DIALECTOLOGY BY COMPUTER

Gordon R. Wood

English Faculty, Humanities Division
Southern Illinois University
Edwardsville, Illinois 62025 USA
1969

A familiar postulate is that computer techniques offer to students of natural language an opportunity to examine linguistic evidence far more exhaustively than was possible in the past. Thus the machine may be viewed as a tool used to sort and count data in the traditional ways and to introduce some new variables such as the attitudes of speakers toward their own language as their contemporaries use it.

To consider this latter possibility is to go against some scientific attitudes as defined in descriptive linguistics. As one statement written for the popular press goes "all types of speech (literary, dialectal, rustic, slang, criminal argot, etc.) are of absolutely equal merit [to descriptive linguists] . That is to say, questions of merit or value just do not enter into the picture of linguistic analysis, however important they may be in the study of literature."[1] But in the United States urban linguistics must not only ask what is the English usage of representatives of the poor, the middle class, and the very rich, but also about the social judgments which color the linguistic worlds of those separate groups. As that branch of dialectology is practiced now, the emphasis is on gathering the spoken patterns of language.[2] For these and other students of natural language, the computer promises a means for classifying both the presence and absence of /t/ in <u>often</u> and measuring that information on a scale of attitudes such as "right," "too fancy," and "ungrammatical" .

This is the golden promise. The state of computer alchemy, however, is such that one can profitably examine what has been done with computers as aids in the study of American English dialects.

<u>Purpose</u> .

This paper will sketch the development of computer procedures which have been used in the analysis of regional American English outside of literary contexts. It will note that lists of words and phrases have grown in complexity of count, and that some larger units of spoken sentences are being described . Having shown what has been done, it will suggest additional uses of the computer in preparing clearer displays of local variations in usage.

Collecting techniques.

If dialectology, defined broadly, is the collection and analysis of natural speech in a community, then American dialectology has gone about this endeavor is three or more ways. One is to conduct a controlled interview during which the investigator, guided by entries on worksheets, asks questions and writes the informant's replies in a phonetic script.[3] A second, based on data collected by the first procedure, is to give an informant a printed questionnaire and ask him to encircle those words which he actually uses and to leave unmarked those which are foreign to his habits.[4] A third, seeking to obtain only pronunciation, asks the informant to read a set passage aloud in his normal speaking voice.[5] And the last one to be named asks a speaker to look at a set of numbered pictures and to say something about them; his remarks are recorded by means of a tape recorder and are transcribed later.[6,7]

Analysis and publication.

The mid-twentieth century marks a turning point in the sorting and publication of data about spoken American English. Manual work was required in the entire preparation of a 1949 study of the vocabulary of the Atlantic communities and of a parallel study of pronunciation there -- the latter published in 1961.[8,9] Between those dates Atwood in Texas and Wood in Tennessee had begun to punch their linguistic data on cards for later machine sorting, counting, and publication.[10,11] Comparisons of the tables in Kurath (1949) with those in Atwood (1962) will show that while both works list a common regional vocabulary, the Atwood study shows an increase in the precision of measurements of occurrence. These calculations and the multiplicity of tables showing gradation in word choice among age groups are a reflection of Atwood's use of computers.

Problems in transcription.

If the researcher employs a printed questionnaire, he has no difficulty in converting those words chosen and conventionally spelled into acceptable punched entries which will emerge in the same spelling in a printout. But if he has made a phonetic transcription or if he has a tape recording as his basic document, then he faces a variety of problems in preparing the document for computer analysis. While science fiction , reversing the computer processes which generate artificial speech, imagines a computer that can read a voice tape , transcribe it into its proper regional representation, and compare it with tapes from other dialects, fact requires linguistic researchers to transcribe the documents within the limits of key punch conventions or other restrictions that apply.

Let us consider the problems of a researcher who is using as basic documents a group of tape recordings which have comparable units of discussion stored on them.[12] Conventional English spelling presents no problem. The upper case Roman alphabet of computer printouts serves all of his needs: *cat* will be punched and printed CAT. But he will need to alter the basic evidence in certain ways since he will return to the natural pattern in a different coding. He will normalize all pronunciations of a word into the conventional spelling. If, for instance, some informants say /r/ and others drop it, he will spell it regularly: *law* + /r/ and *law* appear as LAW; *car* and *mirror* - /r/ appear as CAR and MIRROR. And he will want to consider marking distinctions when written conventions use one sign for two functions: *Tom's here* (contracting *is*) and *Tom's hair* (indicating possession) might be distinguished from each other in this way: TOM/S HERE and TOM*S HAIR. As this audience knows all too well, such normalizings help when one comes to sorting a body of evidence alphabetically by computer.

The pronunciation text has to be prepared separately, a second transcription of the amount of spoken detail that the researcher wants to code and analyze. Whether he is transcribing English or some other language, he has to come to grips with the differences between the Roman alphabet and the International Phonetic Alphabet which serves most of us for phonetic transcriptions of speech. Some letters match well enough. Phonetic [b] is adequately represented by B. For others there is no match: phonetic [æ] has no immediate equivalent in the available letters ; and literal Q has no literal match in IPA. Thus if the researcher has transcribed the American English pronunciation of *cat* as [kæt] , he has K and T at hand but must devise something for [æ] . Two solutions are possible. Some researchers prefer to use numerical rather than alphabetic codes for all of the sounds to be represented. [kæt] might be coded by the number pairs 053303, to use a code that has been employed.[13] This writer finds a code like that hard to remember and to read; his preference is for codes that are almost entirely alphabetic. Thus for the lacking [æ] one could substitute the otherwise unused Q , or he could augment one of the existing vowels so that as A* it would stand for [æ] and as A for [a] of *father*. Final choice is determined by ease of reading rather than computer convenience. Is KQT or KA*T the better becomes then a matter of personal preference.

As for coded annotations about stress and juncture, this writer has felt obliged to leave them and other supersegmental traits out of his preparation of the text. Since these phonemic features will vary from sentence to sentence, from person to person, and perhaps from dialect to dialect, it has seemed to him that if they were encoded, a program would have to be devised to tell the computer under what circumstances to treat different codings as being"the-same." Omission reduces the programming difficulties.

Word geographies of American regional English syntax have not appeared; the nearest that one comes to that desideratum is a study of variant verb forms in the Atlantic states.[14] It is here, then, that computer processing can bring together evidence about strings of words that will enable dialectologists to compare standard and local patterns of sentence structure. But let us dispose of a few other details of editing the spoken text before we turn to its analysis.

It does not seem likely that a grammar and syntax of American regional English will emerge from the phonetic evidence. First, a part of the evidence is missing -- the supersegmental part for instance. Second, the tedium of preparing and proofreading the needed amount of text presents an almost insuperable barrier. As a consequence, regional English grammar and syntax will be constructed from sentences conventionally spelled.

The researcher must establish word boundaries and decide on the must suitable grammar. Let us assume that he has chosen a slot and filler grammar on the Fries model.[15] His word boundaries, then, will allow the word to be put into files in computer memory which correspond to syntactic slots. . That is, the researcher must decide what is "one word" and punch the text accordingly. If the source recording has <u>woodpecker</u> and <u>rail fence</u>, as it does, the decision can often go either way. For computer purposes it may be good to present each as "one word" in the form WOODPECKER and RAILFENCE (or if that looks too strange, RAIL-FENCE). Conversely, it may be more economical to present each as"two words" -- WOOD PECKER and RAIL FENCE.

<u>Discovery procedures</u>.

Having stored the dialectal evidence in sentences made up of strings of conventionally spelled words and in segments of phonetic transcription, the researcher must direct the computer to look for details which characterize that dialect and then to apply the same search techniques to hitherto unexamined words, phrases, and sentences.

The obvious first step for American dialectology is to search computer memory for selected words which have been used to identify the major dialects of that language along the Atlantic seaboard. If one has used a printed questionnaire, some words will be stored in memory with an indication that no one chose them; if, on the other hand, one has used a direct interview procedure, he will have provided computer memory with no guide to the missing lexicon. When the computer is directed to follow a concordance or listing program, it will ultimately report that from the printed questionnaire 0 persons chose *awendaw bread* and some half dozen other words, 10 chose *carbon oil*, 73 chose *double singletree*, and so on. Let us assume that the spoken record had the same numbers of known choices -- 10 and 73, but silence for *awendaw bread*; one cannot conclude that *awendaw* bread is either used or unused among those informants. The available raw evidence, then, helps to shape the model of a linguistic community.

The tabulated responses to a pictorial interview manual may give a wider range of local words, hence a different base for the model, than do other techniques. For the dialectally distinctive *faucet* and *spigot*, Table I provides an illustration of the range of synonyms elicited by a picture of a device to control the flow of water. The computer counted these according to the county in which each informant was living. In Table I # 1 stands for counties in east Tennessee and north Alabama, and # 2 stands for those in central and south Alabama. Under other circumstances this code could stand for young-old, city-country, wealthy-poor, or any other pair that was pertinent.

Table I

	#1	#2
FAUCET	12	8
SPIGOT	3	4
WATER FAUCET	4	3
WATER SPIGOT	1	2
HYDRANT	1	1
WATER HYDRANT	0	1
WATER TAP	1	0
HANDLE	4	2

Handle, the last word listed, presents a recurring puzzle when the interviewer turns to his tape recordings. Did the informants really mean the whole device when they said handle ? Or was their attention directed to one of its parts? Unless the interviewer has made field notes which let him say no; he must accept the tabulated record as it stands.

Once the regional vocabulary has been identified, the research is then directed toward other words in the stored text. As their occurrence is matched with what has already been established, the graded lists of local words will be extended. Along with the tables which show the dialect models of faucet and spigot will be similar tables of such diverse words as beard, whiskers, chin whiskers , or bandit, robber, hold-up man , or the names of numbers -- seventeen ninety two, seventeen nine two, and seventeen hundred and ninety two .

The examination of phonological variants will probably come next. For American English the dialectal differences will hinge on variations in the sound of stressed vowels more often than not. A familiar instance is that of pen and pin ; in some dialects the sounds of the stressed vowel contrast, while in others they come under a common phoneme /I/. The computer count of coded phonetic variation shown in two words in the same parts of Tennessee and Alabama as was shown earlier occurs in Table II .

TABLE II

		#1	#2
BENCH =	BE*NC*		
	E*	0	0
	I*	1	1
	A*	1	2
	I*8	2	1
	E*I*	0	0
NEST =	NE*ST		
	E*	1	4
	I*	5	2
	A*	1	0
	I*8	2	1
	E*I*	2	3

This model of allophonic variations obviously is a more complex structure than is that of synonyms in Table I; thus the comparison of the geographic and social boundaries of these two aspects of common language habits is difficult. At the present time it is done by inspection rather than by computer. Of course some of the difficulty would disappear if the allophones were coded as phonemes; this would reduce five symbols to two -- E* and I*. For a sociolinguist or a dialectologist to do so is to lose details that may be of great importance later. As everyone knows, social distinctions are attached to such things as the presence or absence of /r/, the separation of pen from pin, and so on.

The last stage and the newest sort of computer aided research in dialectology is in the grammar and syntax of local sentences. As has already been said, computer research will be directed toward conventionally spelled transcriptions of the spoken sentences. And the grammatical process itself is essentially one of building a concordance.

The first sentences concorded are those which have in them the words which earlier were used to identify the local dialects. For grammatical analysis, the whole sentence is much too long; rather one selects the two slots on either side of the concorded word itself. The concorded word is viewed as a base. Table III illustrates concording with man (synonym of bandit and robber) with words in numbered slots with an A or C prefix, i.e. antecedent and complement. The coded areas of Tennessee and Alabama are listed at the left.

TABLE III

	A2	A1	BASE	C1	C2
#1	IS	A	MAN	HOLDING	ANOTHER
#1	GUESS	THIS	MAN	IS	HOLDING
#1	STARS	A	MAN	WITH	A
#1	IT/S	A	MAN	HE	HAS
#1	A	HOLDUP	MAN	GOT	A
#1	AT	NIGHT	MAN	WITH	A
#1	HOLDING	UP	MAN	...	...
#2	THE	HOLDUP-	MAN	HAS	A
#2	OF	HOLDUP-	MAN	HAS	GOT
#2	HOLDUP	A	MAN	WITH	A
#2	THE	HOLDUP-	MAN	HOLDING	UP
#	THE	HIGHWAY	MAN	WITH	A

Computer programs removed periods and similar internal punctuation in order that those marks would not be treated as text words. This action results in run on sentences like the one with stars at its beginning; obviously at one state stars and a marked the end of one sentence and the beginning of the next. Thus the structural slots are continuous from sentence to sentence, rather like the intersentence syntax of traditional grammars which discuss pronouns in one sentence and their antecedents in earlier sentences. The solution given here is not very satisfactory, but no better one comes to mind when we are faced with the analysis of syntax that goes from one utterance to the next.

Notice , though, that some markers are retained. The slash with its, and especially the presence or absence of the hyphen with holdup . There the marker serves to set nouns (unmarked) off from modifiers of some kind (marked by a hyphen). Again, the marker serves when the columns are searched; it enables grammatical matchings to go in two different directions for the same word provided the matching for dialectal grammar is extended to that degree of refinement.

The grammar-syntax search does not begin with the base word. Its regional dialectal characteristics have already been identified as part of the lexicon. If it is some special variant -- you'uns, youse in contrast with you, you-all -- the concordance program will already have identified it in its several aspects. The concern, then, is with other words. Since the lists of prepositions, articles, quantifiers, conjunctions, and auxiliaries -- are short, they can be put into the computer memory as short dictionaries, each with its identifying label. Then having alphabetized the contents of each column, the computer will seek to match column by column the words there with those in the dictionaries. Each match in column A2 is counted according to its grammatical category, then in colum A1, and so on to the end.

As grammatical categories compete regionally, the computer keeps a running inventory. At present these inventories are simply stated according to the relative frequency of occurrence of each part of speech in that column whether it is pronouns, conjunctions, or something else. Inspection of the computer printout will show man got a, man with a, and man and a , as well as in the morning, of a morning, and mornings. But the present record of variables in syntax is far from clear.

Results and possibilities.

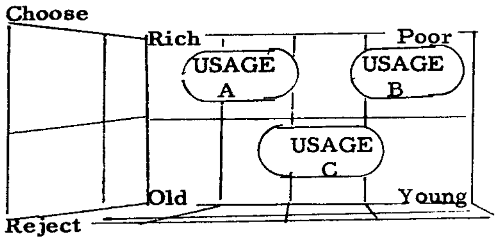
One is not being superciliously critical of the computer based studies of spoken American English when he says that dialectologists have not exploited the full capabilities of computers as research tools in linguistics. In the published work of Atwood and Wood the tables derived from computation differ from those in Kurath chiefly in the refinement of measurements that accompany the word lists. The refinements are certainly needed, but they do not reflect all of the capabilities of the machine to manipulate the data.

In studies that are not intended as dialectology, one finds much of the same thing. The Jones-Wepman word count of spoken English resembles on a small scale the Thorndyke-Lorge word counts of printed English that were assembled several generations ago without the aid of computers. Perhaps the one difference is that Jones and Wepman undertook a part-of-speech classification of their data, a first step toward a grammar and syntax of spoken English. Other related first steps have been taken toward the construction of slot and filler grammars of the main language and a few of its dialects. It does not seemlikely that transformational-generative grammar will contribute much to solving the analytical difficulties largely because that style of grammar is interested in the deep and general structures rather than the surface grammar found in dialectal variations.[16]

As for applicable computer programs, those that serve in the construction of concordances and lists have seemed adequate. In a collection of essays on the use of computers in the study of natural languages, Roger Shuy discussed in broad terms a retrieval program for Linguistic Atlas data once it was in storage.[17] Some recommendations seem to sacrifice the user to the machine: Why code of, with, and from as OFWHFM for a six letter storage of that set of responses when a sub-program could be written to nest OF, WITH, FROM within each other and retrieve them separately or together when needed ? This is a detail, rather on the order of my stated need for a quick and easy way for encoding and storing phonetic information.

The opportunities as I see them are within our grasp -- at least some of them. First is computer mapping of sorted records. In some computer centers there are plotting devices which can use data about the place of origin of a record as a means of drawing maps of several kinds. It is likely that a map of regional English based on the plotted frequencies of choices of local words will differ markedly from the familiar maps of the dialect areas of our language.

Next and of perhaps unsuspected usefulness in urban dialectology (or any dialectology for that matter) is the capability of computers to draw graphs and to rotate these on their axes. Further there is a way of showing measured relationships between things that either differ or are alike but for which no standard scale exists. The researcher can plot (or rather have the computer plot and display it for him) the ranges of agreement and disagreement that the rich and poor in a community have over usages that are "good" or "bad" English. In general form the scale could show agreement at the bottom and disagreement at the top, thus:



[18]

Programming techniques for such non-metrical scaling have been developed. It remains for the dialectologists to discover whether this sort of display is easier to understand when it is presented to general audiences which must be convinced by the evidence than are the usual maps and tabular listings.

And finally some attention should be given to the development of computer programs which allow us to search from vocabulary to pronunciation to sentence structure and back again. In sum, the computer should be brought to a stage of serving as a very extensive search instrument, doing mechanically what the investigator now does by looking and mapping. That is, it should be able to move from lexical BENCH to phonetic BI*NC* to syntactic AN OLD-TIMEY RAIL FENCE. And having moved through this and related evidence should be able to print out that at such and such point a group of different elements converge; e lsewhere they diverge and are replaced by other patterns of relationship.

REFERENCES

1. Hall Robert A., Jr. Linguistics and Your Language. Anchor Books. Garden City: Doubleday, 1960.
2. Shuy, Roger W. and others. Field Techniques in an Urban Language Study. Washington, D. C.: Center for Applied Linguistics, 1968.
3. Kurath, Hans and others. Handbook of the Linguistic Geography of New England. Washington, D. C.: American Council of Learned Societies, 1939.
4. Wood, Gordon R. "Word Distribution in the Interior South," Publication of the American Dialect Society, 35 (1961), 1-16. Also Implicit Change: A Study of Variation in Regional Vocabulary . . . (forthcoming).
5. Thomas, Charles K. An Introduction to the Phonetics of American English. 2nd ed. New York: Ronald, 1958.
6. Sapon, Stanley M. A Pictorial Linguistic Interview Manual.Columbus, O.: (photo-offset) Ohio State University, 1957.
7. Jones, Lyle V. and Joseph M. Wepman. A spoken word count [sic]. Chicago: Language Research Associates, 1966.
8. Kurath, Hans. A Word Geography of the Eastern United States. Ann Arbor: Univ. of Michigan Press, 1949.
9. Kurath, Hans, and Raven I. McDavid, Jr. The Pronunciation of English in the Atlantic States. Ann Arbor: Univ. of Michigan Press, 1961.
10. Atwood, E. Bagby. The Regional Vocabulary of Texas. Austin: Univ. of Texas Press, 1962.
11. Wood. See 7 above.
12. Author's data. Additional examples in his Sub-Regional Variations. . Final Report. Edwardsville: (photo-offset) Southern Illinois University, 1967.
13. Silva, Georgette. "Phontrns: An Automatic Orthographic-to-Phonetic Conversion System for French," Computers and the Humanities, 3, no. 5 (1969), 257-65.
14. Atwood, E. Bagby. A Survey of Verb Forms in the Eastern United States. Ann Arbor: Univ. of Michigan Press, 1953.

15. Fries, Charles C. *The Structure of English, an Introduction to the Construction of English Sentences*. New York: Harcourt, Brace and World, 1952. Efforts to refine this system encounter difficulties: Stolz, Walter S."Syntactic Constraints in Spoken and Written English."(diss.) Univ. of Wisconsin, 1964.

16. Chomsky, Noam. *Syntactic Structures*. Janua Linguarum. The Hague: Mouton, 1966.

17. Shuy, Roger W. "An Automatic Retrieval Program for the Linguistic Atlas of the United States and Canada," *Computation in Linguistics*. edd. Paul L. Garvin and Bernard Spolsky. Bloomington: Indiana University Press, 1966.

18. Kruskal, J.B. "Multidimensional Scaling ..." and "Nonmetric Multidimensional Scaling ... " *Psychometrica*, 29 (1964), 1-27, 115-29.