COMPUTATIONAL ANALYSIS OF PREDICATIONAL STRUCTURES IN ENGLISH

Henry Kučera
Brown University
Providence, R.I., U.S.A.

Summary

The results of a computational analysis of all predications, finite and non-finite, in a one-million-word corpus of present-day American English (the "Brown Corpus") are presented. The analysis shows the nature of the syntactic differences among the various genres of writing represented in the data base, especially between informative prose and imaginative prose. The results also demonstrate that syntactic complexity, if defined as the number of predications per sentence, is not directly predictable from sentence length.

The purpose of this paper is to present an outline of the procedures and the summary of the results of a computational analysis of the structure of predications in a large and representative sample of English texts. This paper is thus intended both as a contribution to the discussion of computational techniques in linguistics and as a study of linguistic performance.

The data base for this research was a one-million-word corpus of present-day American English, originally assembled by W. N. Francis and Henry Kučera at Brown University in the 1960's and thus commonly referred to by researchers interested in text analysis as the Brown Corpus. More recently, the compilers of the Brown Corpus have completed a grammatical annotation of the data base. The entire one million words of the Corpus have been "tagged", with each word given a specific grammatical symbol. The "tagging" procedure, which was semiautomatic, assigned to each running word an unambiguous symbol based on a taxonomy of 82 grammatical categories. The basic principle of our tagging is an expanded set of grammatical word classes; so, for example, modal verbs are identified by a unique tag, differentiating them from other verbs, as are each of the verbs *be*, *have* and *do*. The second principle of our tagging system is morphological, e.g. plurals of nouns are explicitly coded and thus separately retrievable (as are singulars); the same is true of past tense forms of verbs, verbal participles, and so on. We have also introduced some syntactic factors into our coding; so, for example, coordinate conjunctions are differentiated from subordinate ones; sentence boundaries are marked. Because of the system of grammatical annotations used, the retrieval of various types of syntactic information can now be accomplished algorithmically. Our research, partially presented in this article, is concerned both with automatic parsing of the annotated text and with the study of linguistic performance. Specifically, I shall report on the investigation of sentence length and its relation to sentence complexity in written English.[1]

If we disregard headlines and other headings (of chapters, sections, etc.), the Brown Corpus contains 54,724 sentences, with the mean sentence length of 18.49 words. However, both sentence length and sentence structure vary greatly among the 15 genres of writing represented in the Corpus. In general, sentence length differs significantly between "informative" prose and "imaginative" prose, the former exhibiting a substantially higher mean sentence length. In the Brown Corpus, the term *Informative Prose* is applied to all those samples that have been selected from non-fiction sources. This section is divided into nine genres; for convenience of reference, each genre has been assigned a letter code: A. Press: reportage, B. Press: editorial, C. Press: reviews, D. Religion, E. Skills and hobbies, F. Popular lore, G. Belles lettres (biography, memoirs, etc.), H. Miscellaneous (documents and reports of various kinds), and J. Learned and scientific writings. There are altogether 374 samples of Informative Prose in the Corpus; with each sample being approximately 2,000 words long, this part of the Corpus consists of 755,010 words. *Imaginative Prose*, on the other hand, includes samples taken from a variety of fiction sources and is represented by six genres: K. General fiction, L. Mystery and detective, M. Science fiction, N. Adventure and Western, P. Romance and love story, and R. Humor. There are 126 samples of Imaginative Prose, again of about 2,000 words each, accounting for 256,955 words. The entire Corpus thus consists of 500 samples of texts and contains 1,011,965 running words (word tokens), not counting headlines and other headings.

All genres of the Informative Prose portion have a higher mean sentence length than any of the genres in the Imaginative Prose section of the Corpus. The mean sentence length in Informative Prose ranges from a high of 24.23 words (in H. Miscellaneous) to 18.63 words (in E. Skills and hobbies). In Imaginative Prose, on the other hand, the highest mean is only 17.64 words (in R. Humor) and the low is 12.81 words (in L. Mystery and detective fiction). This difference is, to some extent, due to the percentage of quoted material in the two sections of the Corpus. While no genre of Informative Prose has more than 11.9% of quoted material, with Belles lettres having this highest percentage and the learned samples the lowest of only 2.8%, the percentage of quoted material in Imaginative Prose ranges from a low of 12.76% (Mystery and detective fiction) to a high of 26.8% (Science fiction). Moreover, there is a difference in the nature of the quoted material: in Informative Prose it is a mixture of representations of spoken material and quotations from another written source; in Imaginative Prose, virtually all quoted material is fictional dialogue. Two facts should be noted in this regard, however: first, that no sample with more than 50% of quoted material was included in the Brown Corpus; and second, that the correlation between sentence length and the percentage of dialogue is by no means exact. Several discrepancies in such correlation are given in the essay by Marckworth and Bell who studied sentence length distribution in the Brown Corpus in detail.[2]

Sentence length distribution, of course, is bound to have some effect on syntactic complexity of a text. Clearly, a sentence consisting of two words cannot be considered to be syntactically complex by any conceivable standard of measurement. However, neither in theory nor -- as I shall demonstrate below -- in practice, can sentence length be viewed as a reliable indicator of some common sense notion of syntactic complexity which might be useful either in the study of performance in general or in stylistic syntactic characterizations. Consider, for example, the length in words and the syntactic properties of the following two sentences:

(1) John's grandfather left all his oil paintings to the Metropolitan Museum of Fine Arts

(2) Tom planned to ask Alice to dance

The first sentence has fourteen words (by conventional graphic count), the second exactly half that, i.e. seven words. But while the first sentence has only one verbal form, left, the second has three, one finite, planned, and two infinitives, to ask and to dance. In the fairly conservative versions of transformational grammar of the 1960's (such as the 'standard theory'), the first sentence would have had an underlying phrase marker (deep structure) consisting of one S, and thus not very different from the actual sentence. The second sentence, on the other hand, would have had an underlying phrase marker consisting of three S's, supposed to represent the three underlying predications which could be informally given as 'Tom PAST plan', 'Tom ask Alice', 'Alice dance'. In other linguistic theories, of course, the situation might be quite different, with a much more elaborate initial phrase marker in a generative semantic representation, for example. More recently, on the other hand, syntactic solutions have been proposed in which no sentential source at all is required for infinitival phrases. In this kind of syntactic treatment, the infinitival phrases are then directly generated as VP's.

The purpose of this article is not to discuss or evaluate such conflicting syntactic treatments. Rather, I want to discuss first the algorithm for the retrieval of verbal constructions from the data base, and then summarize the results obtained in the analysis of sentential complexity in the entire Corpus as well as in the different genres represented in the data base.

The data analyzed in this study are the actual sentences of the Corpus, which were encoded in the usual standard English graphic form. There is thus no direct information in the data base about "underlying" structure or even about any syntactic bracketing of the surface string. I will therefore avoid the use of the term "surface structure" entirely in referring to my data. Surface structure, in all those linguistic theories that have utilized this concept, includes at least some labeled bracketing of the terminal string. In the "revised extended standard theory" (REST) of transformational grammar, surface structure actually refers to that level of representation which is not only enriched by the so-called traces, but has yet to pass through the deletion rules, the filter component of the grammar and, of course, the stylistic rules.[3] In our case, however, the only information besides the actual sentences is the accompanying

sequence of grammatical tags, described above.

My basic definition of sentence complexity in the present study will be simply the number of predications per sentence. I shall report these results for each of the 15 genres of the Corpus as well as for the Corpus as a whole. Given the form of the analyzed data, the reader should also be aware that my use of the term "predication" is broader than is usually the case in linguistic literature or in general usage. As is customary, I shall consider a predication to be, first of all, any verb or verbal group with a tensed verb that is subject to concord (for person and number) with its grammatical subject. I will refer to these verbal constructions as finite predications. In addition to that, however, I will also include in my analysis what I shall call non-finite predications. These include infinitival complements, gerunds and participles.

My basic taxonomy of verbal groups is thus quite similar to that adopted by structuralist linguists in the analysis of the English verb. All verbal groups exhibiting concord with a subject, including the subject it (as in it rained) will be counted as finite predications, as will interrogatives; those that do not satisfy these conditions will be considered to be non-finite. My only departure from some structural treatments lies in the inclusion of all imperatives in the class of finite predications. This allows me to place imperatives with and without an overt subject (e.g. Don't worry! and Don't you worry!) in the same class of predications.

When it comes to complex verbal strings involving a quasi-auxiliary plus infinitive (such as going to, supposed to, used to + infinitive), I shall follow here the consistent -- although perhaps somewhat controversial -- approach of Joos.[4] Joos treats all quasi-auxiliaries differently from "true" auxiliaries (such as will or may), pointing out that they exhibit different syntactic properties. Joos also argues that including only some of the quasi-auxiliaries with the class of auxiliary verbs would make the whole English verbal system 'incomprehensible'. My adoption of Joos' approach means that in my analysis a sentence such as He used to play tennis will have two predications, one finite and one non-finite.

The retrieval and analysis of verbal forms, which is the subject of this report, represents only a segment of a larger parsing algorithm for the complete syntactic and stylistic analysis of the Brown Corpus. The retrieval has been made possible by the "tagging" system described above. As already mentioned, verbal constructions from all the sentences of the Corpus have been included in this analysis, with the exception of those occurring in headlines and other headings. Headlines and headings, which are identified by a special symbol in the tagged Corpus, were not included because of the particular nature of English "headline grammar", which often omits verbs entirely, e.g. Actor in Critical Condition after Explosion, or omits some verb form, particularly the finite one, e.g. President to Meet Brezhnev in Vienna. All sentences outside headlines are included, however, even those that do not contain any verb at all (e.g. Just our luck!). The number of sentences with a "zero" predication is small: there are 1869 of them in the entire Corpus, accounting for only 3.4% of the Corpus sentences. Nevertheless, they have been included in computing the statistics.

Verbal constructions of both types, finite and non-finite, may consist of a single verbal form (e.g. likes or to like) or of one or more auxiliaries plus the main verb. The longest possible finite verbal group in English can have five elements, e.g. may (might) have been being considered; the longest active verbal group can have only four elements, e.g. may (might) have been considering. A non-finite verbal group can consist of a maximum of four verbal elements, e.g. to have been being considered. Of these, the maximum finite passive verbal group with five elements does not occur in the Corpus at all, nor does the maximum non-finite group with four verbal elements. However, the maximum finite active group with four elements, i.e. the type may have been considering, occurs 8 times, and the second longest passive group, i.e. the type may have been considered, 68 times. The situation is similar with regard to non-finite groups: the one of maximum possible length, i.e. the type to have been being considered, does not occur at all. In three-element groups, i.e. the type to have been considered or to have been considering, only the first (passive) form occurs, 22 times; there are no occurrences of the active type of this three-element group.

Complex verbal groups may be continuous (i.e. not interrupted by a non-verbal element) or discontinuous, i.e.

so interrupted. Discontinuous verbal constructions exhibit a different pattern in declarative sentences on one hand, and in wh-questions and yes/no questions on the other. In declaratives, the number of word-classes that can interrupt a complex verbal group is relatively small: it consists primarily of adverbs, e.g. He will probably consider ..., He has indeed been asked. In declarative sentences, other word-classes, including all the components of a noun phrase, constitute a definite clue that the verb group has terminated. Clues of this sort are of crucial importance in any grammatical retrieval or parsing that uses annotated but otherwise unbracketed strings as input. One of the important facts that such an algorithm has to consider is that, due to various "deletion" rules under conditions of identity, an English verbal group may appear in a truncated form. Consider, for example, the following sentences:

(3) Teddy could not be elected but his cousin could (be)

(4) Teddy could not have been elected but his cousin could (have (been))

(The forms in parentheses indicate optional deletions.)

It is because of this possible truncation phenomenon that the retrieval algorithm needs to allow for the possibility of a verbal group ending in an auxiliary.

The situation is more complex when it comes to the retrieval of verbal groups in wh-questions and yes/no questions. Because of the auxiliary inversion in such cases, a large number of word-class representatives, including entire noun phrases, can be embedded within a verbal group in such sentences. The retrieval of complex verb groups thus needs to take into account a number of variables. Particular attention needs to be paid in the parsing procedure to the fact that an incomplete verbal group may represent either a truncated string or a discontinuous predication which continues later in the sentence.

The retrieval algorithm for all verbal groups, finite and non-finite, and continuous and discontinuous, scanned the tag sequence in each sentence from left to right, without backtracking. The retrieval was thus essentially accomplished by a finite-state automaton (FSA). The complete FSA that can properly handle both continuous and discontinuous verbal contructions (including truncated ones) is quite complicated. Purely for illustration, I give below a small fragment of the FSA, which will retrieve only those verbal groups that begin with a modal or with 'have', and are continuous.

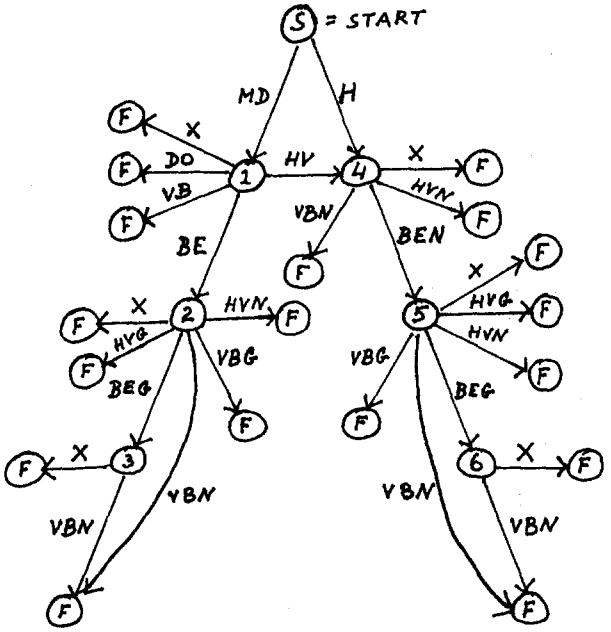

Figure 1

The arcs in the transition diagram in Figure 1 are labeled with the tag symbols of the appropriate classes of items that need to be present for the automaton to reach a final state, and for the string to be thus accepted as a legitimate verbal group. Transition arcs labeled X, all of which terminate in the final state of the automaton, make it possible for truncated groups to be accepted. The symbol X, in this case, thus designates any tag outside of those that may appear in a verbal group. The meaning of the other tag symbols in Figure 1 is as follows: MD = modal; BE = 'be' (base form); BEG = 'being'; BEN = 'been'; H = any form of 'have'; HV = 'have' (base form); HVG = 'having'; HVN = 'had' (past participle); DO = 'do' (base form); VB = main verb (base form); VBG = present participle of main verb; VBN = past participle of main verb.

The basic results, obtained in my analysis, are summarized in Table 1. Three figures are given for each of the fifteen genres and for the Corpus as a whole: mean sentence length in graphic words (i.e. word tokens), mean number of predications per sentence, and the

average number of words of text per predication.

TABLE 1

Genre	Words per Sent.	Pred. per Sent.	Words per Pred.
A. Press, report.	20.81	2.65	7.85
B. Press, edit.	19.73	2.74	7.20
C. Press, reviews	21.11	2.65	7.96
D. Religion	21.23	2.90	7.32
E. Skills	18.63	2.60	7.17
F. Pop. lore	20.29	2.82	7.20
G. Belles lett.	21.37	2.94	7.27
H. Misc.	24.23	2.82	8.59
J. Learned	22.34	2.87	7.78
K. Fiction, gen.	13.92	2.41	5.78
L. Mystery/detect.	12.81	2.29	5.59
M. Science fict.	13.04	2.23	5.85
N. Adv./Western	12.92	2.30	5.62
P. Romance	13.60	2.45	5.55
R. Humor	17.64	2.84	6.21
CORPUS	18.49	2.65	6.97

The three sets of figures, taken jointly, throw a considerable light on the nature of the principal differences among the genres. Particularly revealing is the comparison of the genres of Informative Prose (A through J -- henceforth INFO) as a group with the group encompassing Imaginative Prose (genres K through R -- henceforth IMAG). As already mentioned -- and certainly not unexpectedly -- the mean sentence length, measured in word tokens, is much larger in INFO than in IMAG. The reader should notice especially that all genres of INFO have their sentence-length mean above the Corpus mean, while all genres of IMAG are below the Corpus mean.

The situation is different, in interesting ways, when it comes to predications. Here, too, the number of predications per sentence tends to be greater in INFO than in IMAG, but not consistently so and certainly not to the extent that the differences in sentence length would lead one to expect. No longer are all INFO genres above Corpus mean and all IMAG below it. Within INFO, genre E (Skills and hobbies) is below the Corpus mean, and A (Press, reportage) and C (Press, reviews) are exactly at the mean. On the other hand, in IMAG, genre R (Humor) is well above the Corpus mean.

The lack of correlation between sentence length and the number of predications per sentence, i.e. sentence complexity in my definition, is displayed in a particularly striking manner in the third set of figures, which give the mean number of words per predication. In this case, all genres of INFO show a much larger number of words per predication than the genres of IMAG. As a matter of fact, all genres of INFO are above, and all genres of IMAG below the Corpus mean in this instance. Table 2, which summarizes all the relevant data for the two groups of prose and for the Corpus, shows these results quite clearly.

TABLE 2

Measure	INFO	IMAG	CORPUS
Words/Sent.	21.12	13.55	18.49
Pred./Sent.	2.80	2.38	2.65
Words/Pred.	7.54	5.69	6.97

While Table 2 simply confirms that sentence length is highly genre dependent, it also shows that the predication/sentence figure is not directly correlated with sentence length. The words/predication figures show, in essence, that the number of words needed to express a predication is considerably smaller in those styles of writing in which sentences tend to be shorter. This fact also implies some interesting facts about the overall structure of sentences in INFO as compared to IMAG. Since, aside from the verbal groups, the other major constituents of a sentence are the nominal groups (i.e. NP's), the statistics presented in Table 2 clearly suggest that nominal groups in INFO generally tend to be longer (and, in some sense, thus more complex) than those in IMAG. Both cognitive and automatic parsing of texts of the informational kind will thus put greater demands on noun-phrase processing.

In order to investigate the matter somewhat further and to see what kind of requirements the two groups of prose may impose on the processing of verbal groups, I have also investigated the differences between the ratio of finite vs. non-finite predications in the two groups of writing. The results are given in Table 3, where the symbol F and NF stand for finite and non-finite predications respectively.

TABLE 3

Group	Type	No.	Pred. per sent.	Pct.
INFO	F	68,157	1.91	68.09%
	NF	31,935	0.89	31.91%
		100,092	2.80	100.00%
IMAGE	F	34,329	1.81	75.96%
	NF	10,866	0.57	24.04%
		45,195	2.38	100.00%
CORPUS	F	102,486	1.87	70.54%
	NF	42,801	0.78	29.46%
		145,287	2.65	100.00%

A further examination of the information in Table 3 shows that the greater percentage of non-finite predications in INFO (31.91%) than in IMAG (24.04%) is due largely, although not exclusively, to the greater frequency of gerunds and participles in the INFO texts. There are, on the average, 0.59 gerundival and participal predications per sentence in INFO and only 0.36 in IMAG; the mean for the Corpus is 0.51. This difference is less pronounced with regard to infinitival complements: INFO has a mean of 0.30 infinitives per sentence, IMAG 0.21; the Corpus mean is 0.27 infinitives per sentence.

To summarize then, we can describe the syntactic style of Informative Prose, compared to Imaginative Prose, by at least these three characteristics: longer sentences, more complex nominal structures, and a larger proportion of non-finite predications. In contrast to this, the texts of Imaginative Prose exhibit shorter sentences, a significantly smaller number of word tokens per predication (pointing to less complex nominal groups) and a smaller percentage of non-finite predications. The research which we are now conducting with the Brown Corpus should provide us with further insights into the syntactic structure of English texts and their stylistic properties, as well as into problems of automatic parsing in general.

References

1. The list of tags and the rationale for the grammatical annotation system is given in W.N. Francis and H. Kučera, Manual of Information to Accompany a Standard Corpus of Present-Day American English (Department of Linguistics, Brown University, Providence, 1979). Much of the computer programming required to produce the final form of the tagged Corpus, as well as the retrieval of the predications, was done by Andrew Mackie whose imaginative assistance is gratefully acknowledged.

2. Cf. Mary L. Marckworth and Laura M. Bell, "Sentence-Length Distribution in the Corpus," in Henry Kučera and W. Nelson Francis, Computational Analysis of Present-Day American English (Brown University Press, Providence, 1967).

3. Cf., for example, Noam Chomsky and Howard Lasnik, "Filters and Controls," Linguistic Inquiry, Vol. 8, No. 3 (1977).

4. Martin Joos, The English Verb (The University of Wisconsin Press, Madison 1964).