a transitive bias in the Wall Street Journal data, and an intransitive bias in the other four corpora (see Table 20). This difference is also caused by the corpora having different

Table 19
Cross-corpus subcategorization counts for *charge*

Subcategorization	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
Simple Intransitive	269	44	8	—	7
Prepositional Phrase	667	53	8	—	3
To Infinitive Verb Phrase	18	2	—	—	1
Prepositional Phrase + *To* Infinitive Verb Phrase	7	1	—	—	—
WH Clause	86	8	—	—	1
Sentential Complement with Complementizer	33	1	2	—	16
Sentential Complement (no Complementizer)	166	21	—	—	5
Gerundive Verb Phrase	16	1	2	—	—
Perception Complement	83	23	—	—	—
Simple Transitive	535	98	5	7	9
Ditransitive	25	6	1	6	2
Transitive + Prepositional Phrase	383	72	2	2	31
Transitive + *To* Infinitive Verb Phrase	46	7	—	—	1
Transitive + *WH* clause	25	4	—	1	1
Transitive + Sentential Complement with complementizer	2	—	1	—	—
Transitive + Sentential Complement (no complementizer)	1	—	—	—	—
Passive	1691	54	13	1	24
Other/error	25	2	—	—	9
Particle	185	25	2	—	2
Auxiliary	114	7	—	—	—
Nominal	2176	151	22	1	39
Total	6554	580	66	18	151

Table 20
Cross-corpus word order counts for *run*

	British National Corpus	British National Corpus Spoken	Brown	Switchboard	Wall Street Journal Treebank 2
% Subject verb object or object verb	42	42	33	38	57

distributions of the possible senses of the verb *run*. In the British National Corpus, Brown, and Switchboard data, predominately intransitive senses of the verb *run*, such as the *rapid physical movement* sense, shown in (31), are more common. In the Wall Street Journal, the obligatorily transitive *manage a company/person* sense, shown in (32), is more common.

(31) Duclos **ran** [toward Desprez]$_{\text{prepositional phrase}}$ with fists raised. (Brown)

(32) May Stores, St. Louis, **runs** [such well-known department stores as Lord & Taylor]$_{\text{noun phrase}}$. (Wall Street Journal)

This word-sense based difference and its effects on psycholinguistic experiments is described in several other papers (Hare et al., 2003, 2004; Roland, 2001; Roland & Jurafsky, 2002) which have relied on previous versions of the search strings used in this paper.

Overall word order frequencies

A major goal of this paper is to provide data on the relative frequency of important grammatical structures at different levels of linguistic granularity. In this last section, we sum over all of the structures reported above (as well as certain types of relative clauses not mentioned above—e.g., those where the extracted element is a prepositional phrase) to arrive at a global view of basic word order frequencies in written and spoken English. We have assigned all grammatical structures to one of 6 basic word-order types: subject verb object object (e.g., ditransitive), subject verb object (e.g., simple transitive), subject verb (e.g., intransitive), object verb object (e.g., passive relatives with ditransitive verbs such as *a part of the software market called systems utilities*), object verb (e.g., passive), and object verb subject (e.g., object relative clauses). The correspondence between each structure, search pattern and word-order type is provided as part of the online Supplementary materials accompanying this article.

English is traditionally considered to be a subject verb object word order language. Indeed, when all structures are considered, our data strongly support this view (see Fig. 8). At most, only 15% of verbs are preceded by the object (e.g., passive sentences, object relatives, passive relatives). In the spoken corpora, this number is much smaller—as few as 5% of the verbs are preceded by their objects in the Switchboard corpus.

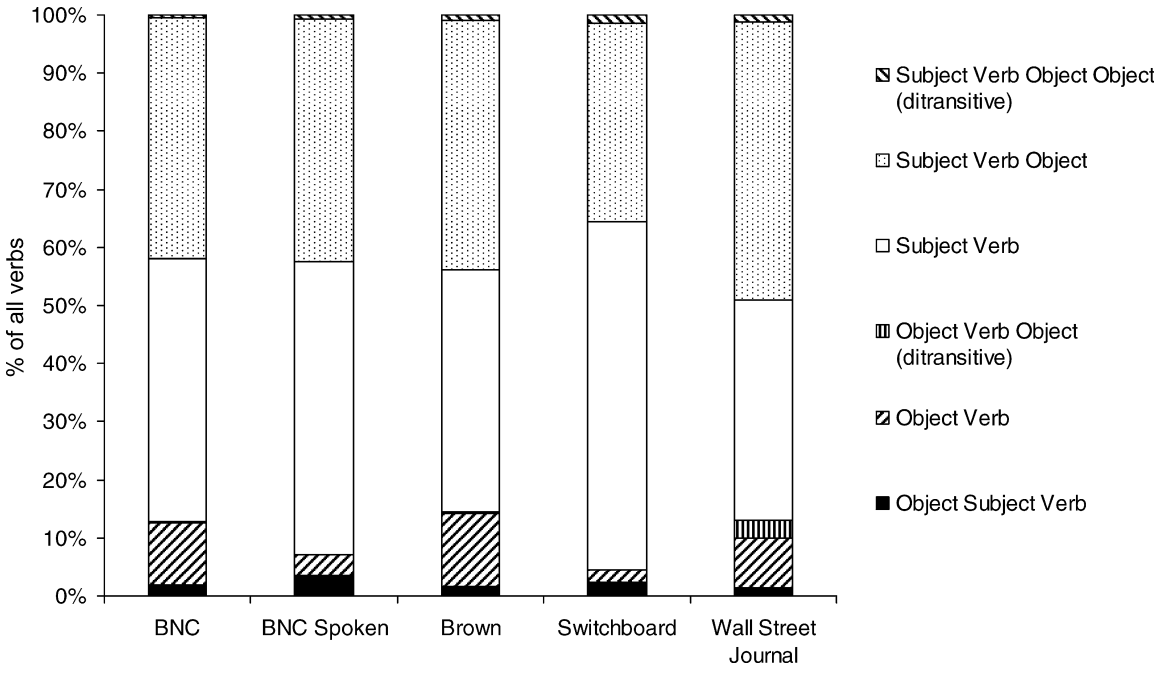

Fig. 8. Distribution of word orders across all structures in each corpus.

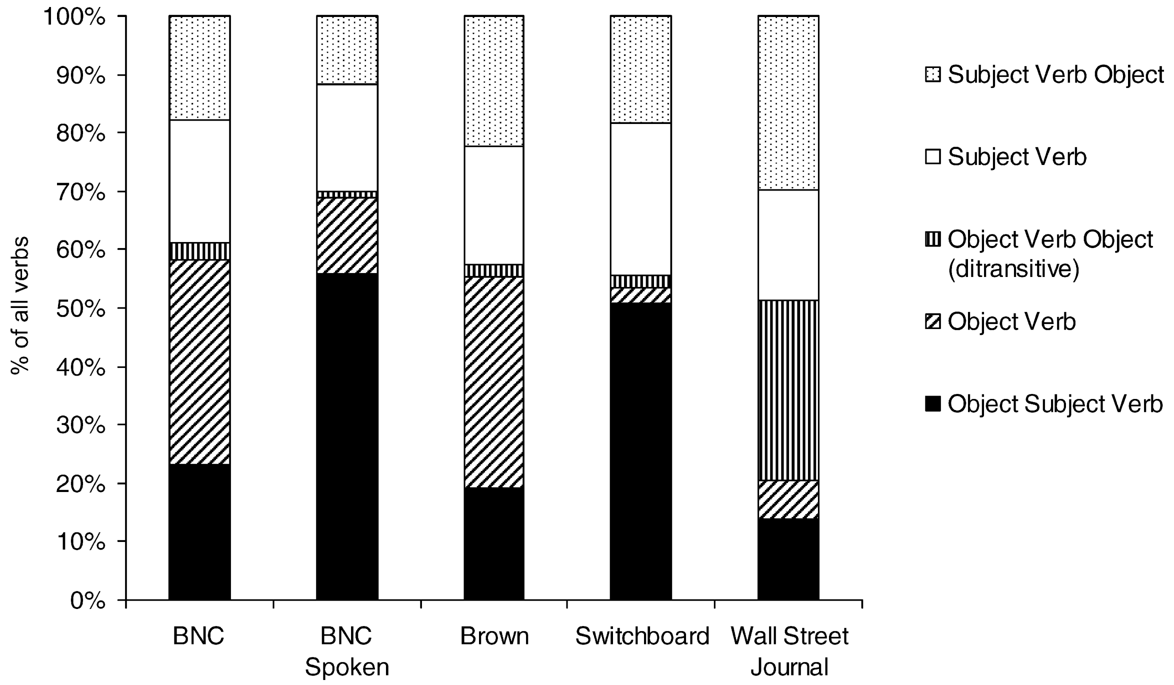

Fig. 9. Distribution of word orders across subject, object, and passive relative clauses in each corpus.

However, when one only considers subject, object, and passive relative clauses, objects are more likely to precede a verb than follow the verb in all corpora—from 50% of the time in the Wall Street Journal written data to up to 80% in the British National Corpus spoken data (see Fig. 9).

Note that in discussing whether the object precedes the verb, we are grouping object verb (primarily various types of passives) and object verb subject (primarily object relatives) together; however, it is not clear that these would necessarily be grouped together by the language comprehension or production system (see Dick et al., 2001; Dick, Butler, Ferreira, & Gernsbacher, submitted for publication).

Conclusions

The primary goal of this paper was to expand upon the set of structural frequency data presently available and to address the issues of how and why the distributions of these structures vary across corpora. In addition, by providing the search patterns and data as part of the online Supplementary materials accompanying this article, we hope to provide a set of tools for other researchers to be able to easily generate their own data from other sets of corpus data. The search patterns described in this paper greatly expand the available structural frequency data available by allowing for the automatic generation of structural frequency information from parsed corpora such as those in the Penn Treebank. Besides the Treebank corpora, the search patterns can be used with any text that has been parsed by statistical parsers such as the Charniak and Collins parsers, although the imperfect nature of current automatic parsing techniques inevitably introduces an additional measure of error. The ability to examine additional texts is particularly useful when looking at the interaction of syntactic structures and particular lexical items (such as when looking at verb subcategorization), especially where the lexical items in question may have a low frequency in the Treebank data.

There remains some error in such data, as is shown by our analysis in Appendix A. We expect that these sorts of errors will be reduced as parsing technology improves. There are also other sources of error that are likely to continue to provide difficulties. These include distinctions which can only be resolved through human coding such as the verb + particle/prepositional phrase distinction (e.g., *look up the address* is a verb particle plus noun phrase construction while *look up the street* is verb plus prepositional phrase construction) and the adjectival passive/true passive distinction (e.g., the island was deserted when we found it is an adjectival passive while the island was deserted by the sailors is a true passive—see Gahl, Jurafsky, & Roland, 2004, for additional discussion). However, there is a cost / benefit tradeoff, since hand coding is expensive in terms of the amount of effort involved.

By providing data from several different corpora, we have also highlighted an important issue in using any such structural frequency data. As the data in this paper and other papers show (e.g., Biber, 1993; Merlo, 1994; Roland & Jurafsky, 2002; Roland et al., 2000), the probabilities of syntactic structures are not universal. Instead, the likelihood of a particular structure is influenced by a wide variety of contextual factors including discourse type, the topics under discussion, the information demands of the situation, the degree of fluency of speech, and the senses of the words being used. We have provided descriptions and examples of some of the differences and their causes, not as an attempt to exhaustively catalogue and account for all possible differences, but instead to illustrate the types of differences and causes that researchers should expect to find.

We suggest that probabilistic theories of language processing must also establish how the frequencies of individual structures combine to form overall representations. The differences in difficulties found in passives, actives, and the various cleft and relative clause constructions may not be explained by the individual frequencies of these constructions, but rather by the overall frequencies of higher level patterns such as *subject verb object* or *agent verb patient*. We feel that language distributional data such as that in this paper will play an important role in converging on an understanding of both the role of frequency in language comprehension and production, and of how higher-level patterns of representation emerge from the representations of specific structures.

Appendix A. Error analysis

Reliability and accuracy is an important issue in the generation of structural frequency information from corpus data. A key problem is that it is not possible to provide a meaningful overall figure such as "all of the numbers are accurate to within X percent". In part, this is because the errors are not random. The automatic techniques that we have used are better at some structures than others. However, it is also the case that the types of errors found in the data affect different research questions in different ways. Therefore, we feel that it is important to provide illustrations of both what types of errors exist in our data, and how these errors might (or might not) affect different research questions.

There are three potential sources of error in our data: errors made by Treebank annotators in assigning syntactic structures within the Treebank corpora, errors made by the Charniak parser in assigning structures to the other corpora, and errors in search pattern design, whereby our patterns assign the wrong category label to correct parse structure.

In general, we found that errors by Treebank annotators (particularly in Treebank II) are minimal, and do not have a major impact on our results. Search string design errors also have only a minimal impact on our results, and those search string design errors that do exist are primarily the result of two design decisions—the decision not to attempt to correctly identify all quote structures, and the decision to limit the extent to which our patterns can correctly capture recursive verb phrase structures (i.e., any structure formed by a rule of the form VP → VP + X). The largest (and most unavoidable) source of error results from errors made by the automatic parser we used for parsing the non Treebank corpora.

We investigated the sources of error in our data using two methods. First, we selected random samples of 100 examples each from all of the corpora, and hand checked the labels which had been assigned by our search patterns for accuracy. In addition, in order to more completely investigate how errors in structure assignment by the Charniak parser affected our results, we aligned all of the examples from the Wall Street Journal-Treebank 2 data and the Wall Street Journal-Charniak data (which contain the same text), and compared the labels which we had given each example in both corpora. Note that because we have labeled all examples of all verbs, the only type of error which we cannot detect at all is one where either the Treebank annotators or the Charniak parser failed to label a verb as being a verb. While errors involving low frequency structures are unlikely to show up in our hand checked samples, we are able to see that we/the parser/the annotators have not missed any major structures.

Table A1 shows the error rates (in terms of percent correct) from three corpora. The error rates for Wall Street Journal-Treebank 2 primarily reflect search pattern design errors, while the increase in error shown for Wall Street Journal-Charniak reflects the additional error resulting from automatic parsing. The error rate for the British National Corpus data reflects additional error resulting from the Charniak parser being used on data that has different properties from the original training data for the parser. Some specific examples of how genre and other cross corpus differences increase parser error are described in this paper, but also see Gildea (2001) for further discussion of this issue.

In the sample of 100 Wall Street Journal-Treebank 2 examples, we found that 92% of the examples had been given the appropriate subcategorization/structure labels. Of the 8 errors, 5 were cases where a sentence containing a quotation had been inappropriately assigned to either the [0] subcategorization or the [other/error] category. This is the result of our decision to sacrifice accuracy on identifying quotes for speech verbs in favor of more accurate identification of subcategorization for all other verbs. Of the other three errors, one was the result of a time noun phrase being counted as the object of a verb, one the result of a failure to correctly deal with a recursive verb phrase structure, and other a miscellaneous failure to correctly classify a low frequency structure. Because most of the errors in our search patterns only involve verbs which take quotes as arguments (primarily *say*), we feel that our data for most other verbs in Treebank corpora is approximately 97% correct.

We can also compare the structural labels that we gave to examples in the Wall Street Journal-Charniak data with those

Table A1
Percent correct labeling of structures in different corpora ($n = 100$ per corpus)

Corpus	% Correct
Wall Street Journal-Treebank 2	92
Wall Street Journal-Charniak	84
British National Corpus	71

that we gave to the same examples in the Wall Street Journal-Treebank 2 data. This allows us to look at the types of errors made by the automatic parser and consider how these errors do (and do not) affect the types of structural frequency counts of interest in psycholinguistics. Out of the 141,675 examples that we were able to map between the two sources (we failed to map approximately 1% of the examples), 87% were given the same label in both Wall Street Journal-Treebank 2 and Wall Street Journal-Charniak. Note that this measure of the degree of error in the Wall Street Journal-Charniak data is slightly different from that provided in Table A1 (84%), since the comparison of the Wall Street Journal-Treebank 2 data and the Wall Street Journal-Charniak data essentially reflects parser errors only, while the hand counting in Table A1 reflects both parser error and search string design error.

The errors did not affect each possible subcategorization label equally. Table A2 shows the error rates in terms of *precision*, *recall*, and *F measure* for assigning the Connine based verb subcategorization labels. Table A3 shows the same data for assigning relative clause type labels. Precision is the percent correct of all examples identified as having a particular structure (i.e., % true positives out of all true positives and false positives). Recall is the percentage of all true examples of a structure that are actually identified as being examples of structure (i.e., % true positives out of all true positives and false negatives). The F measure represents a combination of precision and recall (2PR/(P + R)). All three measures are standard measures of performance in computational linguistics (see Jurafsky & Martin, 2000; Manning & Schutze, 1999).

It is informative to look at the types of differences we found between the labels we assigned to the Wall Street Journal-Treebank 2 data and those we assigned to the Wall Street Journal-Charniak data. Table A4 shows the four most common category label changes between these two data sources. The most frequent difference between the two data sources were 2677 cases where a verb was given a [simple intransitive] subcategorization in the Wall Street Journal-Charniak data while the same example was given a [error/other] label in the Wall Street Journal-Treebank 2 data. All of these cases represent examples of inverted quotes, such as in example (A1). This difference is really a search string design error, because in neither case do our search patterns properly count the verb as having a quote as an argument.

(A1) Surplus power would be sold on the open market, Enron said. (Wall Street Journal)

The three next most frequent category changes all involve prepositional phrase attachment decisions. Hand inspection of the data shows that in nearly all of these cases, the Treebank 2 prepositional phrase attachment decision is correct, and the automatically assigned prepositional phrase attachment is incorrect. This also has an important implication for the interpretation of our results. As in the case of the quote errors which only affect verbs like *say*, the most frequent parser errors also do not affect all structures equally. Thus, if one were interested in prepositional phrase attachment frequencies, one would be

Table A2
Precision and recall for replicating Wall Street Journal-Treebank 2 verb subcategorization results with Wall Street Journal-Charniak data

Connine subcategorization	Treebank 2 total	Charniak total	Same in both	Precision (%)	Recall (%)	*F* (%)
Simple Intransitive	5875	6623	3552	54	60	57
Prepositional Phrase	9351	10,461	8596	82	92	87
To Infinitive Verb Phrase	4376	4579	4153	91	95	93
Prepositional Phrase + *To* Infinitive Verb Phrase	412	293	252	86	61	71
WH Clause	2686	2258	1857	82	69	75
Sentential Complement with Complementizer	3622	3475	3170	91	88	89
Sentential Complement (no Complementizer)	6333	6257	5988	96	95	95
Gerundive Verb Phrase	973	1000	818	82	84	83
Perception Complement	1676	1695	1373	81	82	81
Simple Transitive	25,913	25,863	23,075	89	89	89
Ditransitive	1396	1455	1182	81	85	83
Transitive + Prepositional Phrase	9862	10,144	8644	85	88	86
Transitive + *To* Infinitive Verb Phrase	1969	1975	1670	85	85	85
Transitive + *WH* clause	1751	1704	1401	82	80	81
Transitive + Sentential Complement with complementizer	251	310	221	71	88	79
Transitive + Sentential Complement (no complementizer)	95	154	71	46	75	57
Passive	7804	6525	6362	98	82	89
Other/error	6019	3830	2426	63	40	49
Particle	3160	3478	2959	85	94	89
Auxiliary	17,292	18,282	16,374	90	95	92

Table A3
Precision and recall for replicating Wall Street Journal-Treebank 2 relative clause results with Wall Street Journal-Charniak data

Relative clause type	Treebank 2 relative clause totals (with any Charniak label)	Charniak relative clause totals (with any Treebank 2 label)	Same in both	Precision (%)	Recall (%)	*F* (%)
Subject Relative	5729	5446	5287	97	92	95
Object Relative	557	458	350	76	63	69
Object Relative (reduced)	1047	1110	839	76	80	78
Passive Relative	369	384	338	88	92	90
Passive Relative (reduced)	3903	3825	3754	98	96	97
Subject Infinitive Relative	2169	1981	1698	86	78	82
Object Infinitive Relative	553	606	441	73	80	76
Passive Infinitive Relative	133	109	97	89	73	80

Table A4
Top four changes in category label between Wall Street Journal-Treebank 2 data and Wall Street Journal-Charniak data

Wall Street Journal-Treebank 2 label	Wall Street Journal-Charniak label	Count
Other/error	0	2677
NP	NP PP	851
NP PP	NP	708
0	PP	685

well advised to use the hand labeled Treebank data. For example, Gibson and Schütze (1999) relied on Treebank data for their study of prepositional phrase attachment preferences. Alternatively, if one were interested in distinctions such as the Direct Object/Sentential Complement distinction or verb transitivity, one could more reasonably use automatically generated data without incurring a much greater degree of error. As an example, Roland et al. (2006) report 90% accuracy at labeling the British National Corpus data on a three way Direct Object/Sentential Complement/Other task (using an earlier version of the search patterns from this paper). This figure of 90% is in line with the results shown in Table A2 for the [sentential complement without complementizer], [sentential complement with complementizer] and [simple transitive] labeling accuracy, and is considerably better than the 71% overall accuracy listed in Table A1.

The results in Table A3 show that the automatic parser has a high degree of reliability in identifying reduced passive relative clauses in data parsed by the Charniak parser. When we hand checked random samples of 100 examples which were labeled as passive relative clauses each from the Wall Street Journal Treebank 2 data and the Wall Street Journal Charniak data, we found similar results. Initially, this suggests that the automatically labeled data would be suitable for a corpus study of which verbs appear in reduced passive relatives (e.g., Hare, Tanenhaus, & McRae, 2007). However, as Table A5 shows, there is a large drop off in performance when the parser is used on British National Corpus data. This indicates that the overall accuracy rate of 71% for the British National Corpus data shown in Table A1 can result in both perfectly usable data, such as for investigating direct object/sentential complement ambiguity resolution, and in unusably bad data, such as in the case of reduced passive relatives. Thus, Hare et al. (2007) had to rely on hand coding to identify reduced passive relatives in the British National Corpus data.

One reason for the large difference in performance of the Charniak parser between the Wall Street Journal data and the British National Corpus data is that in both sets of Wall Street Journal data, there are a large number of sentences including the phrase *the period ended*, as in (A2). In the hand labeled Treebank data, these examples are treated as being reduced passive relatives, and given the structure shown in Fig. A1. The *period ended* examples account for a large share of the reduced passive relative examples found in the Wall Street Journal Treebank 2 data and the Wall Street Journal Charniak data, but are relatively rare in the British National Corpus data. Thus, we conclude that the parser was good at identifying the *period ended* examples, but not as good at identifying other (perhaps better) examples of reduced relative clauses.

(A2) During the 5-year period ended 1986, roughly 80% of the names had money tied up in money-losing syndicates, according to Chatset consultants.

Table A5
Percent correct for examples labeled as passive relatives

	Wall Street Journal Treebank 2	Wall Street Journal Charniak	British National Corpus
% correct	96	95	67

```
(TOP (S (S (PP (IN During)
               (NP (NP (DT the)
                       (JJ five-year)
                       (NN period))
                   (VP (VBD ended)
                       (NP (CD 1986)))))
           (, ,)
           …
```

Fig. A1. Structure assigned to *period ended* examples.

Appendix B. Additional description of each target structure

Many target structures required at least three separate search patterns. The reason for this is that the Charniak parser and the two different phases of the Treebank project, Treebank 1 and Treebank 2, rely on slightly different representations for the same structure. One difference is that Treebank 2 relies on a much more detailed tag set than Treebank 1. The more detailed tag set used in the second phase allows one to identify various syntactic structures more accurately. All of the corpora are available with the Treebank 1 tag set, and the Wall Street Journal and Switchboard corpora are also available with the Treebank 2 tag set. (A subset of the Brown Corpus is now available in Treebank 2 format, but was not available in this format during the initial stages of this research.) The results presented in this paper are for the Treebank 1 version of the Brown corpus, and the Treebank 2 versions of the Wall Street Journal and Switchboard corpora. In addition, both versions of Treebank contain nodes and terminals representing empty categories such as traces from noun phrase movement or a 0 terminal to represent the missing *that* from complementizer omission, while the output of the Charniak parser does not contain non-overtly marked terminals.

B.1. Cleft sentences

The cleft sentences of interest include both subject (B1) and object (B2) clefts.

(B1) It was Richard Nixon's first visit to China in 1972 that set in motion the historic rapprochement between Beijing and Washington. (Wall Street Journal)
(B2) It's paper profits I'm losing. (Wall Street Journal)

These sentences are relatively easy to locate in both the Wall Street Journal and Switchboard corpora, since they are tagged according to the Treebank 2 style of notation, which includes an S-CLF tag that signals the presence of a cleft sentence. The overall frequency of these tags was sufficiently low to allow for hand counting of all subject and object cleft examples from the set of items tagged with the S-CLF node. Since the other data is not tagged with the Treebank 2 style of notation, and these structures were found to be very rare in the first two corpora, no effort was made to identify these two structures in the other corpora.

B.2. Verb subcategorization and cross-verb structural frequencies

In addition, by providing the actual search patterns, we allow the reader to generate subcategorization data for verbs (and corpora) that were not included in any of the previous studies.

As noted above, Connine et al. (1984) categorized verb use into 16 possible subcategorization patterns (with additional *other* categories). These subcategorizations are shown in Table 14 in the main body of the paper. Note that we have subdivided the original *That-S* and *NP That-S* categories for sentential complements into separate categories (listed in Table 14) depending on whether the complementizer *that* is present or absent. This is useful because the sentential complements without the complementizer contain the direct object/sentential complement temporary ambiguity, while the examples with the complementizer do not. The individual constituents that make up these categories are described in detail below; we also discuss the status of several types of constructions that do not fit cleanly into this set of subcategorization frames. The search patterns used for identifying instances of each of these types are described in the online Supplementary materials accompanying this article.

B.2.1. Noun phrase

The noun phrase constituent appears in all transitive categories, including the *ditransitive* subcategorization. Examples (B3) and (B4) illustrate transitive and ditransitive uses of verbs, with the target verb written in boldface, and the noun phrase constituents being delimited with square brackets. Cases where the noun phrase has been dislocated are not counted, such as in example (B5). These cases have a trace noun phrase in the Treebank corpora, but not in the automatically parsed corpora, so we designed our search patterns to exclude these cases from our noun phrase counts in the Treebank corpora in order to make our data more comparable across corpora. Passivization is counted in a separate category—see (B23) and (B24) below. Ideally, the noun phrase constituent should exclude cases where the noun phrase is not an argument of the verb, such as in time and measure phrases, shown in (B6) and (B7). In practice, the noun phrase constituent includes all nodes in Treebank labeled NP except those in Treebank which are not lexically filled (marked with a *T* for *T̲race*). Thus, the counts involving the noun phrase constituent differ from the ideal condition. In the Treebank 2 data, it is possible to identify these cases, since they are labeled with an NP-TMP tag rather than a plain NP tag. Approximately 4% of the verb phrase object noun phrases in Wall Street Journal-Treebank 2 are labeled as NP-TMP, suggesting the degree of error introduced by our not taking this distinction into account in the Treebank 2 data, and the distinction not being made in the other corpora. However, hand checking of various random samples in the other corpora suggest that less than 1% of the structural assignments are affected by this issue.

(B3) But, darn it all, why should we **help** [a couple of spoiled snobs who had looked down their noses at us?]$_{\text{noun phrase}}$ (Brown)
(B4) He suggested that a regrouping of forces might **allow** [the average voter]$_{\text{noun phrase}}$ [a better pull at the right lever for him]$_{\text{noun phrase}}$ on election day. (Brown)
(B5) What did he hope to **accomplish** [T]$_0$ here? (Brown)
(B6) Senators unanimously **approved** [Thursday]$_{\text{time noun phrase}}$ [the bill of Sen. George Parkhouse…]$_{\text{noun phrase}}$ (Brown)
(B7) Even though we had **walked** [miles]$_{\text{measure noun phrase}}$ in Kyoto that day … (Brown) (cf. We walked [the dog]$_{\text{noun phrase}}$)

B.2.2. Prepositional phrase

The prepositional phrase constituent appears in both the *prepositional phrase* and *transitive + prepositional phrase* subcategorizations shown in (B8) and (B9) respectively. Not all prepositional phrases are considered to be arguments of the verb. For example, in example (B10), the prepositional phrase *in our spare time* is less closely associated with the verb *paint* than the prepositional phrase *in colors which had never existed* in example (B8). These less closely associated prepositional phrases, or adjuncts, are attached as sisters of the verb phrase rather than as sisters of the verb in the Treebank data, with the Treebank coders being instructed to use high (adjunct) attachment when in doubt.

This category also relies on a distinction between prepositions and particles, indicated within the Treebank data by the PRT and PP tags, respectively. We treated verb particle combinations like (B11) as separate verbs in all of the sections of this paper where we discuss data for individual verbs. Readers who are interested in calculating subcategorization probabilities with verb-particle examples included can do so with minor modification to the search patterns provided in this paper. Gahl et al. (2004) show that the decision to include/exclude verb-particle examples can alter verb transitivity biases.

(B8) ...they **painted** [in colors which had never existed.]$_{\text{prepositional phrase}}$ (Brown)

(B9) One young girl told me how her mother removed a wart from her finger by soaking a copper penny in vinegar for three days and then **painting** [the finger]$_{\text{noun phrase}}$ [with the liquid]$_{\text{prepositional phrase}}$ several times. (Brown)

(B10) We all **painted** [in our spare time]$_{\text{Adjunct}}$.... (Brown)

(B11) All this was unknown to me, and yet I had dared to **ask** her [out]$_{\text{Particle}}$ for the most important night of the year! (Brown)

B.2.3. 'To' marked infinitive verb phrase

The *to*–marked infinitive verb phrase constituent can appear with either with or without a noun phrase before the infinitive.

(B12) However, three of the managers did say that they would **agree** [to attend the proposed meeting]$_{\text{To Infinitive Verb Phrase}}$ if all of the other managers decided to attend. (Brown)

(B13) He **advised** [the poor woman]$_{\text{noun phrase}}$ [not to appear in court]$_{\text{To Infinitive Verb Phrase}}$ as what she was charged with was not in violation of law. (Brown)

B.2.4. PP with 'to' marked infinitive verb phrase

The *prepositional phrase + to infinitive verb phrase* subcategorization consists of a prepositional phrase, typically with the preposition *for*, and a *to* marked infinitive verb phrase. Both our set of subcategorizations, and the original Connine et al. set on which ours are based, contain a total of three possible subcategorizations involving a *to* marked infinitive: *to infinitive verb phrase*, *prepositional phrase + to infinitive verb phrase*, and *transitive + to infinitive verb phrase*. Thus, the *transitive + to infinitive verb phrase* category (potentially) contains examples with and without prepositional phrases.

(B14) A few years before his death Papa had **agreed** [with Mama]$_{\text{prepositional phrase}}$ [to make a joint will with her]$_{\text{to infinitive verb phrase}}$ in which it would be provided that in the event of the death of either of them an accounting would be made to their children whereby each child would receive a bequest of $5000 cash. (Brown)

B.2.5. 'WH' clause

The *WH* clause can occur with either transitive or intransitive verb uses. This category includes the traditional *WH* words (*who*, *what*, *when*, *where*, *why*, *which*, etc.) as well as *if* and *whether* clauses.

(B15) Note where the sun rises and sets, and **ask** [which direction the prevailing winds and storms come from.]$_{\text{WH clause}}$ (Brown)

(B16) About five years ago, Handley came to **ask** [me]$_{\text{noun phrase}}$ [if he could see the tattered register.]$_{\text{WH clause}}$ (Brown)

B.2.6. Sentential complement

The sentential complement constituent consists of a finite clause. Depending on whether the clause is preceded by the complementizer *that* or not, it is either marked as *Sentential Complement with Complementizer* or *Sentential Complement (no Complementizer)*.

(B17) But I insist upon **believing** [that even when it is lost, it may, like paradise, be regained.]$_{\text{Sentential Complement with Complementizer}}$ (Brown)

(B18) A tribe in ancient India **believed** [the earth was a huge tea tray resting on the backs of three giant elephants, which in turn stood on the shell of a great tortoise.]$_{\text{Sentential Complement (no Complementizer)}}$ (Brown)

B.2.7. Gerundive verb phrase

This constituent consists of a gerundive verb phrase. The gerundive verb has no (lexically filled) subject. Items with a subject are included in the *perception complement* category (described in next section).

(B19) He seemed to **remember** [reading somewhere that Abyssinians had large litters]$_{\text{Gerundive Verb Phrase}}$, and suffered a dismaying vision of the apartment overrun with a dozen kittens. (Brown)

(B20) However, he **continued** [experimenting and lecturing]$_{\text{Gerundive Verb Phrase}}$, publishing the results of his experiments in German and Danish periodicals. (Brown)

B.2.8. Perception complement

The perception complement constituent consists of an untensed clause with either a bare stem verb or an *-ing* verb. These are referred to *perception complements* because they are frequently used with perception verbs such as *hear* and *see*. The distinction between the gerundive verb phrase category and this category is that the verb in the perception complement has a (lexically filled) subject – indicated in the examples below in *italics*.

(B21) Winston had **heard** [*her* shaking out the skirt of her new pink silk hostess gown.]$_{\text{Perception Complement.}}$ (Brown)
(B22) Anyhow, I wasn't surprised, early that morning, to **see** [*Handley* himself crossing from Dogtown Common Road to the Back Road.]$_{\text{Perception Complement}}$. (Brown)

B.2.9. *Passive*

Passive verb uses were automatically identified by tgrep search strings. Passives included both *be* and *get* passives. Reduced relative uses of verbs were not included in this part of the paper, but are covered in the section on relative clauses.

(B23) American history should clinch the case when Congress is **asked** to approve. (Brown)
(B24) This selection-rejection process takes place as the file is **read**. (Brown)

B.2.10. *Absence of other target constituent (0)*

The 0 constituent is the absence of any other target constituent, meaning that these examples do not include any of the other target constituents such as a prepositional phrase, infinitival complement or verb phrase, or sentential complement. Thus, examples labeled with the subcategorization *0* can either have no arguments, such as in (B25) and (B26), or can have additional material, such as adverbial phrases or adjectival phrases, as in (B27), (B28), and (B29), which are not otherwise listed in the set of target subcategorizations.

(B25) He oughta **know**. (Brown)
(B26) I ought to **remember**. (Brown)
(B27) The big man **asked** [again]$_{\text{ADVP}}$, taking a step into the boxcar. (Brown)
(B28) I **knew** [better.]$_{\text{ADVP}}$ (Brown)
(B29) He **felt** [light-headed and sick.]$_{\text{ADJP}}$ (Brown)

B.2.11. *Strategies for assigning verb subcategorization categories to difficult cases*

B.2.11.1. Verb + Particle constructions. Verb plus particle constructions, such as (B30), pose a potential problem for the above set of subcategorizations, since, in this particular case, one could either classify this example as a transitive use of the verb *hold up*, or as a verb particle use of the verb *hold*, or potentially as a prepositional phrase use of the verb *hold*. The Treebank coding style provides a separate tag for particles, and our search patterns rely on this tag to separate the verb particle uses from the other uses of each verb. In the subcategorization counts for individual verbs, cases such as example (B30) would appear in the *other* column for the verb *hold*, since we feel that *hold* and *hold up* are not really the same lexical item. Because our present search patterns do not further differentiate between transitive and intransitive uses of verb particle constructions, we exclude all verb particle constructions from our results, as noted above. This seems appropriate when looking at the data on a verb-by verb basis. One could easily modify the present set of search strings to subdivide the verb particle constructions into various sub-groups if such distinctions were relevant. This represents about 2% of the data.

(B30) The spire seemed to **hold up** the sky. (Brown)

B.2.11.2. Auxiliary verbs. We did not consider the uses of auxiliary verbs such as *be* and *have* in (B31) as separate uses of separate verbs, but instead treated such sentences as consisting of a single use of a single verb. Within the Treebank data, depending on whether the corpus was tagged with the Treebank 1 style of notation (Brown) or the Treebank 2 style of notation (Wall Street Journal, Switchboard), auxiliary verbs are tagged in two possible ways, shown in Figs. B1 and B2. Our search patterns do not look for items tagged with AUX, and thus the upper auxiliary verb in Treebank 1 style notation is inherently excluded from all of our counts. The subsequent auxiliary verbs in Treebank 1 style notation, and all auxiliary verbs in Treebank 2 style notation, can be identified by the fact that these all appear in the search patterns as instances of verbs having VP sisters. Since these appear exclusively in a small subset of our search patterns, they also can be removed from consideration (and are removed from all of the data presented in this paper).

(B31) Caldwell's resignation **had been** expected for some time. (Brown)

B.2.11.3. Quotes. The existence of quoted material within a sentence poses special problems for verb subcategorization data. On one hand, there is the philosophical problem of determining the most appropriate label for the quoted material; on the other, there is the severe operational problem of accurately identifying quoted material within corpus data. Part of the difficulty in quotes lies in the arbitrariness of the material allowed

```
(TOP (S (NP (DT The)
            (NNP September-October)
            (NN term)
            (NN jury))
        (AUX (VBD had))
        (VP (VBN been)
            (VP (VBN charged)
```

Fig. B1. Tagging of auxiliary verbs in Treebank 1 notation (Brown).

```
(TOP (S-2 (NP-SBJ-80 (NP (NNS Modifications))
                     (PP (-NONE- *ICH*-3)))
          (VP (VBD had)
              (VP (VBN been)
                  (VP (VBN made)
```

Fig. B2. Tagging of auxiliary verbs in Treebank 2 notation (Wall Street Journal).

in the quote. The contents of the quote can range from a complete sentence, as in (B32), to single word, as in (B33), or even a nonsense word, as in (B34). Likewise, within the Treebank parses, the quoted material appears in a variety of sentential nodes, ranging from NP to PP to S.

(B32) Moreland fixed us each another drink, and said, ''For God's sake, tell me something truly amusing''. (Brown)
(B33) He said, ''Grosse?'' (Brown)
(B34) ''Pap-pap-pap-hey'', I said. (Brown)

An additional problem is that the quoted material can appear nearly arbitrarily within the sentence, both before the verb and after the verb. Thus, unlike with most constituents such as a noun phrase or a sentential complement, there is no fixed attachment point relative to the target verb within the sentence structure. Because of this, any search patterns with enough flexibility to find all of the quotes will also propose a much larger number of false positives, even for verbs which cannot take quotes as arguments.

One strategy, which has been used in some previous papers (such as Roland, 2001; Roland & Jurafsky, 2002; Roland et al., 2000), is to treat quotes as a separate subcategorization, and hand label a limited number of verbs. An alternative strategy, which we will apply here, is simply to treat material within quotes as 'normal' text, thus making it visible to our search strings. Because quotes represent a very small portion of the corpora we are examining, this strategy will not result in a larger error in our overall counts, but will cause some quote-taking verbs such as *say* to have somewhat misleading subcategorization results. In particular, when the quoted material precedes the verb, the example is likely to be classified in the [0] subcategorization. The search strings for the [0] subcategorization included in the online Supplementary materials accompanying this article are somewhat optimized for distinguishing between quote and non-quote examples, in that quotes tend to appear in some patterns, and not others.

B.2.11.4. Disfluencies. The transcriptions in Switchboard include disfluencies. These disfluencies pose a potential problem for determining what to count as an example of a particular structure. In general, these disfluencies are tagged with an *EDITED* node, so that in the sentence shown in Fig. B3, the first instance of the verb *guess* would not be counted at all, while the second instance of *guess* would be counted in the sentential complement category. In other words, the sentence is treated as the equivalent fluent utterance.

```
(TOP (S (EDITED (RM (-DFL- \[))
                (S (NP-SBJ (PRP I))
                   (VP-UNF (VBP guess)))
                (IP (-DFL- \+)))
        (NP-SBJ (PRP I))
        (VP (VBP guess)
            (RS (-DFL- \]))
            (SBAR (-NONE- 0)
                  (S (NP-SBJ (PRP we))
                     (VP (MD can)
                         (VP (VB start))))))
        (. .)))
```

Fig. B3. Labeling of disfluencies in Switchboard.

Appendix C. Supplementary data

Supplementary files including cross corpus verb subcategorization for 204 verbs that have played a role in the psycholinguistic literature, our corpus search patterns, and other details for reproducing our results accompany the online version of this paper. Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jml.2007.03.002.

References

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., et al. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*, 344–380.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language aquisition* (pp. 157–193). Hillsdale, NJ, England: Lawrence Erlbaum, Inc.

Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. (1996). Comprehension of reversible sentences in ''agrammatism'': A meta-analysis. *Cognition, 58*, 289–308.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics, 19*, 219–241.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Limited.

Blumstein, S. E., Byma, G., Kurowski, K., Hourihan, J., Brown, T., & Hutchinson, A. (1998). On-line processing of filler-gap constructions in aphasia. *Brain and Language, 61*, 149–168.

Burnard, L. (1995). *Users reference guide for the British National Corpus*. Oxford: Oxford University Computing Services.

Bybee, J. (1995). Regular morphology and the lexicon. *Language & Cognitive Processes, 10*, 425–455.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral & Brain Sciences, 22*, 77–126.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language, 3*, 572–582.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language* (pp. 35–53). Norwood, New Jersey: Ablex.

Charniak, E. (1995). *Parsing with context free grammars and word statistics* (No. CS-95-28). Providence, Rhode Island: Brown University.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the fourteenth national conference on artificial intelligence* (pp. 598–603). Menlo Park: AAAI Press/MIT Press.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*, 113–124.

Christiansen, M. H., & Chater, N. (1999). Connectionist natural language processing: The state of the art. *Cognitive Science, 23*, 417–437.

Cohen, L., & Mehler, J. (1996). Click monitoring revisited: An on-line study of sentence comprehension. *Memory & Cognition, 24*, 94–102.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics* (pp. 184–191). Santa Cruz, CA.

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

Collins, M. J. (1999). *Head-driven statistical models for natural language processing*. Unpublished Doctoral dissertation, University of Pennsylvania.

Connine, C., Ferreira, F., Jones, C., Clifton, C., & Frazier, L. (1984). Verb frame preference: Descriptive norms. *Journal of Psycholinguistic Research, 13*, 307–319.

Constable, R. T., Pugh, K. R., Berroya, E., Mencl, W. E., Westerveld, M., Ni, W., et al. (2004). Sentence complexity and input modality effects in sentence comprehension: an fMRI study. *Neuroimage, 22*, 11–21.

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition, 30*, 73–105.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498–513.

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (2006). Relative clause attachment in Dutch: Online comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language & Cognitive Processes, 21*, 453–485.

Desmet, T., & Gibson, E. (2003). Disambiguation preferences and corpus frequencies in noun phrase conjunction. *Journal of Memory and Language, 49*, 353–374.

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. *Psychological Review, 108*, 759–788.

Dick, F., Butler, A. C., Ferreira, V. S., Gernsbacher, M. A., & St. John, M. (submitted for publication). *Frequency-evoked syntactic plasticity in artificially induced agrammatism*.

Dick, F., Wulfeck, B., Krupa-Kwiatkowski, M., & Bates, E. (2004). The development of complex sentence interpretation in typically developing children compared with children with specific language impairments or early unilateral focal lesions. *Developmental Science, 7*(3), 360–377.

Dickey, M. W., & Thompson, C. K. (2004). The resolution and recovery of filler-gap dependencies in aphasia: Evidence from on-line anomaly detection. *Brain and Language, 88*, 108–127.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*, 164–203.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*, 296–340.

Ferreira, V. S., & Firato, C. E. (2002). Proactive interference effects on sentence production. *Psychonomic Bulletin & Review, 9*, 795–800.

Fox, B. A. (1987). The noun phrase accessibility hierarchy reinterpreted: Subject primacy or the absolutive hypothesis? *Language, 64*(4), 856–870.

Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language, 66*(2), 297–316.

Fox, B. A., & Thompson, S. A. (2007). Relative clauses in English conversation: Relativizers, frequency and the notion of construction. *Studies in Language, 31*(2), 293–326.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

Friedmann, N., & Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child Language, 31*, 661–681.

Gahl, S., Jurafsky, D., & Roland, D. (2004). Verb subcategorization frequencies: American English corpus data, methodological studies, and cross-corpus comparisons. *Behavior Research Methods, Instruments & Computers, 36*, 432–443.

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.

Gennari, S., & MacDonald, M. (2005). *Parsing indeterminacy in object relative clauses*. Poster presented at the CUNY 2005 Sentence Processing Conference, March 31st–April 2nd, Tucson, AZ.

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

Gibson, E., & Schütze, C. T. (1999). Disambiguation preferences in noun phrase conjunction do not mirror corpus frequency. *Journal of Memory and Language, 40*, 263–279.

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

Gilboy, E., Sopena, J.-M., Clifton, C., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition, 54*, 131–167.

Gildea, D. (2001). Corpus variation and parser performance. In L. Lee & D. Harman (Eds.), In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing* (pp. 167–172). Carnegie Mellon University.

Godfrey, J., Holliman, E., & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92* (pp. 517–520). San Francisco.

Goldberg, A. (1995). *Constructions*. Chicago: University of Chicago Press.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1411–1423.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language, 51*, 97–114.

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science, 13*, 425–430.

Grishman, R., Macleod, C., & Meyers, A. (1994). Complex syntax: Building a computational lexicon. In *Proceedings of the 15th international conference on computational linguistics (COLING 94)* (pp. 268–272). Kyoto, Japan.

Grodner, D., Gibson, E., & Tunstall, S. (2002). Syntactic complexity in ambiguity resolution. *Journal of Memory and Language, 46*, 267–295.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral & Brain Sciences, 23*, 1–71.

Hare, M., McRae, K., & Elman, J. (2003). Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language, 48*, 281–303.

Hare, M., McRae, K., & Elman, J. L. (2004). Admitting that admitting verb sense into corpus analyses makes sense. *Language & Cognitive Processes, 19*, 181–224.

Hare, M., Tanenhaus, M. K., & McRae, K. (2007). Understanding and producing the reduced relative construction: Evidence from ratings, editing and corpora. *Journal of Memory and Language, 56*, 410–435.

Hawkins, J. A. (2002). Symmetries and asymmetries: Their grammar, typology and parsing. *Theoretical Linguistics, 28*(2), 95–150.

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

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition, 90*, 3–27.

Jaeger, T. F. (2005). Optional that indicates production difficulty: Evidence from disfluencies. In *Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop* (pp. 103–109). Aix-en-Provence, France.

Jaeger, T. F. (2006). Phonological optimization and syntactic variation: The case of optional that. In *Proceedings of the 32nd annual meeting of the Berkeley Linguistics Society*.

Jaeger, T. F., & Wasow, T. (2005). *Production complexity driven variation: The case of relativizer distribution in non-subject-extracted relative clauses*. Paper presented at the 18th Annual CUNY Sentence Processing Conference, March 31st–April 2nd, Tuscon, AZ.

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

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J: Prentice Hall.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.

Kempe, V., & MacWhinney, B. (1999). Processing of morphological and semantic cues in Russian and German. *Language & Cognitive Processes, 14*.

Kennison, S. M. (1999). American English usage frequencies for noun phrase and tensed sentence complement-taking verbs. *Journal of Psycholinguistic Research, 28*, 165–177.

Kutas, M., & King, J. W. (1996). The potentials for basic sentence processing: Differentiating integrative processes. In I. Ikeda & J. L. McClelland (Eds.). *Attention and performance* (Vol. XVI, pp. 501–546). Cambridge, MA: The MIT Press.

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

MacDonald, M. C. (1997). Lexical representations and sentence processing: An introduction. *Language & Cognitive Processes, 12*, 121–136.

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney et al. (Eds.), *The emergence of language* (pp. 177–196). Mahwah, NJ, USA: Lawrence Erlbaum, xvii.

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

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marcus, M. P., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., et al. (1994). The Penn Treebank: Annotating predicate argument structure. In *Human language technology: Proceedings of a workshop held at Plainsboro, New Jersey, March 8–11, 1994* (pp. 114–119). Plainsboro, N.J.

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.

McRae, K., Jared, D., & Seidenberg, M. S. (1990). On the roles of frequency and lexical access in word naming. *Journal of Memory and Language, 29*, 43–65.

Merlo, P. (1994). A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research, 23*, 435–447.

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 (nonlexical) statistical records. *Journal of Psycholinguistic Research Special Issue: Sentence processing: I, 24*, 469–488.

Narayanan, S., & Jurafsky, D. (1998). Bayesian models of human sentence processing. In *Proceedings of the twentieth annual meeting of the cognitive science society COGSCI-98* (pp. 752–757).

Narayanan, S., & Jurafsky, D. (2002). A Bayesian Model Predicts Human Parse Preference and Reading Time in Sentence Processing. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.). *Advances in neural information processing systems* (Vol. 14, pp. 59–65). Cambridge, MA: MIT Press.

Penolazzi, B., De Vincenzi, M., Angrilli, A., & Job, R. (2005). Processing of temporary syntactic ambiguity in Italian "who"-questions" a study with event related potentials. *Neuroscience Letters, 377*, 91–96.

Perlmutter, D. (1978). Impersonal passive and the Unaccusative Hypothesis. In *Proceedings of the fourth annual meeting of the Berkeley Linguistics Society* (pp. 157–189).

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21–69.

Plunkett, K., & Marchman, V. A. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition, 61*, 299–308.

Rayner, K., Warren, T., Juhasz, B. J., & Liversedge, S. P. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 1290–1301.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57*, 1–23.

Roland, D. (2001). Verb sense and verb subcategorization probabilities. *Dissertation Abstracts International, 62*(11-A), 3762.

Roland, D., Elman, J. L., & Ferreira, V. S. (2006). Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. *Cognition, 98*, 245–272.

Roland, D., & Jurafsky, D. (1998). How verb subcategorization frequencies are affected by corpus choice. In *Proceedings of COLING-ACL 1998* (pp. 1117–1121). Montreal, Canada.

Roland, D., & Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. In P. Merlo & S. Stevenson (Eds.), *The lexical basis of sentence processing: Formal, computational, and experimental issues*. John Benjamins.

Roland, D., Jurafsky, D., Menn, L., Gahl, S., Elder, E., & Riddoch, C. (2000). Verb subcategorization frequency differences between business-news and balanced corpora: The role of verb sense. In *Proceedings of the Workshop on Comparing Corpora* (pp. 28–34). Hong Kong, October 2000.

Sag, I. A., & Wasow, T. (1999). *Syntactic theory: A formal introduction*. Stanford, CA: C S L I Publications.

Shapiro, L. P., Gordon, B., Hack, N., & Killackey, J. (1993). Verb-argument structure processing in complex sentences in Broca's and Wernicke's aphasia. *Brain and Language, 45*, 423–447.

Sharkey, N. (1992). *Connectionist natural language processing*. Oxford: Intellect Books.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 10*, 174–215.

St. John, M. F., & Gernsbacher, M. A. (1998). Learning and losing syntax: Practice makes perfect and frequency builds fortitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 231–255). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Incorporated.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language, 52*, 452–473.

Temperley, D. (2003). Ambiguity avoidance in English relative clauses. *Language: Journal of the Linguistic Society of America, 79*(3), 464–484.

Thomas, M. S. C., & Redington, M. (2004). Modelling atypical syntax processing. In W. Sakas (Ed.), *Proceedings of the first workshop on psycho-computational models of human language acquisition at the 20th international conference on computational linguistics* (pp. 85–92). Geneva, Switzerland.

Thompson, S. A. (1987). The passive in English: A discourse perspective. In R. Channon & L. Shockey (Eds.). *Honor of Ilse Lehiste/Ilse Lehiste Puhendusteos* (pp. 497–511). Dordrecht: Foris.

Thompson, S. A., & Mulac, A. (1991). The discourse conditions for the use of the complementizer 'that' in conversational English. *Journal of Pragmatics, 15*, 237–251.

Tottie, G. (1995). The man 0 I love: An analysis of factors favouring zero relatives in written British and American English. In G. Melchers & B. Warren (Eds.), *Studies in anglistics* (pp. 201–215). Stockholm: Almqvist and Wiksell.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*, 69–90.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language, 53*, 204–224.

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.

Vos, S. H., & Friederici, A. D. (2003). Intersentential syntactic context effects on comprehension: The role of working memory. *Cognitive Brain Research, 16*, 111–122.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition, 85*, 79–112.

Warren, T., & Gibson, E. (2005). Effects of NP type in reading cleft sentences in English. *Language & Cognitive Processes, 20*, 751–767.

Wasow, T., Jaeger, T. F., & Orr, D. (2005). Lexical variation in relativizer frequency. Paper presented at the workshop on expecting the unexpected: Exceptions in grammar at the 27th annual meeting of the German Linguistic Association, University of Cologne, Germany.