C E N T E R F O R R E S E A R C H I N L A N G U A G E

February 2001 **Vol. 13, No. 1**

The Newsletter of the Center for Research in Language, University of California, San Diego, La Jolla CA 92093-0526
Tel: (858) 534-2536 • E-mail: info@crl.ucsd.edu • WWW: http://crl.ucsd.edu/newsletter

• • •

FEATURE ARTICLE

The Frequency of Major Sentence Types over Discourse Levels: A Corpus Analysis

Frederic Dick
Jeffrey L. Elman

University of California, San Diego

EDITOR'S NOTE

This newsletter is produced and distributed by the **CENTER FOR RESEARCH IN LANGUAGE**, a research center at the University of California, San Diego that unites the efforts of fields such as Cognitive Science, Linguistics, Psychology, Computer Science, Sociology, and Philosophy, all who share an interest in language. We feature papers related to language and cognition distributed via the World Wide Web) and welcome response from friends and colleagues at UCSD as well as other institutions. Please visit our web site at `http://crl.ucsd.edu`.

SUBSCRIPTION INFORMATION

If you know of others who would be interested in receiving the newsletter, you may add them to our email subscription list by sending an email to `majordomo@crl.ucsd.edu` with the line "`subscribe newsletter <email-address>`" in the body of the message (e.g., `subscribe newsletter jdoe@ucsd.edu`).

Please forward correspondence to:

Ayse Pinar Saygin, Editor
Center for Research in Language, 0526
9500 Gilman Drive, University of California, San Diego 92093-0526
Telephone: (858) 534-2536 • E-mail: editor@crl.ucsd.edu

Back issues of this newsletter are available from CRL in hard copy as well as soft copy form. Papers featured in previous issues include the following:

Connectionist Modeling of the Fast Mapping Phenomenon
Jeanne Milostan
Computer Science and Engineering, UCSD
Vol. 9, No. 3, July 1995

Representing the Structure of a Simple Context-Free Language in a Recurrent Neural Network: A Dynamical Systems Approach
Paul Rodriguez
Department of Cognitive Science, UCSD
Vol. 10, No. 1, October 1995

A Brain Potential Whose Latency Indexes the Length and Frequency of Words
Jonathan W. King
Cognitive Science, UCSD
Marta Kutas
Cognitive Science and Neurosciences, UCSD
Vol. 10, No. 2, November 1995

Bilingual Memory: A Re-Revised Version of the Hierarchical Model of Bilingual Memory
Roberto R. Heredia
Center for Research in Language, La Jolla, CA
Vol. 10, No. 3, January 1996

Development in a Connectionist Framework: Rethinking the Nature-Nurture Debate
Kim Plunkett
Oxford University
Vol. 10, No. 4, February 1996

Rapid Word Learning by 15-Month-Olds under Tightly Controlled Conditions
Graham Schafer and Kim Plunkett
Experimental Psychology, Oxford University
Vol. 10, No. 5, March 1996

Learning and the Emergence of Coordinated Communication
Michael Oliphant and John Batali
Department of Cognitive Science, UCSD
Vol. 11, No. 1, February, 1997

Contexts That Pack a Punch: Lexical Class Priming of Picture Naming
Kara Federmeier and Elizabeth Bates
Department of Cognitive Science, UCSD
Vol. 11, No. 2, April, 1997

Lexicons in Contact: A Neural Network Model of Language Change
Lucy Hadden
Department of Cognitive Science, UCSD
Vol. 11, No. 3, January, 1998

On the Compatibility of CogLexicons in Contact: A Neural Network Model of Language Change
Mark Collier
Department of Philosophy, UCSD
Vol. 11, No. 4, June, 1998

Analyzing Semantic Processing Using Event-Related Brain Potentials
Jenny Shao
Department of Speech Pathology, Northwestern University
Helen Neville
Department of Psychology, University of Oregon
Vol. 11, No. 5, December 1998

Blending and Your Bank Account: Conceptual Blending in ATM Design
Barbara E. Holder
Department of Cognitive Science, UCSD
Vol. 11, No. 6, April 1999

Could Sarah Read the Wall Street Journal?
Ezra Van Everbroeck
Department of Linguistics, UCSD
Vol. 11, No. 7, November 1999

Introducing the CRL International Picture-Naming Project (CRL-IPNP)
Elizabeth Bates, et al.
Vol. 12, No. 1, May 2000.

Objective Visual Complexity as a Variable in Studies of Picture Naming
Anna Székely
Eotvos Lorand University, Budapest
Elizabeth Bates
University of California, San Diego
Vol. 12, No.2 , July 2000

The Brain's Language
Kara Federmeier and Marta Kutas
Department of Cognitive Science, UCSD
Vol. 12, No.3 , November 2000

The Frequency of Major Sentence Types over Discourse Levels: A Corpus Analysis

Frederic Dick & Jeffrey L. Elman

University of California, San Diego

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 the absolute as well as relative frequencies of major sentence types. The goal of the present work is to present initial findings from such a study. We report the results of an analysis of parsed versions of two written and one spoken corpus (Wall Street Journal, Brown, and Switchboard, respectively). Frequencies of six major types of grammatical structures were calculated: Active Declaratives, Passives, Subject Relatives, Subject Clefts, Object Clefts, and Object Relatives. We discuss both practical as well as theoretical implications of problems inherent in such an analysis.

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 structure (Bybee, 1995; Dick, Bates, Wulfeck, Utman, & Dronkers, 1999; Kempe & MacWhinney, 1999; MacDonald, 1997; MacDonald, 1999; MacDonald, Pearlmutter, & Seidenberg, 1994; McRae, Jared, & Seidenberg, 1990; Plunkett & Marchman, 1993; Plunkett & Marchman, 1996; St. John & Gernsbacker, 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 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, information theoretic approaches, as well as connectionist models (Manning & Schütze, 1999). Here again, the actual statistical patterns of language play an important role.[1]

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., 2000; Snodgrass & Vanderwart, 1980). There is also a lively debate regarding the role played by construction frequency in on-line processing of sentences with temporary syntactic ambiguities (e.g., Cuetos & Mitchell, 1988; Gilboy, Sopena, Clifton, & Frazier, 1995; Gibson & Schütze, 1999; Gibson, Schütze, & Salomon, 1996; MacDonald, 1994; Mitchell & Cuetos, 1991; Mitchell, Cuetos, Corley, & Brysbaert, 1996). But with the exception of these and a few other studies (e.g., Kempe & MacWhinney, 1999; Bybee, 1995),

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. This is primarily because the relative frequencies of sentence structures is more difficult to obtain, 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. Accordingly, 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.

Previous Work

Several researchers have used manual analyses of relatively small text samples as a way around this problem. 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 & 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.

Other authors have availed themselves of larger electronic corpora in estimating relative sentence type or grammatical construction frequencies. As part of a project examining the role of discourse context on verb subcategorization, Roland and Jurafsky (1998; in press) used subsets of three tagged and parsed online corpora: the Wall Street Journal, Brown (Marcus, Santorini, & Marcinkiewiz, 1993), and Switchboard (Godfrey, Holliman, & McDaniel, 1992). The materials for these corpora are drawn from news articles, mixed written material, and transcribed telephone conversations, respectively. The two written corpora contain at least one million words, with Switchboard containing a quarter as many words; all are fully parsed. The size of these databases allowed Roland and Jurafsky to estimate the subcategorization frequencies of up to 77 different verbs, and in addition compare their relative frequencies over discourse levels.

Goal of Current Work

We had two goals in undertaking the current study. First, we wished to extend the work begun by Roland and Jurafsky in order to cover a larger set of syntactic structures. In particular, we were interested in determining the relative frequencies of the following sentence types:

1. Active declaratives ("The dog is biting the cat")

2. Passives ("The cat was bitten by the dog")

3. Subject Relatives ("The dog who bites the cat barked loudly")

4. Subject Clefts ("It was the dog who bit the cat")

5. Object Clefts ("It was the cat who the dog bit")

6. Object Relatives (full: "The cat who the dog bites ran away"; reduced: "The cat the dog bites ran away")

These sentence types have been the focus of a large number of psycholinguistic studies, and have played a crucial role in characterizing language deficits in aphasic patients (Caplan, 1995; Caplan & Waters, 1999; Hickok & Avrutin, 1995; Just & Carpenter, 1992; Just, Carpenter, & Keller, 1996a; Just et al., 1996b; Stowe et al., 1998). More recently, it has been suggested by Dick, Bates, Wulfeck, Utman, Dronkers, and Gernsbacher (2000) that differential performance by aphasics on different sentence types might arise from differences in their relative frequency of occurrence in the language. Thus, we were particularly interested in seeing whether in fact there were significant differences in the frequency of these syntactic structures.

A secondary goal was to document and place into the literature some of the methodological issues that arise when conducting such a study. The most obvious methodological hurdle is the construction of the syntactic patterns used in searching for different sentence types. As we discuss later, the richness and variety of structures that are found in natural language often make it difficult to formulate a structural pattern that cleanly identifies all and only those examples that are desired. Indeed, we found many cases in which sentences were identified that were not, strictly speaking, examples of our target structures but were similar enough that they were picked up by our patterns. This has interesting consequences not only for the search process, but raises questions about implications for processing models. We address this and other issues in the Discussion.[2]

The Current Study

General Considerations

Like Roland and Jurafsky (1998) we used the Wall Street Journal, Brown, and Switchboard corpora. Although these corpora are not as large in size as one might want, their

availability (through the Linguistic Data Consortium, Univ. of Pennsylvania) and relatively accurate parses make them good starting points for a study of this sort. The Wall Street Journal corpus is a compilation of articles from the Dow Jones Newswire in the 1980's; the Brown corpus draws from several different written texts, including newspaper articles, fiction, and technical writing from the early 1960's (this is the corpus first developed by Francis and Kuçera (1982) for their lexical analyses). The Switchboard database is composed exclusively of telephone conversations between strangers taped by Bell Labs in the early 1990's (with the consent of the participants).

Tgrep Search Patterns

The three electronic corpora we used were parsed to yield a standard phrase structure analysis, in which all grammatical constituents in a sentence (including null elements such as "traces") appear as nodes in the tree. Searches were carried out using the **tgrep** program that is included in the Treebank distribution[3]. **Tgrep** is a tree-oriented search program (analogous to UNIX **egrep**, but sensitive to dominance and precedence relationships) that allows one to selectively extract all examples of a particular sentence type by searching for the grammatical structure underlying it. This was the *modus operandi* for all the full-corpus analyses. Search strings for the different sentence types are shown in the Appendix.

As will become evident in the following section, in many cases neither written nor spoken discourse adheres very closely to tidy linguistic abstractions. In creating and revising search strings for the different grammatical constructions, we were often confronted with (a) sentence types whose grammatical parse was identical to our target sentence type, but which intuitively really did not belong to the target construction; or (b) sentences that were actually examples of our target construction, but deviated in some idiosyncratic way from the canonical version, and were therefore missed in the search. Obviously, increasing the specificity of the pattern would minimize errors of type (a), but at the expense of increasing the number of sentences missed (b). Unfortunately, there was often no pattern that got things just right.

We dealt with the above problem in two ways. First, we adopted an iterative strategy in developing the search strings. For any given sentence type, we began with the most general version of a pattern that would be (reasonably) guaranteed to identify all possible examples of that construction. This typically generated vastly more sentences than were appropriate, and required extensive hand-checking, followed by a number of iterations in which the search pattern was refined. However, by casting the net wide, we discovered many more sentence subtypes and structural variations than we would have anticipated had we begun with the most specific and narrowest definition of the sentence type, using preconceived notions of what the appropriate structure would be. All sentence types ultimately required more than one search pattern to identify the—hopefully—complete set of examples of a given structure.

Second, in addition to the hand-checking of the sentences that were identified by this automatic procedure, we cross-validated one of the searches (for Passives; see below for details) by carrying out a manual analysis of a sub-sample of the Brown corpus. This analysis involved $1/50^{th}$ of the total corpus and produced results that were closely in agreement with the automatic search, thereby providing evidence for the accuracy of the automatic process.

A final complication arose because the parsing style used in the three corpora differs to some extent. In reporting the results for each sentence type, we therefore not only note specific problems and issues that arose in defining each type, but also indicate (with numbers in curly brackets (e.g. {1a-c}) which **tgrep** search pattern was used to generate the result. All **tgrep** search patterns can be found in the Appendix, with additional notes on specific methodological points.

Sentences vs. Exemplars

For most sentence types, we include two counts: (1) the number of complete sentences in each corpus containing at least one example of a particular type, and (2) the total number of exemplars of a type in each corpus. For example, the count for "The dog was hit by the cat and the goose was bitten by the mouse" would be two exemplars and one sentence. It is fairly common for the frequency of a particular grammatical construction to be expressed in terms of the total number of sentences in the analyzed corpus (e.g. Roland & Jurafsky, 1998). However, the sentence *per se* has a very nebulous definition in many discourse situations (particularly spoken). Therefore, we report most results in terms of "exemplar ratios" (i.e., the number of exemplars of construction A) : (the number of exemplars of construction B) rather than relying on percent of total sentences per corpus.

Results

For all results, the reader should refer to the relevant exemplar/sentence counts in Table 1, with percentages/ratios in Table 2. For each sub-result, we have also listed each relevant **tgrep** search command in the *Appendix*, indexed as {1a-c}, {2a-c}, etc.. For several sentence types (Active SVs and OVS citation forms), we have reported results only from the Wall Street Journal and Switchboard, and not from the Brown corpus. This is due

to coding differences for the latter that did not permit us to sufficiently constrain our searches. In general, we briefly comment on the results for each subsearch, then make comparisons between larger sets of structures (e.g. all actives vs. passives, SV vs. OVS word orders).

Table 1: Total number of exemplars per syntactic form. In Tables 1 and 2, cells containing an 'x' are those where a search was either inappropriate or not feasible - see Results for full details).

SYNTACTIC FORM	WSJ exemplar	WSJ sentences	Brown exemplar	Brown sentences	SWBD exemplar	SWBD sentences
SVO	19,947	18,893	29,576	25,341	5,975	5,733
Predicate Nominals (SV)	11,879	x	x	x	7,733	x
Predicate adverbials (SV)	1,403	x	x	x	1,176	x
SVs + preposition	13,773	x	x	x	1,623	x
Passives	7,813	7,326	10,435	9,707	587	575
Inverted Quotation/Citation	1,485	x	x	x	0	x
Subject Relatives with infinitival clauses	2,999	2,838	x	x	593	554
Subject Relatives without infinitival clauses	2,415	2,305	4,164	3,748	501	466
Reduced Object Relatives	599	584	1,286	1,246	124	120
Unreduced Object Relatives	195	188	504	488	187	180
Subject Cleft/Object Cleft	x	40/3	x	x	x	12/1
Total Words	1,102,156		1,002,898		240,041	
Total Sentences	49,208		48,094		20,794	
Mean Words per sentence	22.4		20.85		11.5	

Total Sentences and Words

The total number of sentences in the Wall Street Journal (WSJ) and Brown corpora was almost identical (~50,000); sentences (or "utterances") in the Switchboard (SWBD) corpus numbered about half this (~20,000) {1a-c}. Both written corpora contained approximately the same number of words (~1,000,000), with Switchboard containing about 1/4th as many words{2a-c}; hence, an average sentence from either written corpus had twice the number of words (~20) as did a spoken sentence (~10).

Actives and Passives

We first compared relative frequencies of Active and Passive constructions over discourse context. Actives were perhaps the most difficult to pin down, as their general form often overlapped with other forms, such as truncated Passives. (We avoided the latter by excluding NPs that immediately dominated "traces", coded as (-NONE-)). We first divided our Actives search into transitives of the form Subject-Verb-Object (SVO), e.g. "The boy flunked the exam" {3a-c}, and intransitives of the general form Subject-Verb (SV); these we further divided into three subgroups: nominal or adjectival predicates ("The boy is a student" or "The girl is smart"){4a-b}, adverbial predicates ("The boy walks quickly"){5a-b}, and intransitive prepositionals ("The boy went to the store"){6a-b}. For intransitives, we partially relied on the coding for "predicate" (PRD) provided in WSJ and SWBD; also useful were the various codings of prepositions (see {6a-b}). For all SV/SVO subcategories, we used a family of search strings that

allowed for verb compounds, such as auxiliaries, modals, and so forth ("The boy could have gone to the store").[4] For all Active constructions, we counted exemplars only, as a sentence metric would have seriously underestimated their frequency - for example, sentences in both written corpora often contained multiple Active exemplar.

Like the Active construction, Passives do not take a single immutable form. Hence, we included in our search not only full Passives of the form Object-Verb-Subject (OSV- "The dog was eaten by the cat"), but truncated or agentless Passives of the form Object-Verb (OV - "The dog was eaten"), as well as "get" Passives (OVS - "The dog got eaten by the cat" and OV - "The dog got eaten") {7a&b}. As with Actives, we allowed for verb compounds, variations in tense, and so forth.[5,6]

Table 2: Ratios of Syntactic Forms, in exemplars

Syntactic Comparison	WSJ	Brown	SWBD
Passives : Active (SVO only)	1 : 2.55	1 : 2.83	1 : 10.18
Passives : Actives (SVs plus SVOs)	1 : 6.02	x	1 : 28.12
Passives: Actives plus Subject Relatives (without infinitives)	1 : 6.33	x	1 : 28.98
OVS (Passives plus Inv. quotations) : SVO (Actives plus Subject Relatives w/o inf)	1 : 5.32	x	1 : 28.98
Object Relative (reduced plus unreduced) : Subject Relative (w/o inf)	1 : 3.04	1: 2.33	1 : 1.61
SVO (Actives plus Subject Relatives) : OSV (Object Relatives)	1 : 62.23	x	1 : 54.69

Previous work (reviewed in St. John and Gernsbacher, 1998) pointed to Active/Passive ratios of about 35:1 to 10:1 for spoken discourse (Goldman-Eisler & Cohen, 1970) and 6:1 for written text (Taylor & Taylor, 1983). When we compare Passives to Active transitives (SVO) only, SWBD falls at the lower end of this spectrum, with an SVO/Passive ratio ~10:1; when we include all Active constructions, this ratio increases to ~29:1. In the Brown and WSJ corpora, Passives were somewhat more frequent than we expected, with SVO/Passive ratios of ~2.5 and ~2.8 : 1; for WSJ, the ratio of all Actives/Passives was ~6 : 1.

Another OVS Construction - Inverted Quotations

As mentioned in the Introduction, there is both computational and behavioral evidence suggesting that experience with a particular structure tends to generalize to performance on other structures with similar forms. One of the few syntactic structures in English that does pattern with the OVS word order of Passives is the inverted quotation (e.g., "I did not use any subliminable advertising," said the candidate). This structure's use is absolutely determined by discourse context (as can be seen in Table 1): Newspapers (WSJ) use this construction a fair amount, whereas in telephone conversations (SWBD) they are non-existent {8a&b}. Thus, the impact of these constructions on the overall OVS/Active ratio is small in the case of WSJ (with the ratio moving from ~1 : 6.3 with passives only to ~ 1. 5.3 with passives and inverted quotations), and of course no change in SWBD.

Subject Relatives

As noted above, these sentences are of the form "The *subject* that/who *verbs*," thus sharing the SV word order with actives. We began our search for both subject and object relatives by using a pattern that matched a complex NP—tgrep -w '(NP < (NP . SBAR))'—then narrowed our search using the coding patterns suggested by the subset of target sentences that were found with the broader pattern. The subject relative search string for both WSJ and SWBD {9a,c} explicitly excludes an interesting infinitival construction very similar to relatives, e.g., "they have things to gripe about" and "he has a family to take care of". This construction is coded by WSJ and SWBD (but not Brown {9b}) in the same form as Subject Relatives, but in fact is more similar to an Object Relative - .e.g., "they have

things that they gripe about" and "he has a family that he takes care of." We list the totals for WSJ/SWBD including these infinitival constructions in {10a-b}. We excluded in both these counts an additional construction also similar to relatives, namely "The *noun* whose *noun verbed*" (e.g., "The man whose dog talked won the prize."); these were relatively infrequent (51 exemplars in WSJ, and 1 in SWBD).

Total Object Relatives - Reduced and Full

Full Object Relatives {11a-c}are of the form "The *object* that the *subject verbs*", with reduced relatives omitting the intervening "that" {12a-c}. Our search strings included locatives and prepositionals such as "the place (that) I work at", and "the people I work with". However, we did not include comparatives such as "That was more than I could handle", which is quite similar in structure to an Object Relative. When we compared the frequency of subject and object relatives, we found that, in contrast to what one might have predicted, there was a relatively small difference in subject/object relative frequency , particularly in SWBD (see Table 2).

Subsets of Relatives: Subject and Object Clefts

We used the "S-CLF" tag provided in WSJ and SWBD as a first-pass search for both these forms; because there were very few, we performed a fine-grained hand search. Both subject and object clefts were extremely rare, with 40 subject and 3 Object Clefts in WSJ, and 12 subject and 0 Object Clefts in SWBD.

Discussion

Before considering what we have learned from this study, we point out several limitations to the current work.

First, we regard the problem of finding precise structural definitions for the target constructions to be an open issue. The frequency counts we report must be interpreted with this in mind. At the same time, we do not believe the numbers will change substantially as a result of alternative structural definitions.

Second, our approach has been somewhat coarse-grained, in that there are additional factors we did not take into account that are undoubtedly of interest. For example, we do not consider possible interactions between the frequency of a given structure and main verb tense. Thus, the frequency we report for Passive constructions collapses across all tenses, although intuitively it is likely that the Passive is more frequently used with verbs in the past than in the present tense. Similarly, we do not investigate whether there are interactions between specific lexical items and structures, although we know from other work that (for example) verbs differ with regard to their subcategorization preferences (Connine, Ferreira, Jones, Clifton, & Frazier, 1984; Garnsey, Pearlmutter, Meyers, & Lotocky, 1997; Hare, McRae, & Elman, 2000; Roland & Jurafsky, 1998; Roland & Jurafsky, in press). Some of this work also suggests that comprehenders are sensitive to such contingencies. More generally, it seems clear that syntactic ambiguity resolution can be highly influenced by such lexical factors (MacDonald et al., 1994). Thus, it would be useful in the future to refine the analyses here by calculating statistics separately for these contingencies.

Bearing these qualifications in mind, there are a number of important results that emerge from the present analyses:

- Like Roland and Jurafsky (1998), we find that frequencies vary with corpus, probably reflecting discourse and register factors (cf., Chafe, 1982). In some cases, the differences between corpora are extreme. The inverted quotation occurs 1,485 times in the WSJ, but not at all in SWBD. Such differences may also occur within a corpus. Thus, in our hand count of a subset of Brown, we noted that in several of the technical articles, almost every sentence is in the Passive voice. In narrative prose (e.g., short fiction) on the other hand, the Passive is relatively rare. This suggests it might be informative to break down future analyses of Brown by genre, although the generality of such results may be limited by the relatively small number of examples in the data set. In any case, to the extent that different registers and discourse styles are associated with different choices of syntactic structures, these differences are likely to have processing implications, and the behavior of subjects in experimental situations may or may not be well-predicted by usage statistics that are appropriate to different contexts.

- As we noted in the introduction, given the complex—and often messy—nature of most sentences encountered in naturally occurring written and spoken language, it turns out to be exceedingly difficult to formulate a precise structural definition of most of the syntactic constructions studied here. Nor is it clear that every construction has a single structural description. We feel that the method we used—overgeneration followed by winnowing—was the most likely to identify all examples of the target structures, but there may both outright errors, as well as marginal cases whose inclusion or exclusion remains an open issue. (The problem is compounded by the use of different parsing conventions in the different corpora, as well as occasional and unavoidable misparses and inconsistencies.)

- We were surprised by the relative frequency of some constructions that we had predicted to have been less common. The difference between Subject and Object Relatives was not nearly as great as we anticipated; in fact, in the spoken corpus (SWBD), their frequencies of occurrence are virtually identical (the OR:SR ratio is 1:1.61). The difference between Passives and Transitive Actives was also smaller than expected (in WSJ, 1:2.55; in Brown, 1:2.83). Although this latter pattern changes when the larger set of SV constructions is considered, it is not clear that that broader set is incorporated when Active/Passive differences are studied (e.g., St. John & Gernsbacher, 1998).

- A related issue arises when one attempts to compare the relative frequencies of narrowly defined constructions, such as Subject Relatives vs. Object Relatives. If one looks only at examples of these two constructions in SWBD, one finds a 1.9:1 ratio of Subject to Object Relatives, which is much less than might be expected. But if one includes the broader classes of constructions that share either the SVO or OSV word order, the ratio is 55:1.

- The above observations represent not merely as technical issues, i.e., the difficulty of finding a precise structural definition that exists in principle but in practice is difficult to formulate. The problem is, rather, that there are not always crisp boundaries between what counts as an instance of one construction and what counts as another. For example, the parses of "It was more than he asked for" and "It was the book that he asked for" are virtually identical, and the two structures can only be distinguished on the basis of the specific lexical fillers of the IN constituent. This difficulty has very important consequences for theories in which frequency of constructions figures in explanations of processing difficulty. It is quite likely, for example, that a language user's experience with one construction that structurally overlaps another may facilitate the processing of both. Thus, one construction that occurs with very low frequency of occurrence (e.g., Subject Clefts) may not lead to processing difficulties if there also exist structurally similar constructions (e.g., Active Declaratives, Subject Relatives) that are more frequent (cf., Grodzinsky, 2000)). In a related vein, the disparity between the relative frequencies of the word orders used by Passives (OVS) and Object Relatives (OSV) is echoed in new results from studies of aphasic patients and of college students processing under adverse conditions (Dick et al. 2000). Here, Passive comprehension, although impaired relative to that of SVO sentences , is more preserved than is comprehension of Object Clefts, which pattern with Object Relatives. However, as we note above, comprehension of Subject Clefts, which are equally as rare, but pattern with the frequent SVO word order, is relatively preserved.

Such considerations may also lead to what appears to be the precocious acquisition of rare structures. Chomsky (1975) and Crain (1991), among others, have argued that such patterns of acquisition—"language acquisition in the absence of experience"—provide strong evidence for innate linguistic constraints. Alternatively, it may be that this phenomenon reflects generalization from exposure to other, more frequent constructions. This hypothesis is currently being explored in our laboratory (Dick, in preparation; Lewis, in preparation).

References

Bates, E., Federmeier, K., Herron, D., Iyer, G., Jacobsen, T., Pechmann, T., D'Amico, S., Devescovi, A., Wicha, N., Orozco-Figueroa, A., Kohnert, K., Gutierrez, G., Lu, C. C., Hung, D., Hsu, J., Tzeng, O., Andonova, E., Gerdjikova, I., Mehotcheva, T., Székely, A., & Pléh, C. (2000). Introducing the CRL International Picture-Naming Project (CRL-IPNP). *CRL Newsletter, 12*(1).

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In E. Brian MacWhinney & et al. (Eds.), *Mechanisms of language acquisition*. (pp. 157-193): Hillsdale, NJ, USA.

Bybee, J. (1995). Regular morphology and the lexicon. *Language & Cognitive Processes, 10*(5).

Caplan, D. (1995). Language disorders. In E. Robert L. Mapou, E. Jack Spector, & et al. (Eds.), *Clinical neuropsychological assessment: A cognitive approach*. (pp. 83-113): New York, NY, USA.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral & Brain Sciences, 22*(1), 77-126.

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language*. (pp. 1-16). Norwood, NJ: Ablex.

Chomsky, N. (1975). *Reflections on language*. New York: Parthenon Press.

Connine, C., Ferreira, F., Jones, C., Clifton, C., Jr., & Frazier, L. (1984). Verb Frame preferences: Descriptive norms. *Journal of Psycholinguistic Research, 13*, 307-319.

Crain, S. (1991). Language acquisition in the absence of experience. *Brain and Behavioral Sciences, 14*, 597-611.

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

Dick, F. (in preparation). The effects of training frequency on syntactic comprehension. .

Dick, F., Bates, E., Wulfeck, B., Utman, J., & Dronkers, N. (1999). *Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasics and normals*. (Technical Report 9906). La Jolla, CA: CRL.

Dick, F., Bates, E., Wulfeck, B., Utman, J., Dronkers, N., & Gernsbacher, M. A. (2000). Language deficits, localization, and grammar: evidence for a distributive model of language breakdown in aphasics and normals. *Submitted*.

Francis, W., & Kuçera, H. (1982). *Frequency analysis of English usage*. Boston: Houghton Mifflin.

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

Gibson, E., & Schütze, C. (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 complexity of linguistic structures. *Journal of Psycholinguistic Research, 25*, 59-92.

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

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

Goldberg, A. E. (1995). *Constructions: a construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldman-Eisler, F., & Cohen, M. (1970). Is N, P, and NP difficulty a valid criterion of transformational operations? *Journal of Verbal Learning and Verbal Behavior, 9*, 161-166.

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

Hare, M., McRae, K., & Elman, J. (2000). *Sense and structure: Meaning as a determinant of verb subcategorization preference*. Paper presented at the CUNY Sentence Processing Conference, La Jolla, CA.

Harris, Z. S. (1954). Distributional structure. *Word, 10*.

Hickok, G., & Avrutin, S. (1995). Representation, referentiality, and processing in agrammatic comprehension: Two case studies. *Brain & Language, 50*(1), 10-26.

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

Just, M. A., Carpenter, P. A., & Keller, T. A. (1996a). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review, 103*(4), 773-780.

Just, M. A., Rep, M., Carpenter, P. A., van Dijl, J. M., Keller, T. A., Suda, K., Eddy, W. F., & Schatz, G. (1996b). Brain activation modulated by sentence comprehension. *Science, 274*(5284).

Kempe, V., & MacWhinney, B. (1999). Processing of morphological and semantic cues in Russian and German. *Language & Cognitive Processes, 14*(2).

Lewis, J. (in preparation). *Learning the unlearnable: The problem of AUX inversion*. (CRL Technical Report). La Jolla: Center for Research in Language.

MacDonald, M. C. (1997). Lexical representations and sentence processing: An introduction. *Language & Cognitive Processes, 12*(2&3).

MacDonald, M. C. (1999). Distributional information in language comprehension, production, and

acquisition: Three puzzles and a moral. In E. Brian MacWhinney & et al. (Eds.), *The emergence of language*. (pp. 177-196): Mahwah, NJ, USA.

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

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

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

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

Mitchell, D. C., & Cuetos, F. (1991). *The origins of parsing strategies*. Paper presented at Current Issues in Natural Language Processing Conference.

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1996). Exposure-based models of human parsing: Evidence for the use of coarse-grained (non-lexical) statistical records. *Journal of Psycholinguistic Research, 24*, 469-488

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48*(1).

Plunkett, K., & Marchman, V. A. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition, 61*(3).

Roland, D., & Jurafsky, D. (1998). *How Verb Subcategorization Frequencies are affected by Corpus Choice*. Paper presented at the COLING/ACL.

Roland, D., & Jurafsky, D. (in press). Verb Sense and Verb Subcategorization Probabilities. In S. Stevenson (Ed.), *1998 CUNY Sentence Processing Conference* . Philadelphia: Benjamins.

Sag, I. A., & Wasow, T. (1999). *Syntactic theory : a formal introduction*. Stanford, CA: Center for the Study of Language and Information.

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., & Gernsbacker, M. A. (1998). Learning and losing syntax: Practice makes perfect and frequency builds fortitude. In E. Alice F. Healy, E. Lyle E. Bourne Jr, & et al. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention*. (pp. 231-255): Mahwah, NJ, USA.

Stowe, L. A., Broere, C. A., Paans, A. M., Wijers, A. A., Mulder, G., Vaalburg, W., & Zwarts, F. (1998). Localizing components of a complex task: sentence processing and working memory. *Neuroreport, 9*(13), 2995-2999.

Taylor, I., & Taylor, M. M. (1983). *The psychology of reading*. New York: Academic Press.

Footnotes

[1] Of course, interest in the distributional statistics of language is not new, and has figured prominently in many pre-generative linguistic theories (e.g., Harris, 1954). However, it has only been with the recent advent of fast computers and inexpensive mass storage devices that serious investigations of large scale samples of language have been possible. In particular, the existence of electronic versions of encyclopedias and books, as well as completely new text media (e.g., USENET, email) offers exciting new opportunities for relating detailed patterns of language usage to processing.

[2] Undoubtedly, further refinement and adjustments in the methodology will be useful. Thus, we have placed our search patterns both in the Appendix, and at http://crl.ucsd.edu/corpora/tgrep_patterns.html so that others may more easily experiment with them. We encourage anyone who does so to communicate their experiences with us (fdick@cogsci.ucsd.edu or elman@cogsci.ucsd.edu); in particular, we welcome any modifications or extensions.

[3] A LINUX port of **tgrep** is available at the Center for Research in Language website: http://crl.ucsd.edu/software.

[4] We verified that all our search patterns were mutually exclusive by using the LINUX program **diff** to compare the output of the alternative pattern searches.

[5] The passive search strings are the creation of Dan Jurafsky and Doug Roland; we thank them for allowing us to use and report them.

[6] Because the SVO/Passive ratio was lower than might have been expected, we verified our results by perform a manual analysis of approximately 1/50th of the Brown corpus (1,000 sentences). The sample we drew was composed of 100-200 sentence "chunks", drawn randomly from the corpus. We classified each sentence as either Active (where the sentence had a Subject-Verb-Object or Subject-Verb order) or Passive (as classified by the same standards we used for the TGREP analysis). Any ambiguous sentences or sentences that were in different word orders (such as imperatives) were not included in the Active/Passive count. All sentences that contained a Passive construction (which was often embedded in an Active frame) were counted as "Passives"; we did not count multiple exemplars within sentences as tokens. This method necessarily undercounts Actives, as many Active and Passive sentences tend to be composed of multiple clauses, all in the Active voice. Therefore, the hand count was the most liberal estimate of Passives, and the most conservative with reference to our theoretical hypotheses.

Results are the following: 746 sentences were classified as Actives (where at least one SV or SVO construction was found), and 194 sentences were classified as Passives (where any sentence containing a Passive construction was counted only as Passive)[i] Hence, the Passive:Active ratio for the 940 sentences counted (with 60 sentences ignored because of the above criteria) was 1:3.85. In order to better match our automated counts, we then estimated the ratio of SV to SVO constructions in this sample by classifying a 200-sentence subset (where a sentence with at least one SVO construction was classified as SVO). This subset was composed of sentences containing roughly half SVOs, and half SVs (with 121 SVOs/79 SVs). We then extrapolated to the 1000-sentence set by multiplying this ratio with the original ~1:3.85 ratio, thereby arriving at a 1:2.33 Passive:SVO ratio - one very similar to our Brown **tgrep** analysis.

Unsurprisingly, the distribution of Passives differed dramatically over the texts in the Brown corpus, with some sections (fallout shelter instructions, technical research on cutting surfaces, political reports) containing 50%+ Passive constructions, while others (stories, historical descriptions) containing very few or no Passives whatsoever. The great disparity in the proportion of Passives vs. Actives parallels the qualitative difference between the written (WSJ/Brown) and oral (SWBD) corpora, and is in keeping with the observations of Roland & Jurafsky (1998), who show that levels of discourse differ dramatically over a wide range of variables, from gross numbers of sentence types to the relative distribution of verb subcategorization frames.

Appendix

All strings are preceded by the resulting count in [ ] - this is not part of the command line. We only list strings that produced at least one exemplar in a corpus, although for searches that involved increasing numbers of auxiliaries (such as the SV strings) we continued to add auxiliaries for several iterations after the initial "zero" count to assure that we did not miss any unusually long and complex constructions.

Total number of sentences: (See note (a))

1a - (WSJ): [49,208] tgrep -n 'TOP' | sed -e '/^$/d' | wc

1b - (Brown): [48,094] tgrep -n 'TOP < S ' | sed -e '/^$/d' | wc

1c - (SWBD): [20,794] tgrep -n 'TOP < /S/' | sed -e '/^$/d' | wc

Total number of words in "raw" files: (See note (b))

2a - (WSJ): [1,102,156] cat */* | egrep -v START | egrep -v '^$' | tr -d "[:punct:]" | wc

2b - (Brown): [1,002,892] cat brown.raw | sed -e 's/ T / /g' | sed -e 's/ 0 / /g' | sed -e 's/-[A-Z]*-/ /g' | tr -d "[:punct:]" | egrep -v '^$' | sed -e 's/pseudoattach//g' | wc

2c - (SWBD): [240,041] cat swbd_raw.crp | sed -e 's/*[A-Z]**-[0-9]//g' | sed -e 's/._.$//' | egrep -v Speaker | egrep -v '^#' | egrep -v '^$' | tr -d "[:punct:]" | sed -e 's/ s /s /g' | sed -e 's/ nt /nt /g' | sed -e 's/[0-9]//g' | wc

Total number of active (SVO) [exemplars/sentences] (e.g. "The man stroked the cat") : (See note (c). Search strings for sentences differ from exemplars only in the substitution of the -n option for -an.)

3a - (WSJ):

[11241/10556] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (NP !< -NONE-)))' | sed -e '/^$/d' | wc

[7763/7399] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (NP !< -NONE-))))' | sed -e '/^$/d' | wc

[913/909] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (NP !< -NONE-)))))' | sed -e '/^$/d' | wc

[28/27] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (NP !< -NONE-))))))' | sed -e '/^$/d' | wc

[2/2]tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (VP < (NP !< -NONE-)))))))' | sed -e '/^$/d' | wc

3b - (Brown):

[26177/22003] tgrep -an '(S < (NP !< (-NONE-)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (NP !< -NONE-)))' | sed -e '/^$/d' | wc

[3294/3234] tgrep -an '(S < (NP !< (-NONE-)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (NP !< -NONE-))))' | sed -e '/^$/d' | wc

[98/97] tgrep -an '(S < (NP !< (-NONE-)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (NP !< -NONE-)))))' | sed -e '/^$/d' | wc

[7/7] tgrep -an '(S < (NP !< (-NONE-)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (VP !< (/VB/ < (/is|was|am|were|be|being|been/)) < (NP !< -NONE-))))))' | sed -e '/^$/d' | w

3c - (SWBD):

[3227/3076] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (NP !< -NONE-)))' | sed -e '/^$/d' | wc

[2483/2395] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (NP !< -NONE-))))' | sed -e '/^$/d' | wc

[252/249] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (NP !< -NONE-)))))' | sed -e '/^$/d' | wc

[13/13] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (NP !< -NONE-))))))' | sed -e '/^$/d' | wc

Total number of SV Nominal/Adjectival Predicates (e.g. "The boy is a student" or "The girl is smart")

4a - (WSJ):

[9714] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < /PRD/ !< (-NONE-)))' | sed -e '/^$/d' | wc

[2023] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < /PRD/ !< (-NONE-))))' | sed -e '/^$/d' | wc

[138] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < /PRD/ !< (-NONE-)))))' | sed -e '/^$/d' | wc

[4] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < /PRD/ !< (-NONE-))))))' | sed -e '/^$/d' | w

4b - (SWBD):

[6907] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < /PRD/ !< (-NONE-)))' | sed -e '/^$/d' | wc

[772] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < /PRD/ !< (-NONE-))))' | sed -e '/^$/d' | wc

[53] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < /PRD/ !< (-NONE-)))))' | sed -e '/^$/d' | wc

[1] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < /PRD/ !< (-NONE-))))))' | sed -e '/^$/d' | wc

Total number of Adverbial Predicates (e.g. “The boy walks quickly”)

5a - (WSJ):

[862] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP !<< /NP/ < /ADV/ !< (-NONE-)))' | sed -e '/^$/d' | wc

[473] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-))))' | sed -e '/^$/d' | wc

[62] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-)))))' | sed -e '/^$/d' | wc

[6] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-))))))' | sed -e '/^$/d' | wc

5b - (SWBD):

[860] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP !<< /NP/ < /ADV/ !< (-NONE-)))' | sed -e '/^$/d' | wc

[282] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-))))' | sed -e '/^$/d' | wc

[30] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-)))))' | sed -e '/^$/d' | wc

[4] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP !<< /NP/ < /ADV/ !< (-NONE-))))))' | sed -e '/^$/d' | wc

Intransitive Prepositionals (e.g. “The boy went to the store”)

6a - (WSJ):

[8482] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP)))' | sed -e '/^$/d' | wc

[4666] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP))))' | sed -e '/^$/d' | wc

[594] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP)))))' | sed -e '/^$/d' | wc

[29] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP))))))' | sed -e '/^$/d' | wc

[2] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP)))))))' | sed -e '/^$/d' | wc

6b - (SWBD):

[933] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP)))' | sed -e '/^$/d' | wc

[594] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP))))' | sed -e '/^$/d' | wc

[92] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP)))))' | sed -e '/^$/d' | w

[4] tgrep -an '(S < (NP-SBJ !< (-NONE-)) < (VP < (VP < (VP < (VP < (PP-CLR|PP-LOC|PP-DIR|PP-EXT|PP-PRP|PP-1|PP-MNR|PP--TPC|PP-TMP))))))' | sed -e '/^$/d' | w

Total number of passive (OVS) exemplars and sentences (e.g. "The bird was fed by t - also, command lines were the same for all corpora, and therefore are listed only once. Totals for each corpus are listed in the same order as the search strings; strings for sentences differ from exemplars only in the substitution of the "-n" option for the "-an" options.

7 - passive search strings for all corpora

tgrep -an '(VP<VBN\!<PRT\!>NP%(AUX<< (/is|are|was|were|be|am|been|get|gets|got|gotten|getting|being/)))' | sed -e '/^$/d' | wc

tgrep -an '(VP<VBN\!<PRT\!>NP%(/VB/<(/is|are|was|were|be|am|been|get|gets|got|gotten|getting|being/)))' | sed -e '/^$/d' | wc

tgrep -an '(VP<VBN\!<PRT\!>NP<(VP%(/VB/<(/is|are|was|were|be|am|been|get|gets|got|gotten|getting|being/))))' | sed -e '/^$/d' | wc

tgrep -an '(VP<VBN\!<PRT\!>NP%(VP<(/VB/<(/is|are|was|were|be|am|been|get|gets|got|gotten|getting|being/))))' | sed -e '/^$/d' | w

7a - totals for passive exemplars

WSJ	Brown	SWBD
0	4910	0
6571	4008	440
1223	1452	139
19	65	8

Exemplar totals for all passive searches:

7813	10435	587

7b - totals for passive sentences

WSJ	Brown	SWBD
0	4537	0
6114	3708	431
1193	1399	136
19	63	8

Sentence totals for all passive searches:

7326	9707	575

Number of Inverted Quotations (OVS) (e.g. "'Alpha movement is sci

8a - (WSJ):

[1485] tgrep -an '(SINV < (VP < (S < (-NONE-))))' | sed -e '/^$/d' | wc

8b - (SWBD):

[00] tgrep -an '(SINV < (VP < (S < (-NONE-))))' | sed -e '/^$/d' | wc

Number of Subject Relatives, not including infinitival clauses [exemplars/sentences]

9a - (WSJ): (See note (d))

[2415/2305] tgrep -an '(NP < (NP . (SBAR < (/WHNP/|/WDT/ !< WP\$) < (S < (/NP-SBJ/ < -NONE-) !< (VP < (TO))))))' | sed -e '/^$/d' | wc

9b - (Brown):

[4164/3748] tgrep -an '(NP < (NP . (SBAR < WHNP|WDT < (S <1 (NP < -NONE-)))))' | sed -e '/^$/d' | wc

9c - (SWBD):

[501/466] tgrep -an '(NP < (NP . (SBAR < (/WHNP/|/WDT/ !< WP\$) < (S < (/NP-SBJ/ < -NONE-) !< (VP < (TO))))))' | sed -e '/^$/d' | wc

Number of subject relatives, including infinitival clauses [exemplars/sentences]

10a - (WSJ):

[2999/2838] tgrep -an '(NP < (NP . (SBAR </WHNP/|/WDT/ < (S < (/NP-SBJ/ < -NONE-)))))' | sed -e '/^$/d' | wc

10b - (SWBD):

[593/554] tgrep -an '(NP < (NP . (SBAR </WHNP/|/WDT/ < (S < (/NP-SBJ/ < -NONE-)))))' | sed -e '/^$/d' | wc

Number of unreduced object relatives [exemplars/sentences] (See note (e))

11a - (WSJ): (See note (f))

[195/188] tgrep -an 'NP < (NP . (SBAR < (/WHN/ !< -NONE-) < (S < (NP-SBJ !<< -NONE-) < (VP << ((NP !> S) < -NONE-)))))' | sed -e '/^$/d' | wc

11b - (Brown):

[504/488] tgrep -an '(NP < (NP . (SBAR < WHNP|IN|WDT < (S < (NP !<< -NONE-) < (VP << ((NP !>S) < -NONE-))))))' | sed -e '/^$/d' | wc

11c - (SWBD):

[187/180] tgrep -an 'NP < (NP . (SBAR < (/WHN/ !< -NONE-) < (S < (NP-SBJ !<< -NONE-) < (VP << ((NP !> S) < -NONE-)))))' | sed -e '/^$/d' | wc

Number of reduced object relatives [exemplars/sentences] - see note (e)

12a - (WSJ):

[599/584] tgrep -an 'NP < (NP . (SBAR < (/WHN/ < -NONE-) < (S < (NP-SBJ !<< -NONE-) < (VP << ((NP !> S) < -NONE-)))))' | sed -e '/^$/d' | wc

12b - (Brown):

[1286/1246] tgrep -an '(NP < ((NP !< WP) . (SBAR < -NONE- < (S < (NP !<< -NONE-) < (VP << ((NP !>S) < -NONE-))))))' | sed -e '/^$/d' | wc

12c - (SWBD):

[124/120] tgrep -an 'NP < (NP . (SBAR < (/WHN/ < -NONE-) < (S < (NP-SBJ !<< -NONE-) < (VP << ((NP !> S) < -NONE-)))))' | sed -e '/^$/d' | wc

Number of Subject and Object Cleft sentences (See note (g))

13 a&b (WSJ and SWBD) :

tgrep -an 'S-CLF' < screening_file

WSJ	
Subject Clefts	40
Object Clefts	3
SWBD	
Subject Clefts	12
Object Clefts	0

NOTES:

(a) We did not include "< S" in the WSJ string as it removed many sentence-like fragments.

(b) All three corpora had additional information in the "raw" files that needed to be deleted before a word count could be performed. Brown and Switchboard were especially difficult in this regard. Following is a guide to the various command lines (note that the particular uses of **sed** and **tr** vary depending on LINUX distribution).

Brown: (files from /treebank/tb1_075/raw/brown)

sed -e 's/ T / /g'

Remove codings of traces

sed -e 's/ 0 / /g'

Remove codings of traces

sed -e 's/-[A-Z]*-/ /g'

Remove a few odd grammatical codes of form -LRB-

tr -d "[:punct:]"

Remove all punctuation

egrep -v '^$'

Remove empty lines

sed -e 's/pseudoattach//g'

Remove 'pseudattach' code.

SWBD: (files from /treebank/raw/swbd; note that ordering of commands is important here)

sed -e 's/*[A-Z]**-[0-9]//g'

deletes codes that have * around them, followed by a dash and a number: *ICH*-2 or *T*-1

sed -e 's/._.$//'

deletes a line-final speaker code of the form E_S or N_S

egrep -v Speaker

deletes lines consisting only of Speaker identification

egrep -v '^#'

deletes comment lines (beginning with a #)

egrep -v '^$'

deletes empty lines

tr -d "[:punct:]"

transliterate (here with -d: delete) all punctuation

sed -e 's/ s /s /g'

find and compress things like "John s " to "Johns"

sed -e 's/ nt /nt /g'

find and compress things like "do not" to "dont"

sed -e 's/[0-9]//g'

remove stray numeric codes

(c) Each additional command line includes an extra auxiliary verb. The last search line per corpus indicates the highest number of verb compounds; searches with additional auxiliaries did not come up with any exemplars. The "!< (-NONE-" after the first NP assured that the search pattern did not "double-count" an exemplar. The second "!< (-NONE-" excluded truncated passives. The Brown corpus also required explicit coding of passive-related auxiliary verbs.

(d) The "!WP$" prevents the search from picking up constructions such as "The dog whose owner loved him was happy", of which there were 51 in WSJ and 1 in SWBD.

(e) In all 3 corpora, we include locatives like "the place (that) I work at" or "the situation (that) we are in", as well as prepositionals such as "the teams (that) they are thinking about" and "the people I've talked to". There seem to be fewer of these in Brown than in SWBD.

(f) This string does not include a structure closely related to object relatives - "It was more than I could handle."

(g) The 'S-CLF' marking in WSJ and SWBD is quite general and includes both subject and object clefts, as well as some other constructions. Therefore, the numerical results reflect hand-sorted output of this search string.