

I N F O R M A L S P E E C H

ALPHABETIC AND PHONEMIC TEXTS WITH STATISTICAL ANALYSES AND TABLES

Edward C. Carterette and Margaret Hubbard Jones

University of California Press
Berkeley
1974

Reviewed by John B. Carroll
The L. L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

Is "spoken language" different from "written language", and if so, how? This was the focal question addressed in this monograph. It was the authors' belief that "all previous statistical studies of language ... have derived their material from written language." They hoped to show that "genuine spoken language is actually quite different from written language, even on such a gross level as proportional frequencies of letters and phonemes" (p.3). "Genuine spoken language," in their view, is the sort of language that would be used in free conversations among peers. They proceeded, therefore, to collect large samples of such language.

Being interested also in age trends, they sampled conversational speech from first-, third-, and fifth-grade children, and from "adults" (actually, junior college students from elementary psychology classes, mean age not specified). All were selected as being ostensibly middle-class native speakers of a "Southern California dialect." The conversations recorded might well be described

as "bull sessions" among three participants that, in the case of the children, were put into motion by a friendly adult who faded into the background after the conversation warmed up, and in the case of the "adults," took place in what was purported to be an experiment in small group process. Conversations were tape-recorded and then transcribed both in regular orthography and in phonemes. By the authors' reckoning, the resulting corpus, from 58 peer-group sessions involving 174 different individuals, contained 84,164 lexical words, 313,694 alphabetic letters (in the conventionally spelled form) and 251,360 phonemes (in the phonemically transcribed form). If word and sentence marks (spaces and periods) are included, the figures are 405,906 alphabetic characters and 282,240 phonemes. The total corpus thus gathered occupies pages 57 through 439 of the book; the conventionally printed form and the phonemically transcribed form are on facing pages. No information is given as to whether the corpus is now available in machine-readable form, although it must have existed in that form at some stage of its analysis.

Various types of printed material were also analyzed, to represent "written language": school readers at several grade levels, trade books rated as liked by children, and adult material from a previous study by Newman and Waugh (Information and Control, 1960, 3, 141-153).

The principal mode of analysis--in fact essentially the sole mode of analysis--was inspired by information theory, and concentrates on the phonemes and graphemes of the corpus. Extensive tables (pp. 441-646) give data on first-order frequencies and probabilities

of letters and phonemes, and some of the higher-order sequential frequencies and probabilities (up to "triphones" for phonemes, and word-initial and word-final "tetragrams" for graphemes). The discussion of these statistical tables occupies pages 23-45.

The book itself is a handsome production, printed in a tasteful style on high-quality stock and nicely bound in cloth. On closer examination of its contents, however, one is tempted to conclude that the authors, having completed their manuscript and handed it over to the printer in about 1968, proceeded to divest themselves of any further responsibility for its editing and publication. Only in this way can one explain the many egregious typographical errors, inexcusable editorial changes, and glaring omissions of important materials that are to be noted in the book. The defects are in many instances serious enough to destroy much of the book's potential usefulness.

I infer that the manuscript was completed in 1968 or thereabouts because a reference to a 1968 article is cited as "in press," and there are no references to any publications subsequent to that year. Along with illiterate spellings and typographical errors such as pulication, concensus, idiosyncracies, and diagrams (for digrams, p. vii), we find inconsistent mathematical notation (pp. 18-20) and incorrect plotting of data points (in Figure 3.3.1 the points do not everywhere increase monotonically, although they must, in theory, and do, by the values recorded in Table 7.6, p. 451). But these matters are trivial beside the blatant alterations in some of the tables. On page 43 the authors state that in trigram tables the symbol (/) stands for a word space, in the tables them-

selves (pp. 479-535) no such symbols are to be found. Apparently the original manuscript contained the symbols, but the printer converted all of them to blank spaces and left-justified the entries. Thus, there is no indication as to whether the most frequent trigram, printed as "th", is to be taken as /TH or TH/ In this particular case, it is almost certainly to be taken as /TH (from word-initial 'TH- in frequent determiners and some content words), but what about a less frequent trigram such as what is printed as "en": is this /EN or EN/ ? Fortunately a similar error does not occur in the tables of "triphones" (pp. 537-613) where the printer indicated the space character as a carat (^), although the authors intended use of the slash (/).


The gross omissions are of certain summary statistical tables that were, according to the authors' text discussion (pp. 43-45), supposed to accompany the tables of triphones and trigrams. Tables 8.7.1-8.7.4 are not, as promised, preceded by tables giving frequency distributions of trigrams; similarly, the tables in the 8.9 series are not accompanied by the promised frequency distributions. I gather also that these omitted tables contained reports of certain information-theoretic statistics such as H or H'. With much effort, a user of this book could construct the frequency distributions from the detailed tabulations, but it is still inexcusable for the printer to have omitted them.

But the printer (or the publisher) is not to be blamed for everything. There are also problems with the manuscript, and the research that lay behind it. Questions must be raised about the purposes of the work, the procedures, and the methods of analysis.

Purpose of the work. At the time the work was undertaken, probably in the early 1960's, it may have been correct to say, as the authors do, that all previous statistical analyses of language had been based on written or printed material. Nevertheless, one can think of exceptions: even the authors cite the study of telephone speech by French, Carter and Koenig (Bell System Technical Journal, 1930, 9, 290-324), although this study has its limitations. Concurrently with the work reported in this book, a number of statistical studies of spoken language appeared (e.g., D. Howes, A word count of spoken English, Journal of Verbal Learning and Verbal Behavior, 1966, 5, 572-606; F. Goldman-Eisler, Psycholinguistics: Experiments in Spontaneous Speech, New York, Academic Press, 1968), and several investigators collected samples of spoken language for various purposes (e.g., W. F. Soskin & V. P. John, The study of spontaneous speech, in R.G. Barker (Ed.), The stream of behavior, New York, Appleton-Century-Crofts, 1963). These and other studies could have been cited by Carterette and Jones in their list of references; they were not.

What is relatively unique about Carterette and Jones' samples is that they were collected and analyzed by a uniform, specified procedure, with considerable attention to insuring that the speakers were representative of some defined population at several age levels. The extensive comparative analysis of these texts in both their spoken (phonemic) and written (graphemic) form is also unique. Although one may disagree with the modes of analysis that Carterette and Jones chose to use, we have them to thank for making their corpus available for any other types of analysis that might be desir-

able or feasible.

Are the authors correct, however, in saying they are studying differences between "spoken language: and "written language"? Evidently the authors mean to draw the distinction not between "written" language and "spoken" language as such (for any sample of language can be either written down or spoken aloud), but between "informal language" and "formal language," i.e. edited language. That is, they are concerned with how language is generated. They apparently feel that the most "natural, genuine" language is generated in informal conversational situations where the speakers have little if any pressure to make their speech conform to artificial norms of correctness, grammaticality, or even communicative efficiency. It is probably for this reason that they titled the book Informal Speech. They were also interested in ~~the~~ development of speech generation, at least over a certain age range  from the first grade to the junior college level. Note, however, ~~that~~ they did not sample the informal (relatively unedited) speech of highly educated, mature speakers such as might be found in a congressional cloakroom or the salons of an ivy-league faculty club. It is quite possible that speech sampled in such circumstances would conform fairly closely to the norms of edited, written language, at least in some respects (and probably in the respects studied by the authors).

If one bears in mind, therefore, the limitation that it is not speech as such (as opposed to writing) that these authors have studied, but rather unedited speech of relatively immature speakers,

the work has considerable unique value by virtue of its presentation of extensive samples of such speech. The authors do not really analyze, however, the ways in which unedited language differs from edited language, nor the ways in which speakers develop strategies of editing their speech.

Another objective of the authors was to use "the rather powerful tools of information theory in the description of informal speech over the age range, in an effort to trace the role of redundancy in shaping language as a person uses it and presumably understands it in discourse." I will have more to say about the authors' use of information theory below.

Data collection procedures. For their purposes, the procedures were excellent--certainly superior to procedures (interviews, contrived play situations, classroom discourse) used by other investigators, for the procedures certainly elicited informal, unedited speech full of interrupted sentences, hesitations, false starts, etc. The content covered a wide range of topics. Nevertheless, it was all conversation; the participants were merely exchanging ideas, and declarative and interrogative utterances abound. They were not directing each others' physical activity; thus, there appears to be a low incidence of imperatives, requests, etc. The corpus is certainly large enough for the kinds of analysis employed by the authors at a phonemic or graphemic level, but it might not be sufficiently large or representative for certain types of lexical or syntactical analysis.

Transcription procedures. Transcription of the corpus was a formidable and time-consuming project, not only in terms of a con-

ventional printed form but also and particularly so, in terms of a phonemic version. One can only say that the authors made approximately the best of a very difficult job. They found it impossible always to identify speakers, and decided to omit any speaker identifications, showing only changes of speakers. For the phonemic transcription, they used a modification of the Trager-Smith transcription suggested by Peter Ladefoged, but found it hard to get hired "phoneticians" to adhere to the system consistently. The system used was admittedly only partially phonemic; for example, the glottal stops that were recorded may or may not be phonemic. One wonders how consistently such distinctions as those between /aydownow/, /aydənnow/, and /aydownnow/ were observed. The treatment of pause phenomena was particularly bothersome. Pause phenomena were represented in the phonemic transcriptions only by spaces and periods; thus, the phonemic transcriptions contain a high proportion of very long "phonemic words" like /naktowvərəmɪlkbədəlwənnɛywərgowɪnɛr./, transcribed in graphemic form as "knocked over a milk bottle when they were going in there" (pp. 312-313). But in the printed version the location of the junctures between these "phonemic words" is unfortunately not shown, although it would have been fairly simple to have done so, perhaps by the use of slashes (/). For certain purposes, it is unfortunate, also, that certain kinds of material were omitted from the transcriptions, e.g. repetitions of words when in answer to wh- questions, and certain kinds of interruptions in continued sentences. It is curious that proper names were generally deleted in the printed version but left in the phonemic

transcription (e.g. compare pages 432-433); the authors apparently felt that privacy would be preserved in the mystique of a phonemic transcription but not in conventional spelling.

One can only guess as to what stress and intonational phenomena occurred in the conversations. The printed version contains no question or exclamation marks, and the phonemic transcription contains no indication of intonations. Whether the tape-recorded material is archived somewhere, available for further analysis, is not stated.

Analysis. As mentioned previously, all analyses of the material were at a phonemic or graphemic level. The intention was to use the "powerful tools of information theory" to trace the development of "redundancy." This mission was certainly enough to occupy the authors throughout the course of their study; it was apparently their intention to leave other types of analysis to future workers. One may raise the question, however, whether information theory analysis was really so "powerful", and indeed how informative it was when applied solely to zero- and higher-order phenomena such as word distributions, distributions of syntactical patterns, etc.? Perhaps I can illustrate my attitude by relating my own experiences with such analyses on the phonemic level. In 1951, in connection with an interdisciplinary seminar of psychologists and linguists, a group of the participants decided to investigate the information-theory characteristics or sequences of phonemes in American English speech. Not having readily available any authentic samples of speech, we decided to

make a phonemic transcription of a series of one-act plays written for high school student performances. One of the linguists, (Fred Agard) transcribed some 20,000 phonemes from this corpus and subsequently I did a number of information-theory analyses of the data. The results were incorporated in a mimeographed report that, incidently, was cited by Carterette and Jones in their reference list as, however, "not seen." (A copy of the report could easily have been made available to them if they had asked me for it.) It seemed to me, having done these computations, that they meant very little. I tabulated zero-order, first-order, and second-order sequential probabilities, estimates of information (H), and the like, but it seemed to me that all that was being shown was that certain phonemes and combinations of phonemes were more frequent than others because of their appearance in words or sequences of words having the higher frequencies. People generate language, I reasoned, not by selecting phonemes but by selecting words and sequences of words; therefore, the frequencies of phonemes and their combinations were mere epiphenomena. Tabulations of these sequences might conceivably have some uses in designing stimulus material for psycholinguistic experiments, to control for the frequencies of habit patterns, but beyond that they would be of little interest either linguistically or psychologically. When Lee Hultzen requested use of my material for his analysis (Hultzen, Allen, and Miron, Tables of transitional frequencies of English phonemes, University of Illinois Press, 1964), I was only too happy to turn it over to him.

Now, what do we find in the work of Carterette and Jones?

These authors must have been disappointed with their findings on the zero-order distributions of letters and phonemes. They find that for the distributions of letters in the conventionally printed version of their conversational samples there is very little change over age. As they say, "the largest change is in m, which decreases steadily from first grade to adult speech." This is "partly accounted for by a decrease in the use of 'um' as a noise word, with the concomitant [sic] rise in the use of 'you know.'" (p. 23). Furthermore, the distribution of letters is about what many other investigators have found for samples of printed English. Etaoin Shrdlu can still be the linotyper's friend! Changes in zero-order distributions over different age levels seem mainly to reflect changes in the frequencies of certain "noise words" like-"um", "well," etc. Yet the authors take pains to compare their results to those of other investigators of phoneme distributions, claiming that "the highest correlations usually occur between phonemic systems derived from material closest to natural speech, whereas the lowest correlations occur with phonemic systems based on material furthest from natural speech" (p. 29). Their argument is not convincing, however, for they seem to underestimate the effect of different systems of phonemic transcriptions used in various studies, and also the effect of the "editing" that occurs as one goes from highly informal speech (with its "noise" words) to more highly edited speech (e.g., contrived speech in high-school plays).

What the authors make most point of are the differences among various types of material, spoken or written, in "redundancy"

or "relative sequential constraint" as defined by information-theoretic statistics. Actually, there are no differences between first-grade speech and the "adult" speech samples in phoneme redundancy--the curves of relative sequential constraint across second-symbol positions (Figure 3.3.1) are virtually identical, leveling off at about .30. I would interpret this to mean merely that both first-graders and adults are using the same (Southern California English) language, and that the same system of phonemic transcription has been used in the two cases. I would be much surprised to learn that first-grade and adult speech samples could not be differentiated in many ways--in lexical selection, in grammatical patterns, etc. Sequential constraint of phonemes is probably not a sensitive way of indexing anything useful or interesting about language samples. It is at least misleading for the authors to state that "[i]n terms of simple sound pattern redundancies, therefore, 6-year-old speech is already adult" (p. 30).

The case is slightly different when the redundancy statistics are applied to letters (graphemes) in the transcriptions of speech, or in printed materials. First grade speech is shown to be slightly more redundant than adult speech; I would think that a large part of the difference could be traced to differences in lexical distributions--differences that show up in letters but not in phonemes because lexical boundaries are observed in the letter statistics but are rarely preserved in the phonemic transcriptions. First grade readers are also much more redundant in letter distributions than even first-grade speech; but it is well known that

first-grade readers are typically highly redundant in their lexical distributions. Redundancy statistics based on grapheme distributions, apparently reflect these lexical distributions, but how reliably, it is difficult to tell. There is the suggestion, arising from these results, that redundancy statistics based on grapheme distributions and their sequences could be a useful surrogate for other types of indexes based on incidence and sequences of words, grammatical patterns, etc. But the authors' suggestion that differences thus revealed between natural speech and written material are somehow important to take into account in the teaching of reading seems rather forced and gratuitous.

The authors also pay some attention to words and sentence lengths, both in the printed material and the phonemic transcriptions. Their "phonemic word" is defined "only in terms of a prosodical feature, specifically a pause in the flow of sound:" They find phonemic words to be three times as long, on the average, as lexical words. They suggest, "Insofar as the units of spoken language and written language are different, the learning of written language (reading) will be difficult," but do not explore the implications of this suggestion further.

Final evaluation. The authors at several places state that the results presented should be used "with great caution." I would say that this caveat must be taken to apply to the whole work. Some linguists, and psychologists, and educators may find uses for the transcribed speech samples, but the limitations of

these samples--particularly in the way in which they are presented--must be borne in mind. One can conceive uses for the statistical tables, perhaps by psychologists seeking ways to control experimental stimulus material for phoneme frequencies. In general, however, one wonders whether it was worthwhile to pursue the statistical analyses and tabulations of phoneme and grapheme frequencies to the extremes reached by Carterette and Jones. It is little wonder that these authors seemed to abandon their interest in their research after completing the manuscript represented in this strangely unfinished book. But more importantly, the authors have not persuaded me that "spoken language" is different from "written language" in any interesting ways. It is conceivable that interesting differences exist between "informal" and "formal" speech, but the authors' analysis has not revealed them.

Pp. xiv + 646

\$25.00.