DOCUMENT RESUME

ED 111 204 FL 007 071

AUTHOR Francis, W. N.
TITLE Problems of Assembling, Describing, and Computerizing Corpora. Research Techniques and Prospects. Papers in Southwest English, No. 1.
INSTITUTION Trinity Univ., San Antonio, Tex.
PUB DATE 75
NOTE 25p.
AVAILABLE FROM Trinity University, San Antonio, Texas 78222 ($2.00)

EDRS PRICE MF-$0.76 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Comparative Analysis; *Computational Linguistics; Contrastive Linguistics; *Data Collection; Descriptive Linguistics; Language Research; Linguistic Competence; Linguistic Performance; *Research Tools; Semantics; *Word Frequency; Word Lists
IDENTIFIERS *Brown Standard Corpus

ABSTRACT

The paper investigates the problems of assembling, describing and computerizing corpora, defined as collections of "texts assumed to be representative of a given language, dialect or other subject of a language, to be used for linguistic analysis." Specific reference is made to the formation of the Brown Standard Corpus. The formation of a corpus is justified in terms of saving effort and in providing a compilation of data that will serve as a research tool in comparative studies. Important questions in the process concern the body of language from which the sample will be drawn, the size of the sample, and its structure. These, in turn, are dependent on the purpose for which the corpus is assembled: graphic analysis will require a different corpus than will phonological or grammatical analysis, for example, the latter presenting the most problems. Practical constraints on the size of the corpus, including time, energy and money are mentioned. The organization of the corpus is discussed, underlining such factors as the size of the base units, mode of selection and collection, assembly of the corpus and computerization. The question of how much additional explanatory material should accompany the corpus is raised, with particular reference to lexical and semantic analyses. (CLK)

PROBLEMS OF ASSEMBLING, DESCRIBING, AND COMPUTERIZING CORPORA

W. N. Francis
Brown University

Random House Dictionary gives as its fourth definition of corpus "Ling. a body of utterances or sentences assumed to be representative of and used for grammatical analysis of a given language or dial ·." Being a Merriam-Webster fan, I must admit with regret that this is a good deal better than Webster's Eighth New Collegiate does by our subject; Webster gives only "all the writings of a particular kind or on a particular subject; esp. the complete works of an author." As I shall demonstrate shortly, a corpus for linguistic purposes is almost never complete; it is virtually always a sample. In fact, I should like to broaden the Random House definition a bit. A corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis. This takes account of the facts that the corpus may be purposely skewed--toward legal or scientific language, for instance--and that it may be used for phonological, graphemic, lexical, or semantic, as well as grammatical analysis. In fact, the most frequent recent use of corpora has been for lexical analysis, particularly the statistical study of word frequencies.

When I was just beginning the job of collecting the Brown Standard Corpus (about which more later), I encountered Professor Robert Lees at some linguistic orgy or other. In response to his polite question as to what I was doing lately, I answered that I had a grant from the Office of Education to assemble a million-word corpus of present-day American English. He looked at me in amazement and asked: "What in the

world do you want to do that for?" I burbled something about finding out the true facts about English grammar. I have never forgotten his reply: "That is a complete waste of your time and the government's money. You are a native speaker of English; in ten minutes you can come up with more illustrations of any point in English grammar than you'll find in many millions of words of random text." Now beyond the fact that Bob Lees is (or at least was) one of the great put-down artists, this remark has important implications for our subject. I don't think that Chomsky had yet coined the terms competence and performance, but that's what Lees (who, you will remember, was Chomsky's first and most articulate disciple) was talking about. A corpus--a collection of texts-- is inevitably a record of performance, while the linguist, according to earlier transformational theory, at least, is supposed to be interested only in competence. And the best way to arrive at a knowledge of competence, it was held, is not by looking at texts but by exploiting the intuitions of the native speaker. So it is quite legitimate to ask why one should expend time, energy, and funds on accumulating a corpus. The answer is, of course, that a good many linguists are interested in performance, whether for its own sake or as a way in to an understanding of competence. As evidence, I can cite the fact that over 130 copies of the computer tape containing the million-word Brown corpus are in use all over the world, from Wellington, New Zealand, to Oslo. I don't know what all those people are doing with it, but I get enough news back to convince me that our original idea--to accumulate a corpus which would be a standard research tool for a variety of linguistic studies--has been amply vindicated. For example, a Swedish scholar has used it to make counts of letter frequencies in printed English

(Zettersten 1969); a philosopher in Hong Kong is studying the collocations of the word good; a lady in Jerusalem is studying word-families; and my own students have used it in studies of English modal auxiliaries (Ehrman 1966), and the progressive aspect of English verbs (Recktenwald 1974). These are matters which can and in some cases must be studied from the performance end. It was our hope that by making available a carefully collected and prepared corpus we would serve two ends: to save others the labor of collecting and computerizing their own, and to supply a standard body of data which would permit comparative studies. It looks as though we were right on both counts.

Not all corpora are as general in purpose as ours. The 5,000,000 word American Heritage Intermediate corpus, for instance (Carroll et al. 1971), was accumulated for the express purpose of making available citations on which to base a dictionary for grades 4-8. The Survey of English Usage, under the direction of Randolph Quirk at University College London, was primarily intended to supply material for grammars, and it has already led to the most comprehensive English grammar since Jespersen (Quirk et al. 1972), as well as several other more limited studies. Professor Cassidy at Wisconsin has gotten together a large and diverse corpus of present-day and older American English, both spoken and written, from which DARE, the Dictionary of American Regional English, is being written (Cassidy 1970). Professor Bessinger has used a corpus of Old English poetry for computer study of Old English versification (for other Old English corpora see Cameron et al. 1970). Most corpora, it is safe to say, have some such specific purpose, though they are often put to use for other purposes than those originally intended.

The purpose for which a corpus is planned should help to supply the answers to some questions which must be answered in the early planning stage. The most important of these are three: (1) what is the body of language, the universe, from which the sample is to be drawn; (2) what is to be the size of the sample; (3) what is to be the structure of the sample. Suppose, for example, that we wish a corpus to be representative simply of "mid-twentieth century American English." Immediately the size of the universe becomes astronomically vast. The other morning, having waked up earlier than I wanted to get up, I amused myself by speculating on just how big it might be. Setting aside written and printed language, the amount of which though great is only a small fraction of the whole, I guessed that the average person might talk enough in one day to make up two continuous hours. Obviously this amount varies greatly with occupation, from professors and politicians at one end to firewardens and lighthouse-keepers at the other, but two hours seemed like a possible average. As you will see, it doesn't much matter if I over-estimated by 100%; the total is still inconceivably vast. Two hours per day at, say, 200 words per minute (about the rate that I am speaking to you now) is 24,000 words per day, or about 8 3/4 million words per person per year. Multiply that by 200 million people and you get a yearly total of 1,752 trillion words. Even if each person spoke only 12 minutes per day, that would only reduce my estimate by a factor of ten, to 175 million million words. It is obvious that any sample conceivably processable would be as a bucket to the ocean, and it doesn't seem to matter much how large or small the bucket. A 5,000,000 word sample, for instance, would give us one part out of 350,000,000.

Most corpora, of course, are drawn from a more restricted universe, appropriate to the particular purpose for which they are assembled. The Brown corpus, for example, is restricted to prose printed in the United States in the calendar year 1961, excluding drama and fiction containing more than 50% dialogue (see Francis 1965). I have no idea how large a body of text this is--certainly many millions of words, so that our sample, while perhaps larger than a bucket from the ocean, is still only a drop in the bucket. The American Heritage universe was considerably smaller--it was limited to textbooks and readers used in grades 3 through 9 of American schools, as determined by a questionnaire sent to over 200 school officials. From some 6,000 titles in the responses, a total of 1,045 were selected as the universe from which about 10,000 500-word samples were chosen. In a few cases the universe is so small that it can be used entire--as in the case of Spevack's Shakespeare concordance, where both universe and sample consisted of all of Shakespeare's authenticated writings.

One factor of theoretical interest which is seldom if ever taken into account in calculating the size of the universe is what might be called the reception index. Usually each text is considered as one, on a par with every other. But if we try to measure not the number of times a text is produced (normally once) but the number of times it is received, the figure for a single spoken utterance might range from one (the informant speaking to the linguist) to 50,000,000 or more (a presidential TV address or Howard Cosell broadcasting a football game). Somewhat the same kind of spread would appear in written material as well. I was once asked by what right we included in the Brown corpus selections

from journals like Archives of Neurology on a par with newspapers like the Chicago Tribune. I could only answer that making an accurate estimate of the readership of every one of our 500 selections in order to weight them would be exceedingly difficult, if not impossible. It can still be done if anyone wants to try it. Certainly the weighting factors would swing the preponderance of evidence way over toward the daily papers, which must constitute an overwhelming majority of the printed material read in the U.S. in any given year.

The size of the sample to be taken from the previously determined universe, which is to constitute the ultimate corpus, is dictated by various considerations, some logistic and practical and some theoretical. Let us consider the theoretical ones first. These are chiefly statistical, having to do principally with the purposes for which the corpus is selected. I am no statistician, so for detailed theoretical treatment of statistical matters I will refer you to two comprehensive articles by John B. Carroll, one based on the Brown Corpus ("On sampling from a lognormal model of word-frequency distribution" in Kučera and Francis 1967:406-24) and the other based on the American Heritage corpus ("Statistical analysis of the corpus" in Carroll et al. 1971:xxi-xl). But I can point out one fairly obvious fact, that the size of the sample needed for accurate representation of the universe depends on the particular aspect of language under consideration. The graphic system, for example, is probably the simplest, with the smallest number of discrete units (letters and punctuation marks) and, in spite of the vagaries of English spelling, the fewest tactical rules of collocation. Hence the sample of printed material needed to establish the frequencies of single characters

and the types and frequencies of permitted combinations is presumably rather small. I don't know how large samples were used by Poe for the cryptography in his famous story "The gold bug" or by the designers of the linotype to establish the familiar sequence ETAOIN SHRDLU which we sometimes see in a newspaper when the operator has intended to cancel a slug by running his finger across the keys. I am sure these samples were smaller than 1,000,000 words, yet their findings are borne out by Zettersten's study already referred to, though he did discover that the frequencies and even the rank among the first ten characters vary with the genre of material sampled. Similarly, the phonological system, again with relatively few discrete units and (pace Chomsky and Halle) fairly simple rules of combination, can be studied on the basis of a quite small sample. The Linguistic Survey of Scotland, for example, used a list of about 1,000 words, which could be pronounced and recorded in a few hours, to establish the phonemic repertories of its informants (Catford 1958). Tactical rules and prosodic data, of course, require larger samples of continuous speech, such as those collected for the Survey of English Usage (Quirk 1960, 1974:167ff). But it is still probable that a practically complete inventory can be made on the basis of a not overwhelmingly unwieldy sample.

When we come to the grammatical system, however, it is a different story. Though theoretically finite, as the transformational grammarians tell us daily, the number of possible constructions is certainly very large, and it is a truism that no grammar, not even the vast compendia of Jespersen, Kruisinga, and Poutsma, captures all the possibilities. And these grammars are based on lifetime collections of

examples--corpora admittedly skewed in the direction of the grammatically interesting. Here I must agree that Bob Lees was right--the native speaker can conjure up constructions that will be accepted by other native speakers as grammatical even though they may be so statistically infrequent as not to appear at all in very large samples of the language. A case in point is the fully marked verb phrase--marked for modality, past tense, perfective phase, progressive aspect, and passive voice. One that I once caught and remembered is you should have been being paid, which has excellent credentials--it occurred in a conversation between the Executive Secretary of the Modern Language Association and the President of the National Council of Teachers of English. Yet no such verb phrase appears in all the million words of the Brown Corpus--I know because I looked (or at least got the computer to look).

This consideration--the difficulty, if not impossibility, of finding all possible constructions in a limited sample--led Professor Quirk and his colleagues to the notion of the extended corpus--a random sample which is extended by directed elicitation. As he describes it (Quirk 1974:167-8):

> Central to the entire operation (the Survey of English Usage) is the notion of assembling for analysis a large corpus of English as it is actually used, a corpus that is reasonably representative of the repertoire of educated professional men and women in their activities, public and private, at work and at leisure, writing and speaking. By basing our description on such a range of actual usage we have sought to avoid both of the deficiencies in older descriptions that I have mentioned (i.e. eclectic, and hence skewed, sampling and omission of spoken language). But, of course, no corpus, however large, could be expected to give information in the requisite degree of detail on all the grammatical structures of English. Indeed, it is difficult to conceive of a body

> of texts large enough to inform us fully even on inflexional variations. . . . . We have, therefore, additionally developed psycholinguistic techniques to investigate the native English speaker's potential performance, to catalogue items in his linguistic repertoire that do not necessarily emerge in the actual instances of usage that an observer may assemble.

I do not have time here to describe the "psycholinguistic techniques" to which Quirk refers, but you will find a full discussion in Investigating linguistic acceptability by Randolph Quirk and Jan Svartvik (1966). Their purpose is to extend the corpus in a selected direction by subtle and indirect methods of elicitation. In a sense they may be considered an attempt to get at competence more directly than is possible by the description and analysis of a large sample of performance.

When the purpose of the corpus is lexical, all thought of complete coverage must be abandoned. So large is the lexicon of a language and so almost infinitely numerous the possibilities of collocation that we cannot imagine a corpus, however large, that can contain it all. Of course, if the universe is sufficiently restricted, completeness is possible. The Spevack concordance, whose universe is the works of Shakespeare, lists all of the 29,066 words that Shakespeare uses and all their collocations, at least out to the limits of the line of verse. But that is a very special case. If our universe is language in general, or printed language, as in the Brown and American Heritage corpora, the numbers are simply too large. The reason is the large size of the vocabulary of any known language and the very uneven way that the words are distributed. With a tightly closed system such as the graphic one, it is likely that one would encounter all of the characters within a few

pages of print and no more would be encountered as we made the sample larger. The lexicon, on the other hand, is virtually open-ended. No matter how far we extend the sample, we keep encountering hitherto unrepresented words. In slightly over a million running words (tokens) in the Brown corpus, there are about 50,000 different words (types) for a type-token ratio of about 1:20. In the 5,000,000-token American Heritage corpus there are about 87,000 types, or a type-token ratio of about 1:57 (remember that the universe from which this corpus was drawn was limited to books for children below the ninth grade). It is clear that the curve is leveling off, since increasing the number of tokens fivefold increased the number of types by only 75%. It seems reasonable to suppose that at some point the whole vocabulary would be represented. By rather elaborate mathematical methods, Professor Carroll has calculated that a corpus with the type and token characteristics of the American Heritage one must be drawn from a universe with over 600,000 types. In order to include them all, he estimates that a corpus 100 times as large, or 500,000,000 words, would be necessary. (Note: This is, of course, theoretical. Since the universe from which this corpus was taken is limited to about 1,000 books, its total size is much less than half a billion words. Thus as the sample was increased it would eventually come to include the whole universe, and we would have a situation similar to the Shakespeare concordance, with a restricted vocabulary exhaustively listed.)

The reason for this, of course, is the very uneven distribution of words in natural language. In corpora of the size of the Brown and American Heritage ones, frequencies range from the, which accounts for about 7% of all the tokens--every fourteenth word--to the hapax legomena,

the words which occur only once, which make up around 40% of the types but only a tiny proportion of the tokens. In the American Heritage corpus, for example, the first 1,000 types, or only a bit over 1% of the vocabulary, account for 75% of the total corpus, while the approximately 35,000 hapax legomena, or about 40% of the vocabulary, account for only .7% of the total corpus. The figures for the Brown corpus are comparable. It seems to be true that for large corpora the number of hapax legomena runs around 40%. To most people this seems counter-intuitive. In fact you can probably make a bit of money if you will wave a novel at somebody and bet he can't estimate within 10% the proportion of words in it that occur only once. But you'd better make the wager fairly high to compensate you for the time it will take to make the count.

In addition to these theoretical considerations, it is obvious that there are some practical constraints on the size of corpora. The time and energy that can be devoted to collecting and processing a corpus are limited, and both of these can ultimately be translated into money. Corpora are accumulated in various ways by different agencies: by individuals, by groups centered at academic institutions, like Quirk's staff at University College, by governmental agencies (increasingly such documents as statutes and patents are being computerized for ease of retrieval), and by various commercial enterprises, especially the publishers of dictionaries. In all cases, it is largely the availability of funds which imposes the outer limit on size.

The cost, of course, varies greatly depending on the type of material and the amount of pre-editing and other processing it needs. The easiest and cheapest are those that are already available, having been

prepared for some other purpose. The increasing use of computerized composition, for example, means that large bodies of text--Time magazine, for example--are first put onto punched paper tape, which is immediately available for input. Lawrence Urdang, editor of the Random House Dictionary, has a 25,000,000-word corpus of this kind of material. I believe, however, that it is not readily available because of copyright restrictions. Another problem is that the compiler has no control over the universe or the sampling methods; he must take what he can get. Nevertheless, I look for considerable use to be made of this kind of material in the future.

At the other extreme is spoken material, which is usually collected by tape recorder and transcribed, sometimes directly onto punch cards or paper tape, more often through an intermediary typescript, which may or may not be readable by an optical scanner. Transcribing may take anywhere from ten to forty times as much time as written material, depending on the amount of phonetic detail to be preserved. Occasionally, though rarely, a spoken corpus may be transcribed at someone else's expense for another purpose. I have not seen any proposals to use the White House tape transcripts in this way, though it might be done. To judge from the newspapers, however, the quality of both recording and transcribing was so poor as to render their value questionable for linguistic purposes, however powerful they might be legally and politically.

Printed text falls somewhere in between, so far as cost goes. The million words of the Brown corpus cost the U.S. Office of Education about $23,000 in 1963-64, or about 2.3¢ per word. I suppose that it would be twice that today. So if we multiply by a factor of 20 for spoken

material, it might approach a dollar per word, which is getting pretty high. Of course the cost might be less if volunteer or cheap labor (such as graduate students) can be used, which is the case with both the Survey of English Usage and Geoffrey Leech's British analogue to the Brown corpus, now nearing completion at Lancaster. But it's still an expensive business. The most expensive of all, I suppose, is the dialect survey, which requires putting fieldworkers into the field to seek out, interview, and record informants in their own homes or places of business. It is indeed remarkable that corpora such as the field records of the Linguistic Atlas of the U.S. and Canada or the Survey of English Dialects get collected in the first place. None of them have yet reached the computer, though plans are under way for some of them to do so.

Once the basic theoretical and practical questions as to purpose, size of the universe, and desirable and feasible size of the sample have been answered, decisions must be made as to the nature of the corpus. What, for example, is to be the size of the basic units? It might be individual words or short phrases, as in most dialect surveys. Or sentences, as in the Wenker-Wrede Deutscher Sprachatlas. Or single lines of verse, as in Spevack and most other concordances. Corpora to be used for syntactic analysis usually consist either of complete texts or, as in the Brown and American Heritage corpora, of samples of fixed length (2,000 and 500 words respectively). In some cases a corpus may be composite: the Survey of English Usage, for example, contains four main types of spoken material and three types of written material, all of them measured off into texts of 5,000 words each.

Next, the mode of selection and collection must be established. Usually this is a combination of designation and random choice: that is, it is decided in advance how the universe is to be subdivided and how large a sample or number of samples is to be taken from each subdivision, and then choice of the actual texts or portions of texts to be used is made by some random procedure. In the case of the Brown corpus, we convened a conference of such corpus-wise scholars as Randolph Quirk, Philip Gove, editor of the Webster's Third International dictionary (by the way, I have it in writing from him that he prefers the plural corpuses, though he put corpora into the big book), and John B. Carroll. This group decided the size (1,000,000 words), the number of texts (500 of 2,000 words each), the universe (material in English, by American writers, first printed in the United States in the calendar year 1961), the subdivisions (15 genres, nine of "informative prose" and six of "imaginative prose"), and by a fascinating process of individual vote and average consensus, how many samples from each genre (ranging from six in science fiction to 80 in learned and scientific). Having made these pontifical decisions, they went their way, leaving me and my small staff to manipulate book lists, library catalogs, and a random number table to select the actual samples. Logistic factors led us to make some arbitrary decisions, such as restricting the selections from the daily press to those newspapers--30 or so--of which the New York Public Library keeps microfilm files. The American Heritage corpus made use of a similar combination of plan, practical necessity, and randomness. As I said above, they wrote to 221 school officials, chosen carefully with regard to geographic distribution, type of school (some Roman Catholic schools and some independent schools

were included), and other relevant considerations. The 90 completed questionnaires which they used supplied about 19,000 titles, which came to somewhat over 6,000 different books, divided among 22 pre-established categories. From these they selected slightly more than 1,000, as being "the largest number of texts that could be acquired in the time available and kept under sufficient control through sampling." (Carroll et al. 1971:xv). By methods too complicated for me to describe here but fully explained in The American Heritage Word Frequency Book, they decided how many samples to take from each text, and finally they took the samples at evenly spaced intervals through the texts.

In contrast to these larger corpora with their attempts to be more or less proportionally representative of their universes, the traditional dialect survey is strongly skewed. It usually samples a very small part of the language and a minute proportion of the population. It is strongly biased toward the individually and regionally variable. The dialect fieldworker is not like a geographer surveying the topography of a landscape but more like a prospector looking for gold. He looks for what he wants in the places he believes he will be most likely to find it. One result is that nobody knows what proportion of a regional vocabulary is unique to the region.

Once all these preliminary decisions have been made, the actual assembly of the corpus can begin. If the planning has been done properly, the mode of collection and assembly is pretty well indicated and the actual extraction of the samples becomes routine. For corpora of printed materials such as the Brown and American Heritage corpora it was simply a matter of finding the texts, using the appropriate random

procedures to locate the sample, and counting out the words. In our case we let logistic considerations override complete randomness, in the interest of saving money. Thus if our random selection procedure led us to a title which was not in the Brown University Library, we simply moved down the list to the next title that was--hence our universe was not actually all the prose printed in the United States in 1961, but the selection of that body of prose contained in our library. For a few of the genres, comprising material not usually found in university libraries, we used other collections: as I mentioned above, our newspaper selections were drawn from the microfilm files of the New York Public Library. Our detective fiction came from the extensive holdings of a well known private library in Providence, and our popular magazine materials from the large stock of a New York second-hand magazine store. For other types of material, other methods are used. In connection with a study of Mississippi Black English, both spoken and written, as used by freshmen at Tougaloo College, we made tape recordings of small-group discussions for the spoken corpus and collected class themes from the same students for the written corpus. For his work on lexis, Professor Sinclair collected about a million words of free conversation simply by sitting three or four people around the fire in his Edinburgh office, plying them with tea, and keeping the tape recorder going (Sinclair et al. 1970:22-23). In contrast to these informal but staged sources, Professor Quirk made considerable use of radio material from the BBC, including panel discussions as well as formal and informal talks. Professor Spevack only had to decide what edition of Shakespeare to use. Since he was concerned more with literary than linguistic users, he picked a modern one; some of us wish he had used

the First Folio. Walter Loban, in collecting a large corpus of the language of school children at intervals through the first twelve years of schooling, made use of individual interviews centering on pictures which the informants described or used as starting points for stories (see Loban 1966). The usual method for dialect surveys is to send a trained field-worker into the field with questionnaire, notebook, and tape recorder. But other methods, such as the postal questionnaire, have been used successfully in collecting, particularly lexical data. For his study of the regional vocabulary of Texas, Bagby Atwood equipped vacationing students with check lists which they completed back home; in this way he accumulated a large corpus of lexical material from all the counties of Texas (he was, by the way, the first, so far as I know, to sort his material with IBM equipment) (Atwood 1962). One of the most ingenious devices for quickly accumulating a very specialized corpus was used by Professor William Labov. Interested in the distribution of post-vocalic r, both word-final and preconsonantal, in New York City, he or one of his assistants went to a large department store, consulted the directory to find an item sold on the fourth floor, and then asked all the salespersons and floorwalkers he could find where that item was handled. In this way he got several hundred examples of the words fourth floor in a few hours.

I come now to the last part of my topic, computerizing. I shall not go into much technical detail on this, first because I don't know a great deal about it, and secondly because we have a resource person here who does. But I should like to recommend emphatically that anyone contemplating putting a corpus on the computer (and in these days any corpus of any size at all should be planned with that in view) should find

out enough about the technicalities of programming and computer operation so that he knows what he can expect, what he can reasonably ask for, and what he should do to make the manipulation of his corpus as easy and inexpensive as possible. Probably the easiest way to do this is to take an introductory course in programming, such as is given at most colleges and universities these days. This will allow him to take part in intelligent decisions on some important questions. What, for example, is to be the method of input--punch cards, paper tape, direct typing on a terminal, or optical scanner? How shall it be stored? How indexed? The basic question, the answer to which will guide the programmer to the most efficient and economical format, is, what is to be retrieved? For a special-purpose corpus this question is not difficult to answer. But for a general-purpose one, like ours at Brown, one must try to envisage the uses to which it will be put. We made one mistake in planning the coding for our corpus which has caused a good deal of trouble ever since. Following our model, the Patent Office coding system, we put marks of punctuation immediately after the last letter of the word they followed, instead of leaving a blank space. This meant that to the computer any word followed, say, by a comma was different from that same word without a mark of punctuation or followed by a period. This created problems in sorting and counting words for frequency tables. Eventually we produced a second form of our corpus with all external punctuation stripped off.

A basic question that must be answered at this stage concerns how much of the information contained in the texts, whether spoken or written, which have been sampled is to be preserved in the computer coding. The transfer of linguistic material from one medium to another inevitably

occasions some loss of detail. I know of no way, for example, by which the individual voice qualities that allow us to identify an individual speaker even over the telephone can be indicated in a written medium. Yet in such corpora as the White House transcripts, for instance, this information might be vital. Similarly, the vast variety in printed material cannot all be carried over onto the computer, at least not without making the coding incredibly cumbersome. Yet the size of type, the particular fonts used, the density on the page, even, I suppose, the feel of the paper, all influence us at least subconsciously as we read. What is to be done about more or less obvious mistakes--misspellings and typos in print, or the repetitions, backtrackings, hesitations, Spoonerisms, and all the other performance characteristics par excellence which mark unscripted informal speech? Again the purposes for which the corpus will be used should suggest the answer. Spoken material which is to be used for lexical or grammatical purposes need not be given full phonetic transcription, but can be transcribed in standard orthography: we did this with the Tougaloo corpus I mentioned earlier, and so did Sinclair with his Edinburgh conversations and Loban with his Oakland interviews. But punctuation is another matter. Loban's tapes were transcribed by a stenographer who followed standard punctuation practice, making her own decisions about sentence boundaries. Some of the material contained a lot of what the handbooks call "and . . . and sentences"--long strings of independent clauses linked by and. But when some linguists listened to the tapes, it became apparent from the intonation patterns that most of the and's were sentence-initial. In fact, for some speakers and was the normal way to begin a new sentence in a connected narrative--somewhat

similar to the use of kai in New Testament Greek. Since one of the parameters Loban was using in analyzing his material was sentence length, it is obvious that this one apparently minor point of punctuation made a tremendous difference. To get around this problem, the spoken material in Quirk's Survey of English Usage is transcribed almost entirely in standard orthography, phonetics being used only for special details, but with an elaborate system of notation for prosodic features such as tempo, loudness, pitch, voice quality, and so forth (see Crystal and Quirk 1964). This preserves the grammatically significant speech signals while not encumbering the transcriptions with phonetic detail irrelevant to the purpose of the survey. On the other hand, most dialect corpora are concerned with phonological, as well as grammatical and lexical detail, which means that phonetic transcriptions must be coded for computer input and retrieval. This has been a knotty problem for some time. I know of two ongoing attempts to solve it, one at Leeds for the Survey of English Dialects (Keil n.d.) and the other by Rex Wilson and his colleagues in Canada. They have devised a new phonetic alphabet, designed with a view to a special IBM-type "golf ball" which will input phonetic transcription directly through a terminal (Wilson et al. 1973).

I could go on at some length discussing problems connected with the transfer of the material from speech or print to the computer, but I won't try your patience further. I should like to close by referring to one more question that faces the corpus-collector at the computerization stage. That is, how much additional information, if any, is to be supplied? This particularly concerns material which is to be used for grammatical and lexical, particularly semantic, analysis. Written English, for

example, presents two problems which affect frequency counts especially: the separation of homographs, and lemmatization. English, as you know, is particularly rich in (or plagued with) homographs of two kinds. One kind consists of the cognate words which are different parts of speech but without overt morphological markers of the difference: notably noun-verb pairs like walk, run, love, hate, find, blow, drive, etc., etc. The other kind includes the accidental homographs--unrelated words which happen to come out with the same spelling, such as bloom ("blossom" or "ingot of wrought iron"), sow, row, etc. Even five of the eight modal auxiliaries--can, will, must, dare, and need--have homographic mates, and if we disregard capitalizing we can add a sixth, may. This makes it impossible to find out the absolute and relative frequencies of these important grammatical elements from corpora like Brown and American Heritage, which do not distinguish homographs, without devising rather elaborate algorithms or making the counts manually from concordances. On the other hand, English has some inflectional morphemes--relatively few compared with Latin or Greek, but rather high frequency ones--which separate the graphic forms of plurals, past tenses, and the like from the base form. In a lemmatized list (such as the usual dictionary) all forms of a given word are brought together under a single entry. This can be done fairly easily by the computer, but it is not worth much unless homographs are separated. Thus to know the actual frequency of the verb love, for example, the computer must count all occurrences of LOVE, LOVES, LOVING, LOVED (no problem) and then eliminate instances of LOVE and LOVES that are nouns and of LOVING and LOVED that are nouns or adjectives (a real problem).

Polysemy presents a problem related to homography. Everybody (but not the computer) knows that FAST is at least five different words--a verb meaning "to starve oneself," a noun meaning "a period of self-starvation," two adjectives meaning respectively "swift" and "firm," and an adverb meaning "swiftly." We are all also pretty well agreed that BOARD is one noun, even though its meanings range from "plank" to "governing body of an institution." But how many different words is STALL? The Random House Dictionary has two entries with that spelling; the Eighth New Collegiate has five!

This last question, of course, goes beyond a mere computer problem. But the homograph and lemmatization problems suggest that a great deal more could be done easily with a corpus if it were tagged in some way, by grammatical abbreviations following each word, by subscript numerals, or whatever--some kind of overt mark which the computer could use to put together what should be together and to separate what should be separated, functions which the reader does on the basis of the whole structure and context of the utterance. Such tagging can be done at least semi-automatically, and we are working on it at Brown (Greene and Rubin 1971). But in some cases it might better be done manually as part of a pre-editing process.

I was given the word Problems as part of my title, so if I seem to have been emphasizing the problems and difficulties of corpus-collecting, it is only that I have been carrying out my assignment. But I hope I have not totally discouraged you from embarking on the enterprise. Because, in spite of Bob Lees' derogatory remark with which I began, it is a worthwhile enterprise. And he who not only compiles a corpus for his

own use but makes it available to his fellow researchers performs a public service of immeasurable value. I am happy to be included in the fellowship of such harmless drudges. But you won't catch me doing it again!

REFERENCES

Atwood, E. Bagby. 1962. The regional vocabulary of Texas. Austin: University of Texas.

Cameron, Angus, Roberta Frank, and John Leyerle. 1970. Computers and Old English concordances. Toronto: University of Toronto.

Carroll, John B. 1967. "On sampling from a lognormal model of word-frequency distribution." In Kučera and Francis 1967:406-24.

_____. 1971. "Statistical analysis of the corpus." In Carroll, Davies, and Richman 1971:xxi-xl.

Carroll, John B., Peter Davies, and Barry Richman. 1971. The American Heritage word frequency book. New York: American Heritage Publishing Co.; Boston: Houghton Mifflin Company.

Cassidy, F. G. 1970. "The DARE project at the end of 1970." ERIC Document ED 060 006.

Catford, J. C. 1958. "Vowel systems of Scots dialects." TPS (1958): 107-17.

Crystal, David, and Randolph Quirk. 1964. Systems of prosodic and paralinguistic features in English. The Hague: Mouton.

Ehrman, Madeline. 1966. The meanings of the modals in present-day American English. The Hague: Mouton.

Francis, W. Nelson. 1964. A standard sample of present-day English for use with digital computers. Report to U.S. Office of Education on Cooperative Research Project No. E-007. Providence: Brown University.

_____. 1965. "A standard corpus of edited present-day American English for computer use." College English 26:267-73.

Greene, Barbara B., and Gerald M. Rubin. 1971. Automatic grammatical tagging of English. Providence: Department of Linguistics, Brown University.

Keil, Gerald. n.d. "Narrow transcription on the computer: Taking the phone off the hook." Unpublished ms.

Kučera, Henry, and W. Nelson Francis. 1967. Computational analysis of present-day American English. Providence: Brown University.

Loban, Walter. 1966. Problems in oral English. NCTE Research Report No. 5. Urbana: National Council of Teachers of English.

Quirk, Randolph. 1960. "Towards a description of English usage." TPS (1960):40-61.

_____. 1974. The linguist and the English language. London: Edward Arnold.

Quirk, Randolph, and Jan Svartvik. 1966. Investigating linguistic acceptability. The Hague: Mouton.

Quirk, Randolph, S. Greenbaum, G. N. Leech, and J. Svartvik. 1972. A grammar of contemporary English. London: Longman.

Recktenwald, Robert P. 1974. The English progressive: Semantics and history. Unpublished Brown University dissertation.

Sinclair, J. McH., S. Jones, and R. Daley. 1970. English lexical studies. Report to OSTI on Project C/LP/08. Birmingham: Department of English, University of Birmingham.

Spevack, Marvin. 1968-70. Complete and systematic concordance to the works of Shakespeare. Hildesheim: G. Olms.

_____. 1972. "Shakespeare's English: The core vocabulary." RNL 3.ii:106-22.

Wilson, H. Rex, M. G. Wanamaker, and A. M. Kinloch. 1973. "A revised phonetic alphabet, proposed by the Maritimes Dialect Survey." IPAJ 3.1:29-35.

Zettersten, Arne. 1969. A statistical study of the graphic system of present-day American English. Lund: Studentlitteratur.