DOCUMENT RESUME

ED 098 815 FL 006 606

AUTHOR	Miron, Murray S.; Pratt, Charles C.
TITLE	Manual for the Development of Language Frequency Counts.
INSTITUTION	Syracuse Univ. Research Corp., N.Y.
SPONS AGENCY	Defense Language Inst., Washington, D.C.
REPORT NO	AD-775-923; SURC-TR-73-235
PUB DATE	Jun 73
NOTE	60p.
AVAILABLE FROM	National Technical Information Service, Springfield, Virginia 22151 (Order No. AD-775 923, MF-$1.45, HC-$6.00)
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Computational Linguistics; Descriptive Linguistics; Diachronic Linguistics; Etymology; *Language Research; *Manuals; *Mathematical Linguistics; Research Methodology; Structural Analysis; Vocabulary; *Word Frequency; *Word Lists

ABSTRACT

As part of a continuing project of language analysis, SURC presents its final manual. This manual is an explanation of the procedures used to collect and analyze data for this project. After explaining the theory and application of the methodology, the manual discusses specific problems encountered in the design, administration, and analysis of the language data collected. (Author/NTIS)

ED 098315

CONTRACT NO. DAAG 05-72C-0574

SURC TR-73-235

MANUAL FOR THE DEVELOPMENT OF LANGUAGE FREQUENCY COUNTS

JUNE 1973

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FL006606

THE PSYCHO-PHYSICS LABORATORY
SYRACUSE UNIVERSITY RESEARCH CORPORATION
MERRILL LANE, SKYTOP ROAD
SYRACUSE, NEW YORK 13210

MANUAL FOR CONSTRUCTING, INTERPRETING AND APPLYING LANGUAGE DATA

An American tourist visiting Japan often finds himself outside a little shop being followed by a group of curious students. He listens as they gather in a circle giggling and finally pushing one of their number toward the foreigner. "Hello" he says smiling and laughing to hide his embarrassment. "May I practice my English with you?" The surprised tourist usually agrees and there are more student giggles. The topic may be the weather or where one comes from, but rarely does the conversation exchange more than two or three phrases. The student is suddenly limited not only by vocabulary recall but also by what to talk about with his English speaking source. As they have limited time, the students say thank you and move on.

This serves to illustrate one of the greatest problems in the teaching of a language other than one's own. The language that is taught must be a functional one that can be used in an environment that actually exists. That is, it is not enough to store vocabulary and sentence structure because one must store vocabulary and sentence structures that are useful. To this end, word frequency count methods have been used as one way to determine what is useful or functional. The purpose of this manual is to explain a methodology that can be used to obtain these data and how to analyse it once it has been gathered. By using this method, one can devise a functional language pedagogy that will allow a person, within a reasonable period of time, to learn to use a language other than his own effectively. Thus, when a person

"practices" his English or Russian, French, Macedonian, Thai, Ashanti or whatever, he will have the ability to speak more than a few phrases.

Language like other natural systems has been an object of study since man has engaged in such enterprises, or at least for as long as we have preserved record of such study. As a natural phenomenon, it presents a uniquely different challenge to the naturalist, however, which is not shared by those systems which can be construed as purely physical. As with other aspects of human behaviour, it pre-eminently involves intentional motivations which underlie and give purpose to the objective manifestations which are open to study. Thus, the early studies of language as system concentrated almost exclusively upon its intentional aspects; the meanings and symbol processes in whose service it was employed. Two developments, however, presaged a different but parallel method of investigation; the invention of movable type and the rise of enumeration as a measurement tool.

It is the original invention of writing which in very large measure has defined the word entities for which the modern scientist has sought laws. The definition of this entity has remained moot since scribes have sought to record the continuous stream of sound which is language. But, with the invention of movable type and the consequent wide distribution of printed language, the definitions employed by the makers of books if nonetheless arbitrary became at least conventional and consistent. For

without such consistency their products could not have been successful. Thus, in a sense the problem which this paper seeks to address is a man-made problem. We invented a unit called the word for largely commercial purposes and then decided that we should study our own invention by application of another of our inventions, namely counting. Once set in motion, however, the process appears to have assumed a life of its own -- in all regards words appear to have a natural life which share the characteristics of those systems we did not create and their counting has become a scholarly discipline of its own commercial and intrinsic value.

Although measurement by enumeration itself stretches far back into man's time, its early uses were more linguistic and qualitative than quantitative. Measurements of sacks of grain, wealth or live-stock required only that the measurement scale enumerate the finite and directly countable. Such scales have the characteristic that they isomorphically map the objects of enumeration explicitly to an only nominally representative set of numbers. The nominal use of numbers as a measurement device is exemplified by such modern devices as numbering the members of a football team or labeling our coinage with denominations as qualitative categories which only partially reflect their extrinsic values. In such measurement, one moves from few to some through many to too many to count. One speaks of a lot of money or more money than can be counted. The enumeration remains limited by the mechanics of physically mapping the objects into their numerical representations. A clay tablet which is to be used to record the number of animals

involved in a business transaction serves only because its size and the number of potential mappings are well suited. Notions of an infinite number of animals or of negative amounts of wealth were as meaningless as they were impractical. It is meaningless to declare that I have two and a half cents in my pocket or that if you have five cents in yours that you are twice as wealthly as I. And although you may declare that I owe you three cents in exchange for an article set at that value when I protest that I have zero wealth; i.e., that I have -3 elemental units of money, having minus a billion maximal units of enumeration would be treated in precisely the same way, that is, as without meaning. Such vagaries are inconsistent with the precision which is required of enumeration as a measurement tool. The post Renaissance development and acceptance, however, of arithmetic manipulations which bore no extrinsic relationship to the practical usefulness of enumeration suddenly opened a fertile field of speculative and theoretical implications of the natural lawfulness of the countable. It was not, surprisingly enough given the modern acceptance of such operations, until the 16th century that such arithmetic operations as 3-7=-4 were accepted as other than an absurdity. And only still more recently with the introduction of the Calculus that the succession rule defining infinity has been accepted.

The Word as An Attribute of Measurement

Any discussion of measurement as specifically applied to vocabulary must grapple with the definition of the attribute of "wordness." Although

it is clear that the user of a language finds the notion of word psychologically meaningful, attempts to make the notion linguistically explicit have not been successful. Greenberg (1957, p. 27)* has summarized the linguists' position on this matter as follows: "Some linguists deny any validity to the word as a unit, relegating it to folk linguistics. Others believe that the word must be defined separately for each language and that there are probably some languages to which the concept is inapplicable." Nonetheless, Sapir (see Ulman, 1962, p. 39) has observed that "The naive Indian, quite unaccustomed to the concept of the written word, has nevertheless no serious difficulty in dictating a text to a linguist student word by word; he tends, of course, to run his words together as in actual speech, but if he is called to halt and is made to understand what is desired, he can readily isolate the words as such, repeating them as units." For those languages in which there is a rich cultural tradition of writing and literacy, the word as apprehended by its speakers might be construed as little more than the propogation of the conventions of writing. It is in this sense that most counts which define the word as that which is conventionally bounded by spaces in printing define their measurement units. Even in this narrowest of senses, the study of such conventions might be of interest. But, it is the search for the word more broadly considered which is of particular interest: its psychological and linguistic significance.

*References cited in text may be found in the Annotated Bibliography of The Counting of Words, SURC Report, 1973.

The word as psychological unit. The child's original exposure to language is solely vocal in form, we reserve instruction in writing and reading until rather late in the child's development, or at least until the spoken language is reasonably in hand. But even the child's earliest vocal experience involves considerable emphasis on those isolatable units of speech which have unitary symbolic value. The child seeks and is given names of things and these names are typically those conventionalized units which we normally call words. In those high cultures for which literacy is sanctified, the parents of children anticipate and transmit the written language conventions. Thus, it is the rare parent of a high culture child who would respond to a query regarding a drum, with the response "Thatsthethingthatgoesboom." It is much more likely that the parent will pare the response down to the minimally isolatable unit of semantic intent which comes closest to the conventional lexical entry for that object: i.e., "drum" accompanied by an appropriate pointing gesture rather than "Thatscalledadrum" or even "Thatsadrum." Further, once the child is made literate, what may have begun as a printer's convention is perceived as a psychological neccessity which takes on its own significance. Later on should this now literate child be required to learn a second language, he will find it both efficacious and satisfying to learn a vocabulary of words for that language and even to expand his own tongue by study of its lexicon. Finally, the adult speaker of a language with written traditions will unerringly identify upon request what is or isn't a word. And even, according to Greenberg, those adults without writing can do the same.

The word as linguistic unit. Assuming then that a reasonable case can be made for the word as a psychological reality, there remains the question as to whether or not there exists a linguistic definition which can serve as an explication of the concept. That is to say, can we provide an explicit theory of "wordness" which is independent of the user's perceptions of the conventions he employs. Such an explication implies what Chomsky (1957) has called "explanatory adequacy" in contrast to the descriptive adequacy which might be served by reference to the conventions of a particular language, whether written or spoken. Accordingly, it is clear that when considered in this light, the answer to such a question lies at the heart of the complete theory of any language and as such will be extraordinarily difficult to attain. The most modern of grammatical treatments which would seek an account of the structure of language typically eschew the problem as premature, choosing instead to assume the weaker requirement embodied in the presumption of a commonsensical appreciation of what a word is as commonly understood (i.e., psychologically apprehended) by the users of the language.

It is thus not accident that one must search backward into the Bloomfieldian era to find attempts at a linguistic definition of the word, an era for which descriptive adequacy was the prime consideration. Bloomfield (1933) attempted to define the word by reference to the formal characteristics of syntactic boundedness. Those minimal forms which can occur as sentences he termed free forms and those which are never used as sentences bound forms.

Words, as they are commonly used, are those minimal utterance units which can occur as free forms. What distinguishes words from other free forms is that words cannot be split into still smaller forms without leaving a bound form residue.

It is not difficult to find instances of what the speaker of this language would psychological call words which would not be called words by Bloomfield's definition. All compound forms composed of two or more independent words (by either the conventional definition or the definition under test) such as penknife or yardstick provide paradoxical exceptions to the definition. Similarly, the functors such as a or the must occur as bound forms under the definition and yet they are clearly apprehended as psychologically defined words. The meta-language arguments cannot serve to rescue the definition, for all such arguments neccessarily involve the definition we seek as a presumption. Thus, to say that "The." is a permissable sentential response to the question "What is the third word of this sentence?" would only make the issue more cloudly than it already is.

The word as lexical entry. Lexicography at its best represents the structural and functional characteristics of a language as it is conventionally employed, at least, by those who are largely responsible for shaping the culture defined by that language. At its worst, it represents a set of normative prescriptions regarding its language hardly even characterizing its use by those pedants who would prefer proscription to description. The

conventionality of either the description or prescription of its source books is largely dictated by the vicissitudes of publishing and data collection. But such conventionality serves, nonetheless, to represent the conventions of the language usage and as a normative model of such usage to itself perpetuate those conventions. The conventions, in turn, capture the aggregate distillation of the psychological realities by which the language user accounts for his language. New words and new usages replace old conventions at the leisurely pace of slow moving publishers who thus assure that the changes have already been accepted as conventions by the majority of their users. All of these factors in combination serve to make the lexicographer's source book an unequaled arbiter of the problems of defining wordness.

The word as grammatical form. Conventionality in language usage extends beyond the boundries of wordness and arbitrary meaning to function and structure. Grammatical classes or parts of speech as they are more traditionally called, codify by label the functional elements which the language user deems essential to his account of the structures he employs. Whether or not such labels have real explanatory meaning in the theory of language is moot. But, again, as conventions they do have at least psychological meaning which even if without linguistic validity at least deserve recognition by dint of the universality of their acceptance in instruction and perception. And, as before, such purposes are best served by the conventionality of the language's dictionary or alternatively as in this research by a structural definition derived from the mutual substitutability of speech parts in language frames which model their usage.

A special set of problems. There exists a grey area of wordness for which no solutions are readily available. Compound forms that have not as yet made the complete and preferred transition from multiple words through hyphenated forms to single units or fixed collocations too extensive in length to move into the hyphenated life form but nonetheless function as if they were single units, and learned forms which because of the pedantry of their users cannot be tolerated to change, all represent exceptional cases for which it is difficult to devise other than ad hoc and arbitrary solutions.

Inflectional forms in those languages for which such grammatical mechanisms are productive do not, however, represent a particularly difficult problem. It is possible to identify variant forms of the simpler root form on the basis of their derivation from a paradigm. Such a paradigm has the characteristics of regularness and of limiting the number of variants to an absolutely small number. Adverbs, in English, for example, are very largely paradigmatically derived from their more productive adjectival roots by the single pattern form of -ly.

A functor may be defined as any free standing word form in analytic languages which is lexically defined as serving strictly grammatical rather than referential functions and for inflectional languages as that morphological change of the stem which carries such meaning. This definition facilitates the counting of both lexical forms and grammatical patterns.

In the first instance, the working definition of functor is used to suppress those elements which, occurring with such overwhelmingly high frequency, tend to usurp the lower-frequency, but higher information-content forms. In this sense, "functor' is a convenient catch-all for those terms in a language which are finite in number, but which account for a greatly disproportionate frequency of occurrence. In English, nouns, adjectives, verbs, and "pure" adverbs (i.e., those not paradigmatically derived from adjectives) comprise over 99 percent of the total available vocabulary presented in the Shorter Oxford English Dictionary; in contrast to this, we have for all the remaining parts of speech not more than 650 words. Yet these two groups provide approximately equal proportions of the total word usage. While all of the words of the remaining parts of speech (pronouns, prepositions, conjunctions, auxiliary verbs and articles) in English are not strictly "functors", they all share three features of functors: (1) they belong to a small, limited, isolatable class; (2) they have paradigmatic features; (3) they occur with extremely high frequency and, thus, suppress non-functorlike Group I words. It is, therefore, our contention that functorlike words should be treated separately, both for lexical counts and, as it turns out, for grammatical pattern counts.

In the case of strictly inflectional languages, the paradigmatic functors will occur as bound forms in traditional orthography. This presents no problem other than identifying these forms and coding them in such a manner that the "root" form will be the entry into the frequency count. In Latin, for example, *agricolae* would be subsumed into *agricola*.

If the text contained one instance of *agricolae* and one of *agricola*, the frequency tally would show *agricola* as occurring twice.

However, pure analytic and pure inflectional languages are the exception, not the rule. Therefore, the treatment of "functors" in the hybrid languages must allow for the uncluttered tally of words, yet preserve the grammatical patterning of occurrences. Thus, in German, for example, *zum kleinen Kind* would be coded as preposition-definite article-adjective-noun for grammatical pattern and *klein* would be tabulated in its base form for frequency tally. Similarly, in English, *cat* and *cats* would appear as two occurrences of *cat*, since, in English the two forms can be considered as co-occurring items of a paradigm. Verbs would be treated similarly for frequency counts. The total tally for the verb, *run*, for example, would include occurrences of paradigmatic forms such as *runs*, *ran*, and *running*.

There are other common words which should be given separate treatment. For example, numbers, certain kinship terms, days of the week, month of the year, and the like require special attention. The term *Monday* should be taken to include the terms for the other days of the week as though it were a root form from which the others are derived. Thus, all names for the days of the week which are elicited would contribute to the frequency total for the base form, arbitrarily taken to be *Monday*. Similarly, in English, the terms for the members of the nuclear family *(father, mother, son, daughter, brother, sister, husband, wife)* should share a position in

the frequency count equivalent to their total occurrence in the elicitied samples. Words which can be generated paradigmatically from a base form can be collapsed into the base form which will then receive a frequency count equivalent to the total occurrence of the paradigm membership; thus, all variations of verbs due to inflections for person, number, and tense can be counted as instances of the base form.

A final word about word. In the end, the final definition of wordness rests entirely upon the conventions of usage in two senses of use. First, we may interpret and operationalize the psychological apperceptions of the language user for an answer to the meaning of word. We require only that the user recognize and distinguish those units which he would construe as words. We do not require that the user explicitly define or understand the processes by which such recognition is achieved. Where dictionaries exist, these source books provide the best aggregate judgements of such recognition, where they do not we shall have to compile such judgements directly from the speakers themselves. In the second sense of use, it is the purposes of our definition of wordness which must be examined. When the uses of our definition of word are pedogical rather than theoretical, it is surely certain that we shall at least require other tests of that definition; tests which will involve considerations which are as practical as the model tests are theoretical.

Methodological issues in sampling. If one wished to construe a selected corpus of language to be representative of some larger body of language of

which that corpus is sample, the researcher is compelled to provide a rational defense of the sample's representativeness. Selection by random strategy is designed to provide such justification on the grounds that a random sample requires that all members of the population had equal probability of being selected as members of that sample. Under such rationale, the occurrence frequencies of the units of analysis are both efficient and unbiased estimators of the population probabilities of those units. But then two problems arise, random with respect to what and how are we to translate random into a set of explicit procedures? The overwhelming bulk of research on vocabulary has concentrated on the written forms of language, the number of worthwhile spoken analyses numbers less than half a dozen. The preceding work* has reviewed and evaluated these studies. The populations represented by the spoken and written forms of a language are both different and same when viewed from differing standpoints. We have argued that at the level of the functor, the vocabularies of speech and writing are as alike as the linguistic code is inflexible with respect to their grammatical function. At the level of substantive choices, the two are as separate as the distinction made by the culture between informal and formal styles of communication, with an extensive penumbra area of overlap between those styles at the level of the higher frequency substantives. And from still another viewpoint, the two communication forms may or may not be different with respect to their schemapiric lawfulness, a consideration which has already been considered in our previous work.

*See The Counting of Words, SURC, 1973.

But in a sense even the distinction being made between speech and writing is itself artificial for some purposes. Plays are written to be spoken and all writing must be speakable if it is to conform to its parent linguistic code. Nor is it simple to classify the procedure which this research proposes as the optimal sampling strategy; that of eliciting restricted associations from the users of a language. That procedure is designed to bypass the written-spoken dichotomy by sampling from the highest frequency items of the users vocabulary. The rationale of that assumption, in turn, rests upon the spew hypothesis. Under that rationale the problem of corpus length is also largely avoided, for no attempt is being made to fully sample the entire frequency range of vocabulary items as they appear in the population. The spew hypothesis quite simply posits that "...the order of emission of verbal units is directly related to frequency of experience with those units." (Underwood and Schulz, 1960.)

A number of studies have provided strong support for such an assertion. Johnson (1956) demonstrated that 84% of the most frequent associations to the Kent-Rosanoff stimuli occured with a frequency of 50 times or more per million in the Thorndike-Lorge list, whereas only 48% of the least frequent responses had equally high ratings. Howes (1957) computed the correlation between frequency of associations to the Kent-Rosanoff list and frequency of words in the language to be .94 if functors are excluded from consideration. The effect has even been demonstrated when subjects are asked to provide male given names; those names which occur most frequently in the

written language are also those most likely to be given by a subject (Cromwell, 1956). Bousfield and Barclay (1950) have also demonstrated that the order of emission of verbal units is directly correlated with their frequency of occurrence in the language.

Taken in its weakest sense, the spew hypothesis is not an hypothesis at all. It is obvious that if emission of verbal units is taken to include all uses of the language, the complete tabulations of such emissions are the frequencies of those units. But in its strongest sense, the spew hypothesis provides a sampling strategy for estimating the total linguistic probability of verbal units. Construed as ad libitum responses, associational responses obtained from subjects provide a higher face validity procedure for estimating the frequency of spoken language units.

Either spoken or written data suffer from several inherent difficulties which accrue to the nature of natural language codes. The lawful statistical nature of such counts always produces a frequency ordering in which roughly half of the occurrence types have token realizations which are at the limits of measurement; i.e., have single occurrence frequencies. Probability estimates of population frequencies from such inherently errorful sample frequencies are statistically unreliable. At the high frequency end of the distribution of such word samples one consistently finds that function and interstitial words account for disproportionately high percentages of the total sample. The situation is roughly analogous to using the Wall Street

Journal to determine the frequency of English units. From such a data base, ordinal numbers and fractions would dominate the frequency distribution of the count. Functors, as is the case with numbers, are important to the language, but they displace and minimize the importance of the other substantive form classes because of their overwhelming prominence in natural languages. Foreign language instruction has typically met this difficulty by sub-dividing the lexical units of the language into separate form classes. Such form classes are fundamental to any description of a language. They function at elemental levels in both phrase structure and transformational rules. The speaker of a language only rarely can make explicit the category rules which define such grammatical classes and, even in these rare cases, such explicitness is typically incorrect. However, the speaker does use such rules in the construction of any utterance, his inability to provide an explicit account of the nature of those rules is not evidence against their functional utility. If the speaker is given a contextual frame which calls for a unit from a particular grammatical class, the speaker can provide an appropriate completion. Further, the choice of the particular completion within that functional class is apparently determined by the frequency of experience of that unit. Thus, elicitation procedures which call for grammatical class associations in specified frames simultaneously solve two problems otherwise encountered in frequency counts: 1) all token frequencies are automatically marked by function class and 2) frequency determinations of unit types are separately determined within function class, thus increasing the pay-off yield of the data collection.

For the continuous language samples which we have also taken, the issue of corpus length is as crucial as it is difficult to answer. Rapoport (1965) addressing himself to this problem with regard to speech samples has argued that:

"In the selection of speech to be analyzed, the question of how long the transcript should be, though practically important, is not easy to answer. Intuitively it would be nice to have very long transcripts, 5000 words or more, in order to get a substantial sample of the subject's vocabulary. Practical considerations, on the other hand, call for smaller samples. In addition, it might not be feasible to obtain very long samples of connected discourse from the subject. Without considering some exceptions, people usually do not utter 5000 words and more in one session on the same topic. It seems that a proper solution to the length of the transcript is an empirical one. Sample sizes should be considered within the range where the mathematical form of the observed distribution of word-frequencies is not markedly changed."

And then after reviewing data similarly collected by Howes and Geschwind (1962) who claimed that: "These data show that even for samples of 1000 words, there is excellent correspondence between the theoretical equation and the empirical distributions. The considerations suggest that for most purposes samples of 2000 words are adequate for estimating parameters of [spoken] word-frequency.", Rapoport concludes that: "It thus seems that the sample sizes used [in the Rapoport study] (between 1000 and 5000 tokens) are appropriate."

In addition to the question of optimal sample size, there still remains the question concerning the method of sampling. The samples must be sufficiently scattered with respect to subject matter so as to avoid the vocabulary biases inherent in the ideational clumping which characterizes language. Yule (1944) has specifically rejected the random strategy of sampling in favor of spread sampling. This technique spreads the sample as uniformly as possible over the whole range of the work to be sampled. Yule's suggestion was to select a sample of words from each page, the words being samples within the page unit taken either at random or from a continuous passage of a prespecified number of lines. It should be observed that the technique which we have employed for the sampling of American television in this research is a spread sample based upon randomly selected continous segments of five minute duration. The procedure is quite straightforward. A clock activates a tape recorder for a five-minute interval during each hour of total speech time. The specific five-minute interval is varied in a psuedo-random fashion so that different five-minute segments are sampled at each hour. The technique for accomplishing this sampling is instrumentally simple. The minute and hour hands of a normal clock coincide at a different locus during each hour of a day. The specific time of coincidence is given by the equation:

$$(0) \quad h{:}5h + \frac{5h}{12}$$

assuming the clock is started with the hands at 12 midnight. Thus, for example, the first coincidence of the hands would occur at 1:05.3, the second at 2:10.8 and so on. As real time progresses through the day, the

five-minute sampling segments precess further into the hours. In order to avoid this consistent precession, the clock is randomly started at a different clock time each day.

DERIVATION AND PROCEDURE

The analysis of the extensive literature on vocabulary counts which SURC has prepared and submitted in conjunction with the annotated bibliography of that literature indicates that no single procedure for ascertaining the minimal instructional vocabulary of a language adequately answers the problems associated with that task. Several difficulties accrue to a procedure based upon the counting of written language forms. (1) Huge corpora must be assembled in order to achieve sufficient reliability for low frequency but high information substantives uniformly usurped in such counts by the ubiquitously occurring grammatical functors. (2) Stylistic differences in the written forms of the language cannot adequately be guaged because of the requirement that each stylistic source by represented by a data base at least as extensive as that required by the entire count irrespective of style. (3) Such a procedure is entirely insensitive to the expected style variance which accrues to the spoken form of the language even when such expedients as using quasi-spoken samples (e.g., plays) are employed. On the other hand, this procedure has been employed despite its obvious flaws because it does solve those problems which are intrinsic to a procedure which employs samples of the spoken language. Such purely spoken counts encounter difficulties in defining word units, entail inordinate

difficulties in recording and transcribing the speech and are rarely extensive enough to permit reliable measurement of infrequently occurring lexical units. Both of these procedures share the inadequacies which must necessarily accrue to the tabulation of single lexical units whether orthographically or phonetically defined. Such lexical counts destroy the syntactical organization of the data base and as a consequence lose both the linguistic and situational contexts which serve to distinguish homologous forms which may differ either grammatically or semantically. Neither procedure is sensitive to the compound lexical forms which characterize collocations whose meanings are emergently different from their constituent parts.

The third procedure which has most recently been devised and which formed the essential basis of SURC's attack on these problems involves a form of subjective frequency estimation. A frequency count of either the spoken or written forms of a language purports to define a representative sample of the population of occurrences of lexical units in that language. Any natural language, by definition, syntactically organizes the elements of communication into lawfully regular patterns of rule governed sequences. Such regularities are reflected in the language samples by disproportionate occurrences of the elemental communication units. For most purposes, these units are apprehended by the speaker of a language to be those units, i.e., words, which are set off by space bounding symbols in the written form of the language. Retrieval of such word forms, even in the solely spoken forms

of the language, is both easy and efficient for the users of the language. Since the lawful differences in occurrences of such word forms is substantially dictated by the linguistic structure of the language and that structure must be part of the competence of the speaker of a language, it is possible to use this speaker's competence to subjectively ascertain the predictive differences in usage of these units. Viewed in this manner, the speaker of a language can be conceived as a form of automaton whose structure dictates a particular order of retrieval and usage for the lexical units of its communication capabilities. There can be no basis for quarrel with the assumption that the usage of the lexicon of a language on the part of this automaton reflects the occurrence probabilities of the code which it shares with other members of its linguistic community. Clearly, considered in the aggregate, it is this usage which determines those occurrence probabilities and the assumption that the order of retrieval of these units will reflect the usage probabilities is, in turn, strongly supported by a number of empirical investigations. Retrieval order, usage, and subjective frequency estimation thus become interchangeable methodologies which both empirically and logically can be shown to be covariates dictated by the structure of the language.

Language structure, however, must be considered from at least three standpoints. The purely linguistic structure of a natural language code provides the organizational framework within whose rule boundaries idea-

tional content may be expressed in a style which is constrained and is slave to the former two. Communication which does not conform to syntactic structure is empty. Communication styles which do not serve ideational intent are disordered. These ordered constraints upon communication successively delimit the patterns of usage in a language. All communication within a linguistic community share a common syntactic structure for it is precisely that commonality which defines the mutual intelligibility of the members of that community. Ideational content within this common code is bounded by the communicative intent of the speaker, his experiential universe and his social needs. The style of communication is, in turn, bounded by personal variables of idiosyncratic origin, values shared by only smaller sub-groupings of the community, or identifiable attributes recognized by the entire community as serving particular ideational purposes. Both style and ideational differences in communication are largely reflected by substantive lexical choices or at least by communication choices which are markedly different from those dictated by syntactic structure. Style and ideation do not govern the correct usage of the subject-verb agreement in English. Proceeding from the level of maximum constraint to least; the speaker assesses the style requirements of his communication of a particular content in a syntactically well-formed utterance. Viewed in this manner, patterns of language usage successively are controlled by narrower and narrower constraints which move from universal characteristics of the entire linguistic community to the individual differences of each member of that community.

Thus, the overwhelmingly high frequency of occurrences of grammatical functors in any frequency count of a language merely reflects the universality of the syntactic constraints which bind all members of that community together. Commonality of the nominal elements of a language reflect ideational contents of styles of communications which are temporally, topically and situationally common to segments of the linguistic community. Since the diversity of these situational contexts of communication is quite large, it requires an extraordinarily large data base to begin to detect any commonality in usages. Style and ideational content differences are reflected in specialized vocabularies which are best revealed by methods other than those of the large vocabulary counts. For such determinations, closer control of the situational contexts can more efficiently and parsimoniously be utilized. By contrast, the grammatical regularities which are common to those differing contexts can readily be ascertained from modestly small data bases precisely because they are less sensitive to situational control.

Thus, what is required is a combination of approaches which maximizes the advantages which accrue to each of the separate techniques. Function word usage can be assessed by relatively short samples of running speech with the expectation that differences in source informants or situations will not produce substantial reorderings of their frequencies of occurrence. Substantive or content forms can be assessed by frame elicitations which can be modified to capture those aspects of the communication which are of interest.

Sentence Frame Elicitations: Procedure I

In order to assess such differences in content of the communication sources, it is clear that the raw frequency counting method is not the method of choice. For that purpose, it would be more efficacious to elicit the content forms from the speaking automaton by controlling situational contexts. Such a procedure provides a specific stimulus to the informant and asks that he retrieve an association to that stimulus which he judges to be apposite both stylistically and semantically. By varying the informant characteristics, e.g., military personnel, students, salesmen, etc., one may capture those salient vocabulary differences which characterize their groups. By varying the stimulus contexts either within or across subject groupings, one may tap particular discourse domains of interest, e.g., situational differences characteristic of various cultural settings, occupations or subject matters. If the stimulus contexts are constructed as sentence frames or fragments, the elicitation method simultaneously defines the form class of the semantically apposite association. For the general purposes of the methodology best suited to the instructional needs of the DLI, SURC proposes a standard list of 100 substantive stimuli ascertained from previous research to be culture fair and heterogeneous in semantic content. The list of elicitation stimuli as used in the three test languages is presented in Table 1.

Choice of productive sentence frames in a language constitutes a particular problem for the elicitation methodology. What sentence patterns are

TABLE I

HORSE	MARRIAGE	GAME	COLOR	HEART	FRIEND
DEATH	KNOWLEDGE	FREEDOM	BELIEF	SUCCESS	ROPE
HAND	MOTHER	KNOT	LIFE	HEAD	THUNDER
TRUTH	AUTHOR	MUSIC	SLEEP	FUTURE	EGG
ROOT	SUN	DOG	MONEY	SMOKE	FISH
MAN	WEDNESDAY	CHAIR	GUILT	LUCK	PEACE
HAIR	FOOD	SEED	POLICEMAN	FATHER	FEAR
PLEASURE	PURPOSE	FIRE	DOCTOR	POWER	WINDOW
RIVER	WATER	HOUSE	GIRL	PICTURE	MEAT
TRUST	PAIN	DEFEAT	BOOK	LAKE	STAR
BATTLE	DANGER	SYMPATHY	PROGRESS	CUP	COURAGE
THIEF	BREAD	LOVE	FRUIT	BIRD	SNAKE
HEAT	MAP	HUSBAND	RAIN	TREE	STONE
TOOTH	EAR	RESPECT	LAUGHTER	MOON	WIND
WORK	STORY	PUNISHMENT	WEALTH	WOMAN	CLOUD
CAT	POISON	CRIME	HUNGER	CHOICE	NOISE
NEED	HOPE	ANGER	TONGUE		

One hundred words used as stimuli in the English portion of the project.

FARASI	YAI	NGUVU (pawa)	TUNDA	KUKOSANA NA SHERIA
NDOA	MZIZI	DIRISHA	NDEGE	NJAA
MCHEZO	JUA	MTO (river)	NYOKA	UCHAGUZI
RANGI	MBWA	MAJI	JOTO	KELELE
MOYO	PESA	NYUMBA	RAMANI	HITAJI
RAFIKI	MOSHI	MSICHANA	MUME	TUMAINI
KIFO	SAMAKI	PICHA	MVUA	HASIRA
UJUZI	MWANAMUME	NYAMA	MTI	ULIMI
UHURU	JUMATANO	IMANI	JIWE	
IMANI	KITI	MAUMIVU	JINO	
MAENDELEO	KOSA	KUSHINDWA	SIKIO	
MAMBA	BAHATI	KITABU	HESHIMA	
MKONO	AMANI	ZIWA (lake)	UCHEKO	
MAMA	NYWELE	NYOTA	MWEZI	
FUNDO	CHAKULA	VITA	UPEPO	
MAISHA	MBEGU	HATARI	KAZI	
KICHWA	ASKARI POLISI	HURUMA	HADITHI	
NGURUMO	BABA	MAENDELEO	ADHABU	
UKWELI	WOGA	KIKOMBE	UTAJIRI	
MWAWDISHI	RAHA	USHUJAA	MWANAMKE	
MUZIKI	NIA	MWIZI	MAWINGU	
KULALA	MOTO	MKATE	PAKA	
SIKU ZIJAZO	DAKTARI	MAPENZI	SUMU	

One hundred words used as stimuli in the Swahili portion of the project.

TABLE I (con't)

暑さ
地図
御主人
雨
木
石
歯
耳
尊敬
笑い
月
風
仕事
お話
豊富
女の人
雪
ねこ
毒
犯罪
飢え
選択
騒音
必要
希望
怒り
舌

馬
結婚
ゲーム
色彩
心
友達
死
知識
自由
信仰
成功
絹
手
母
結び目
人生
頭
かみなり
真優
作者
音楽
眠り
未来
たまご
根
太陽
大
お金

煙
魚
人間
水曜日
いす
罪
運
平和
髪
食べ物
植
警官
文
恐怖
快楽
目的
火
医者
力
窓
川
水

家
女の子
絵
肉
信用
痛さ
敗北
本
湖
星
戦い
危険
同情
進歩
コップ
勇気
泥棒
パン
愛
果物
鳥
蛇

One hundred words used as stimuli in the Japanese portion of the project.

TABLE I (con't)

employed and what are the most productive form classes in a language? SURC dealt with this problem from two approaches.

Linguistically trained informants in the language of interest were used to devise sentence patterns which in their judgment were critical for that language. In addition, native speakers were asked to supply sentences in ad libitum fashion which employed the stimulus substantives. The construction of this frame is fairly easy. Instructions to the subjects should request them to construct a grammatically sound sentence using the 100 stimulus word which should then be listed. It should be pointed out to the subject that they are to respond to the word as quickly as possible and to construct sentences that they would use in everyday speech. Table II details the frames employed in each of the three test languages.

Procedure I: Statistical Analyses

For each elicited response within a specified frame as supplied by 30 informants in each language, the information theory statistic, H, was computed. Statistical information in this instance is defined as the probability distribution of the word responses across stimulus items. Figure 1 illustrates a sample calculation for two adjective responses. Linguistically, H implies greater productivity for the higher information terms. In general, the absolute value of H increases as the total frequency and diversity of the word (i.e., the number of different stimulus items with which a given response is associated) increase and as the distribution of frequencies of responses to each stimulus becomes more nearly equal or rectangular. For two words with equal diversity and frequency values, the word with more equal distribution of frequency of responding will produce the greater H value. A value of H = 0 is obtained whenever a word occurs as response to only a single stimulus frame; i.e., when the word has diversity equal to 1, regardless of the relative frequency of that qualifier.

Frame Constructions and

Examples Used to Help Subjects Understand Desired Responses

	English	Swahili	Japanese
Adjectives	(The) Box is square.	Msichana ni mrembo.	あのテーブルは ＿まる＿ い。
	(The) Box is too heavy.	Msichana ni mwembamba mno.	あの小包は ＿小さ＿ い。
	(The) Box is large.	Msichana ni mjanja.	あのかさは ＿くろ＿ い。
			動作は ＿にぶ＿ い。
Transitive Verbs	The man sat on the box.	Mwanamke amekaa juu ya meza.	あの人は かさ (を. に. で. が) ＿さす＿。
	The man was hit by the box.	Mwanamke aliamshwa na mwanawe.	
	The man burned the box.	Mwanamke anasuka nywele.	あの人は 雑誌 (を. に. で. が) ＿言さました＿。
Intransitive Verbs	The box broke.	Kiti kimevunjika.	テーブル (は. が) ＿こわれた＿。
	The box fell.	Kiti kimeanguka.	小包 (は. が) ＿倒いた＿。
	The box was empty.	Kiti kilikuwa kibovu.	かさ (は が) ＿やぶれた＿。
			動作 (は. が) ＿止む＿。

Frame Constructions and

Examples Used to Help Subjects Understand Desired Responses

		English*	Swahili	Japanese
Nominals	Box	car	Myumba chumba	小包 郵便
	Man	plumber	Msichana mvulana	雑誌 広告
	Tar	road	Picha sinema	肉 牛肉

* While frames for verbal and adjectival types are fairly obvious, a frame to elicit nouns is potentially difficult for the subject to use. While no problems occurred in Swahili or Japanese, subjects in English did not always respond with usable noun forms. Therefore, it may be advisable to instruct subjects as to the part of speech one is looking for in addition to the several examples as suggested in the table.

TABLE II (Cont)

Example H and PHI Calculations for English Adjective Frame.

"This S_i is _____."

Stimulus Item (S_i)	*good*	*nice*
1	10	2
2	0	0
3	0	0
4	4	5
5	0	0
6	8	0
7	0	0
8	0	1
9	0	0
10	6	3
.		
.		
.		
100		

A_J

f_{IJ}

$$H_J = -\sum_{I=1}^{100} \frac{f_{IJ}}{N_J} \text{Log} \frac{f_{IJ}}{N_J} \qquad \text{Where } N_J = \sum_{I=1}^{100} f_{IJ}$$

$$H_{good} = -\frac{10}{28}\text{Log}\frac{10}{28} + \frac{4}{28}\text{Log}\frac{4}{28} + \frac{8}{28}\text{Log}\frac{8}{28} + \frac{6}{28}\text{Log}\frac{6}{28} = .58$$

$$H_{nice} = -\frac{2}{11}\text{Log}\frac{2}{11} + \frac{5}{11}\text{Log}\frac{5}{11} + \frac{1}{11}\text{Log}\frac{1}{11} + \frac{3}{11}\text{Log}\frac{3}{11} = .54$$

$$\text{Phi} = \frac{AD - BC}{\sqrt{(A+B)\times(C+D)\times(A+C)\times(B+D)}}$$

Where:

	+	-
+	A	B
-	C	D

$$\text{Phi}_{\substack{good \\ nice}} = \frac{15 - 1}{\sqrt{(4)\times(6)\times(4)\times(6)}} = .58$$

FIGURE I

31

Should a foreigner, with limited time for study, learn both words of the example? Both words have high information content, both words are relatively frequent responses and both words have high linguistic diversity of application to a wide range of concepts. The answer to the question lies in the proposed calculation of the correlation between usage in different contexts. It will be observed that the example words good and nice with only one exception always co-occur in the same conceptual environment, i.e., they exhibit non-complementary concept distribution. By analogy to phonemic discovery techniques, this would mean that they are alloforms of the same meaning class. For the purposes of language instruction, this means that one of the forms could be dropped, since each can be used in the same environment; each is linked to a specific concept category. The phi coefficient which indexes these distributions as illustrated in Figure 4 was calculated for each word form in comparison with all other word forms elicited by a given frame. Initially, the words elicited by a particular sentence frame are ordered from highest information to lowest. Then, in turn, each higher ranking word is correlated with all lower ranking words. Words with similar, or identical, distributions of occurrence are culled from the list as the process proceeds. The computer print-out records the value of the correlation and the highest correlating word in all instances. This procedure thus culls from the list those words which are distributionally similar and, hence, meaningfully similar so that the retained list comprises the minimally efficient set of forms.

Frequency Counts of Spoken Language: Procedure II

In order to obtain the list of functors in each of the languages, a subset of the informants used in the elicitation tasks were assembled for a prompted discussion session of two hours duration. These sessions were recorded in their entirety. Discussion was allowed to range freely over any subject the informants wished to discuss and minimally prompted by questions from the experimenter only at those points where the discussion lagged. In English, these recording sessions were supplemented by samples taken from wireless microphone and television broadcasts. Sampling from these latter sources is achieved by means of a clock which automatically samples five-minute segments from each hour.

The same strategy has been used for recording the live speech of subjects who have agreed to wear a wireless microphone. Familiarization runs of two days in which no recording was taken were instituted to avoid content biases of the speech.

The wireless microphone speech sources were not taken for Swahili and Japanese. All other data sources including a media sample were, however, included. It should be noted that we do not recommend the use of a wireless microphone speech source. The main reason is legal since the use of such devices places one in a questionable legal status as to invasion of privacy. We found that candid remarks can be obtained just as easily by using the discussion session.

Procedure II: Statistical Analyses

Straight frequency counts of all word units as obtained from the data sources were made. The frequency ordered listings of the resultant word types are coded as instances of the base forms from which they derive. In addition, the context of each word type is listed as a key word in context (KWIC) tabulation. Computer print-outs of these contexts provide an exhaustive listing of every contextual usage of the obtained vocabulary which should be of material aid to the language instructor seeking example utterances which employ the Key Words.

SPECIAL PROBLEMS

Some problems were noted in these various methods. In the discussion sessions, attention should be paid to speech speed. All of these sessions were two hours in length for this project. Experience with English showed that that particular length produced a reasonably sized corpus to analyse. Thus, it was suggested that Japanese and Swahili be the same length. What we did not take into account is that Japanese is spoken much faster and Swahili a little slower than English. This produced a large corpus in the former and smaller corpus in the latter. Thus, in gauging one's time, it is advisable to judge corpus length in terms of the normal speaking speed of the language.

The sentences and frames produced some problems also. Subjects should be throughly briefed before being given the elicitation. It should be fully explained in terms of one's objectives and it should be emphasized that you are after normal conversation. Encourage the subjects to answer every stimulus but also allow them to skip a stimulus if they cannot achieve an answer. The instructions we employed are presented in the accompanying display. If mechanical recording equipment is used, insure that the subjects are instructed in its operation. For example, if cassettes are used, remember that they have a long lead tape. As many as three responses may be lost before the subject is actually recording. Emphasize to the subject that this is not a test or some trick to measure their language comprehension. We found, especially with Swhaili, that this was a hampering factor. Our Swahili dis-

INSTRUCTIONS EMPLOYED IN ENGLISH FRAME ELICITATIONS

This task is designed to see what words people use in certain sentences. The booklets given to each of you contain a series of sentence fragments along with a set of key words. For example, on the first page of the booklet you will see a sentence fragment and below it a series of words and blanks:

THE_____IS_____.

HOUSE _____
GIRL _____
PICTURE _____
etc.

We would like you to complete each of these sentence fragments with whatever single word first comes to your mind as being most appropriate in your language. For example, you might have completed the first sentence fragment as follows: THE HOUSE IS large, and the second fragment with THE GIRL IS pretty. You should continue on through each of the sentence fragments in this manner until you have completed all of the items of the booklet. There are a total of 100 key words for each of the different types of sentence fragments. You should work quickly without puzzling unduly over any of the items, we would like to have your first impressions; the words which immediately come to your mind in each case. You may, of course, use the same word to complete as many of the sentence fragments as you wish, we only want the words which you think are most appropriate. Do not try to be literary or technical in your response, try to think in terms of your normal conversational style of speech. If you feel that a short phrase or combination of words seems to you to be most appropriate for some of the sentence fragments, you may use that response. Remember it is whatever response seems to you to be most appropriate that we want.

Because of the nature of this task, you may find a number of instances for which you simply cannot find an appropriate completion for the sentence fragment, or for which the sentence fragment cannot make sense. In such cases simply leave the response space blank and go on to the next item. Each new type of sentence fragment will be given at the top of the page of a new section of your booklet along with an example completion to aid you.

Work at your own speed and if you become tired take as long a rest as you would like. Are there any questions?

cussion session was cancelled several times because subjects were afraid to respond for personal and, in some cases, political reasons. In Japanese, some subjects felt they could only use their best Japanese and gave responses that were very classical, almost poetic in nature. We found that the subjects should be paid for their efforts and the recording should be set up in a relaxing environment that is free from disturbance. It is advisable to provide some refreshments to further add to the atmosphere of informality. This will free the individual to talk about things he feels are important and topical.

PREPARING THE DATA

There are two ways to collect one's data samples. One way is cassette or reel to reel tape devices. Such allows the subject the convenience of merely sitting and responding. More over, we found that subjects seldom went back and did sections over again. That is, the first response was recorded which is in keeping with the rational of the spew thesis. The problems with this approach, however, are several. Technical problems often occur with these machines as subjects fail to record or record at the wrong levels. This means that close supervision has to be maintained, but such is liable to upset the subject or add to any imagined fears discussed above. Cassette problems are familiar to anyone who has ever owned or worked with one. In addition to the lead tape problem, cassettes have a tendency to twist, get off their tracks, or foul around the spindal mechanism. Such problems can

destroy an entire subject's response and reduce one's overall corpus size. Thus, if one wants to use these technical devices it is advisable to have more subject data then one needs. We found that at least 10% of our cassette recorded data was lost and had to be replaced with extra data.

An alternative to this is to have the subject write his responses on the stimulus questionnaire. This gives a set of responses that can be checked before one releases the subject. The chief problem with this approach is that some of the spontaneity may be lost or the subject may go back over his work and put in answers he thinks you want to see. Cassettes are much easier to store and less likely to be lost while papers can be lost in the piles in one's office if not carefully filed. Our experience showed that papers gave us more responses because subjects using our cassette system seemed to feel the pressure of the turning tape and tended to give no response if they couldn't think of one (even though these machines had pause devices). Paper responses tended to be filled out completely. Cassette reproduction into a computer system required a cassette recorder with extra equipment to enable a typist or keypuncher to produce a transcription. Papers were just carried in and work begun.

One final consideration in this area of preparation is necessary. If one is measuring a language that uses script other than Romanized characters, one must include an additional step in preparation. It is important that this transcription be a uniform one. In the Japanese portion of the

project, the pressures of time forced us to find as many transcribers as possible. One person used the Hepburn system while the others used the Block-Jorden system. This meant an additional step to produce a uniform transcription.

COMPUTER ANALYSES

The data that one has compiled must now be analyzed. Getting it correctly prepared for the computer is the first step. One method is to keypunch the responses on cards and to read them directly into the system. This step is somewhat time consuming and very cumbersome. We dealt with 30 subjects at 100 responses per frame, four frames each, which produces 12,000 cards for the frames alone. If one were to copy sentences and discussions at one sentence per card, one would have a monsterous collection of computer cards. To avoid this problem, we used the IBM Magnetic Tape Selectric Typewriter (MTST). The MTST produces magnetic images of the material which can then be converted to a nine track magnetic computer tape. All that has to be done is to type the information and code it in a manner that will make manual retrieval easier. It should be pointed out, however, that both of these methods are very error prone owing to the gargantuan size of the material. It is an absolute requirement that it be edited thoroughly before it is placed in the program. A simple error like spelling horse hrose will produce two counts in which both horse and hrose appear as unique forms. Errors are easier to correct on the MTST

since no manipulation of cards is required, but one must also insure that problems like field definition and carriage return definition are also tightly controlled or items will fall out of sequence or disappear. We found that at least two people are needed to edit these items and that each step's results need to be edited to insure an error free print-out product.

Once the data have been collected and stored on a 9 track tape or cards, the process is ready to begin. Our first step was to list the text. This produces a print-out of everything on the tape. The next step is to count the words. The machine is commanded to print out each word alphabetically in the context in which it appears, add the total number of times each word occurs, and print the number of occurrences. This is called the Key Word In Context (KWIC) print-out. Using these data, the machine then produces another print-out of each occurrence by frequency and without the context. The problem is that this count is not a true semantic count since all forms are printed as unique. Thus, one finds can, can't, could, and so forth. The next step is to code such word types as instances of a common base form, in this instance as the verb can, this requires an editing step that must be done by hand. A base form is hand tallied along with the total number of words that are variants of that base form. Thus, one might find the entry of boy 5 and variants of boy 2, boys 3. Once these data are assembled, it can be presented in many formats. We chose to present it once by frequency and then again in alpha-

betic order with the variants as sub-listings. The question of what is a base form and what is a variant is entirely up to the analyst and his own particular theory of language. For example, we chose to leave nominal forms of verbs as unique cases, thus combinations is listed as a variant of combination, even though both might be construed to be derived from the verb to combine. Our reasoning is based upon the assumption that it is a very questionable decision to place forms which function as nouns in the verbal form. Other linguists might not agree. The point is, that the method is flexible enough to allow this sort of divergent opinion.

One other area should be noted in reference to the selection of data to be analyzed. The methods of analysis and the items to be analyzed mentioned in our study are forms which we found acceptable for our purposes. Other individuals might not agree, but the program is flexible enough to allow variation. For example, in Japanese and Swahili, we analyzed our media sample, discussion, and sentence elicitation words as one unit to produce a combined semantic frequency count. Our reason for doing this was to produce one list of total words obtained rather than three lists. The point is simply that one can combine or separate the analysis steps as one chooses. However, it is our recommendation that the final frequency count because of the enhanced diversity of sampling sources should combine the separate sources.

Once these first steps have been taken, the analysis of the syntatic patterns is ready to begin. The first step is to code each lexical unit as a part of speech. The code one uses is up to the individual, but we found single letter codes to be the easiest to work with. We found two methods of coding to be useful. The first is to have the sentences punched onto cards and numbered or printed out and numbered. Then, the cards or print-out is coded and cards made up that contain the codes are inserted into the text. The machine is then commanded to list out the text with codes. Another method, was to tag the sentences as they were being printed on the MTST machine. Then all one has to do is command the retrieval system

to bring them forward when this step is arrived at. Of course, this latter method implies that no mistakes were made on the MTST original.

We found the two methods of tagging sentences discussed here to be useful. However, it is clearly easier to sort, correct and edit these data when they are prepared in punch-card form; one card for each sentence and a corresponding card for its speech-part tags.

Although it is possible to attempt to tag all sentences as derived from the media, discussion and sentence elicitation sources, our findings indicate that only the sentence elicitation source can be expected to contain regular and pedagogically useful sentence patterns. Spontaneous speech as recorded from either media or discussion sources exhibit aberrant, run-on, ungrammatical and disfluent constructions which even after extensive editing, do not display the characteristics most desirable for either instruction or frequency tabulation.

The specific coding systems as employed in the three test languages represent those speech parts which were judged to be linguistically productive. They represent a compromise between what could have been a more detailed subcategorization and that of a system employing subcategories. The data indicate that this compromise produced a coherent set of sentence pattern matches which have high face validity and usefulness. Either finer or broader categorization would be expected to provide either too few or too many pattern matches. For other languages, we would recommend that those parts of speech be selected for which there is common linguistic agreement and common native speaker validity and understanding not to exceed under normal circumstances approximately 15 categorists or fewer than 5.

Once the print-out is obtained, it should be edited and corrections made. Then a list out is made that counts the syntactic patterns and records how that pattern occurred in the texts. Alternate means could be a listing of each code type, an alphabetic sort to determine what type starts a sentence, or any other type of sort the user desires.

Analyzing the word frames is a different process. Once the associates have been gathered, the first step is to base form the data. Thus, the response faded is changed to fade. Again the criteria used is up to the individual. Then an H-Rank is assigned to each word association in base form. Those forms having an H-Rank greater than zero are then correlated with all other word types of greater than zero information in order to find that lower ranking word with greatest similarity of associative distribution.

Figure II (A, B and C) and its accompanying explanatory notes summarizes the specific procedural steps for these data analyses.

Concluding Observations

It is our belief and the substance of our recommendation that the elicitation procedures which we have outlined and which form the basis of our research have strong justification for their assumptions. Further, it would seem that those procedures largely obviate the conceptual abstractions of the data which those who would only analyze printed text are required to make in order to satisfy the assumptions of the model they employ. We require only the assumption that the speaker of a language will "spew" his vocabulary items by frequency priority. In order to select those items which have greatest utility over as wide a linguistic context as possible, we conceptually abstract the notion of high information over stimulus environments within a form class frame. This is the equivalent of the abstraction which other researchers have made with respect to content sources. It differs in that in our research we have defined informational uncertainty in terms of differences across the speakers of a defined subject population rather than across the authors of differing content texts.

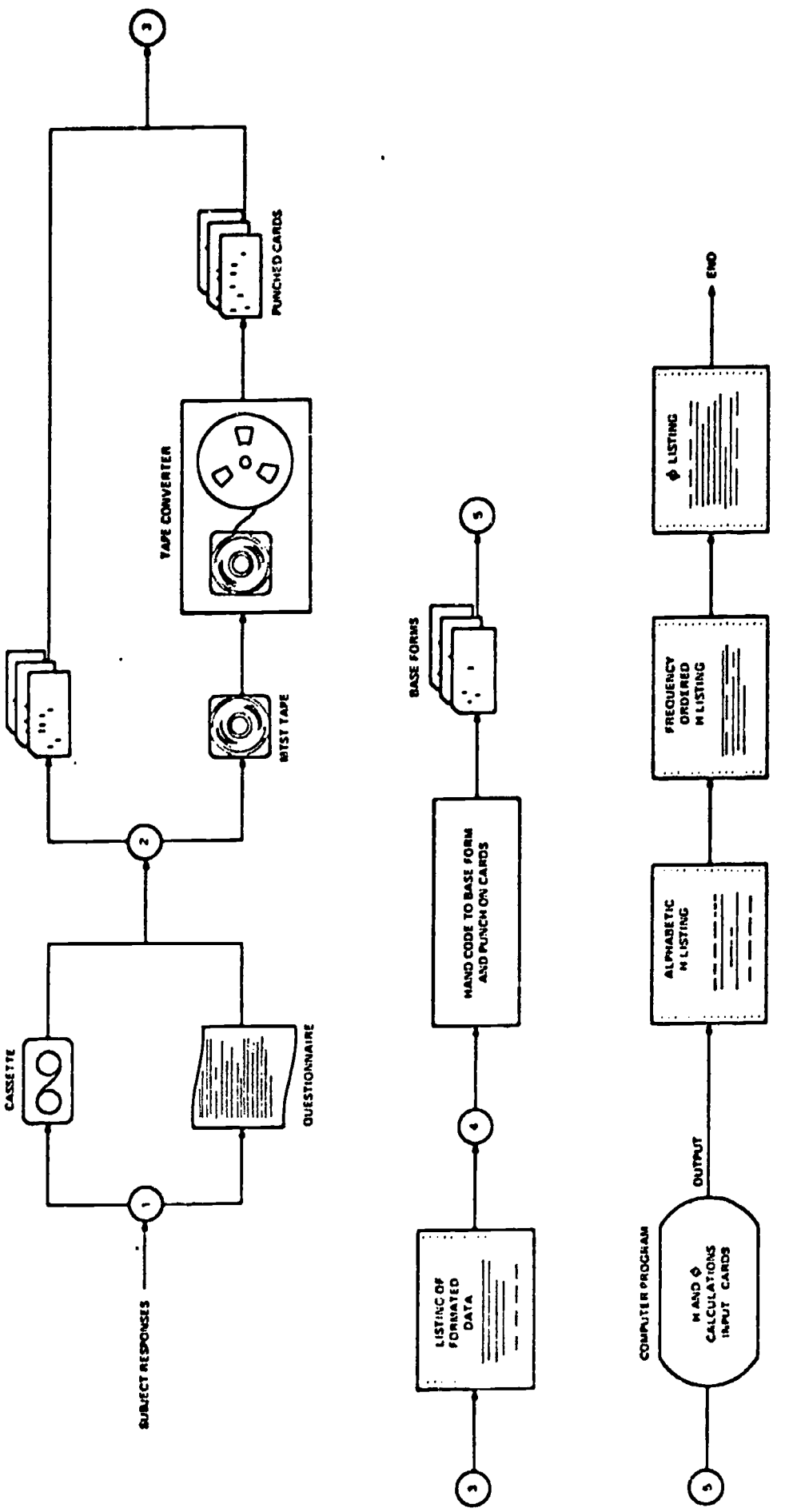

FIGURE IIA. ELICITATION FRAMES

Figure IIa References

Note 1.

Subject responses are collected and recorded for each of the four frames: nouns, adjectives, transitive verbs, and intransitive verbs. As mentioned above each method of recording data has its own potential hazards.

a. Transcribing from cassettes requires special technical help who not only understand the language but can listen, spell correctly and type accurately, all at a reasonable speed. Typing or punching from hard copy does not necessarily require language proficiency.

b. Filling out a printed questionnaire may create a filing problem and detract from the desired spontaneity but it does provide an original document which is much easier to reference and can best be checked for missing responses, even missing pages, and the correct order of responses. When designing the form, each stimulus listed on the questionnaire should be numbered and this index then included in the transcribing process. In addition, each set of forms and each page should be uniquely identified so that responses from different subjects or frames are correctly associated. All documents generated during this process should be saved until the final editing is finished.

Note 2.

The two options for preparing data for a computer must each be carefully monitored. Again it is important to correctly identify each response as to frame number, subject designator and stimulus number.

a. If punched cards are created directly a single format must be defined and adhered to rigorously, especially if more than one key-puncher is involved.

b. If MTST tapes are to be used, the procedure becomes a little more complicated. Again a precise set of rules must be followed and each correction cycle should be proofread by someone other than the typist. Some intermediate identification should be included to denote a break at the beginning of each frame and each set of subject responses. Special symbols such as **, $$, or // are easy to spot and could be followed by the actual index. In the next step the scan program can then recognize these sentinels and produce a sequential listing with appropriate headings. When the final tape is produced it is then converted to 9-track computer tape (see footnote), reformatted to the specified card image and punched.

Note 3.

Once the punched cards are available, the next step is to get a listing for each frame. At this point, a final check should be made for spelling errors, missing responses (represented by BLNX, or some other unique "word"), and the correct sequence numbers for each subject.

Note 4.

Coding back to the base form is now done by hand. Again certain conventions should be established, and clearly explained to each person doing the coding. Based forms should be written on the listing opposite the response and then the corresponding cards pulled out of the deck and the base form

punched starting in column 45. If the cards have subject and stimulus punched in the correct columns it is not necessary to refile these cards. Prudence, however, dictates otherwise.

Note 5.

The present version of the computer program for H and PHI assumes cards are in order of response by subject (on tape in card format). Comment cards in the program deck indicate the input statement and format number. Output listings for each frame consists of:

a. Alphabetic summary of all words encountered (base form, or by default, original response) including the value of H_j, for each word j; N_j, total number of responses; and D_j, number of different subjects using word j.

b. Frequency-ordered summary of reduced dictionary for those words with $H_k > 0$.

c. Highest Q_k correlation for the pair of words (W_k, W_m) found by calculating Q_m for all $m > k$.

Footnote - Tape Converter

The hardware used for type conversion reads the entire MTST tape and converts all information to EBCDIC, 8-bit codes. Special characters are: E0 for carriage return, E1 for backspace, and 00 for word fill after a carriage return. All other numbers, letters, and punctuation marks are standard EBCDIC. All letters will appear as upper case. One page becomes a variable length record on tape so it is important to know the exact number of pages typed on each MTST tape. The remainder of the data can be ignored by skipping to an EOF.

A simple, but hand-tailored, FORTRAN program was written for each language, and in some cases for each frame, another good reason to standardize on data format. The program basically must DECODE 4 characters per computer word and scan each character of the record. Depending on conventions established there may be a need to eliminate extraneous punctuation, consecutive spaces, or in the case of Japanese, to recognize two or more words as a single response.

BEST COPY AVAILABLE

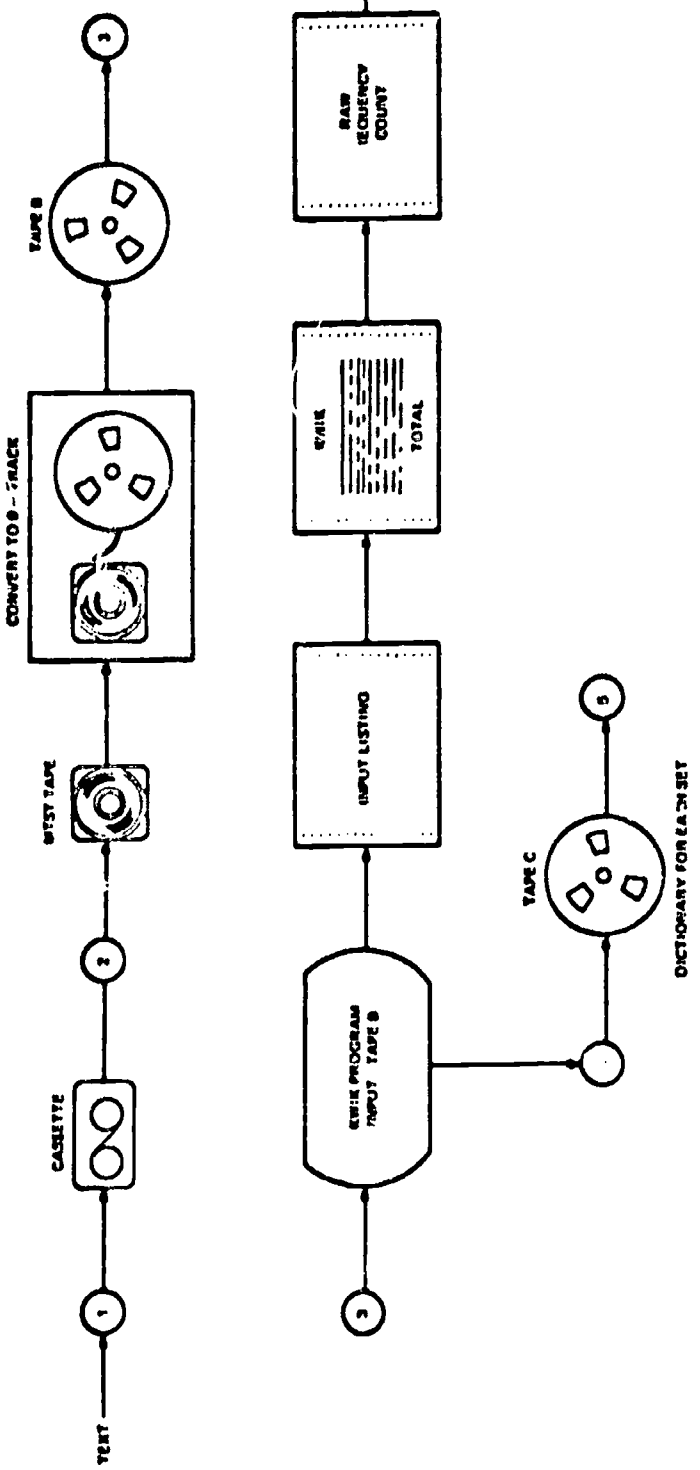

FIGURE IIB. TEXT ANALYSIS

Figure IIb References

Note 1.

Each set of data, discussion, sentence elicitation, and media (TV or other) are recorded. Previous comments about identification also apply here with the additional reminder that the sentences are to be used later in the breakdown of parts of speech.

Note 2.

Transcribing and editing MTST tapes again require a certain amount of control. For example, margins on the typewriter should be set for $\leq$ 80 columns. The editing program will also need rules for punctuation, hyphens, commas, and pronunciation marks. Tape A containing sentence elicitations should be typed double-spaced, and the typist alerted to the second phase when the hand coding of sentence patterns must be inserted.

Note 3.

The program KWIK is a FORTRAN program that could be used as a substitute for the General Inquirer routines. It is limited in the amount of text it can handle but with addition of tape storage could easily be expanded. It reads in data records from Tape B, lists them and prints the raw frequency counts. Part II will produce a KWIC-type list for specified words, rather then the entire dictionary which produces volumes of output. The return to step 1 indicates that this procedure is repeated for each of the three sets of data.

Note 4.

Tape C should save the words and frequency counts form step 3, one set per file, to be used in the final dictionary merge.

Note 5.

A system sort can now be used to combine the three files in alphabetical order, list and punch a complete dictionary. Although cumbersome, punched cards seem to be the best way to handle the variety of combinations and changes involved in coding back to base form. The rationale for coding can then be fairly flexible.

Card Format

Col. 1-28		Word
38-41		Frequency count
79-80		Code to distinguish language and source, such as JS to refer to Japanese Sentences

Note 6.

Hand coding to base form will probably require more than one pass since words such as <u>can</u> and <u>could</u> will not necessarily be adjacent in the deck and much card shuffling will result. In order to produce a uniform listing of base forms and each variant it was necessary to punch and insert the base form card in the correct place and also put in sentinel cards (X in column 1) to separate each set of variants whether one or several.

If the base form already appears in the original deck (columns 31-34 are non-zero) only the X-cards are needed. If a base form card is needed, punch the word starting in column 1 and insert just ahead of the variants. In this case the X-card is not necessary.

Example

```
.
.
X
GO          3    GO will be used as base form
```

```
GOES      6
GONE      4
X                 Optional
HEAR              Hand punched
HEARD     2
HEARS     7
X
STOP      3)
X          )      Base form with no variants
TRAVEL    9)
X
.
.
.
```

Note 7.

A simple FORTRAN Program reads in the data cards, checks for X in column 1, and lists base form indicated by *'s with total count, followed by variants and their counts. X is an arbitrary sentinel, and could have been * or some other character, depending on the language.

BEST COPY AVAILABLE

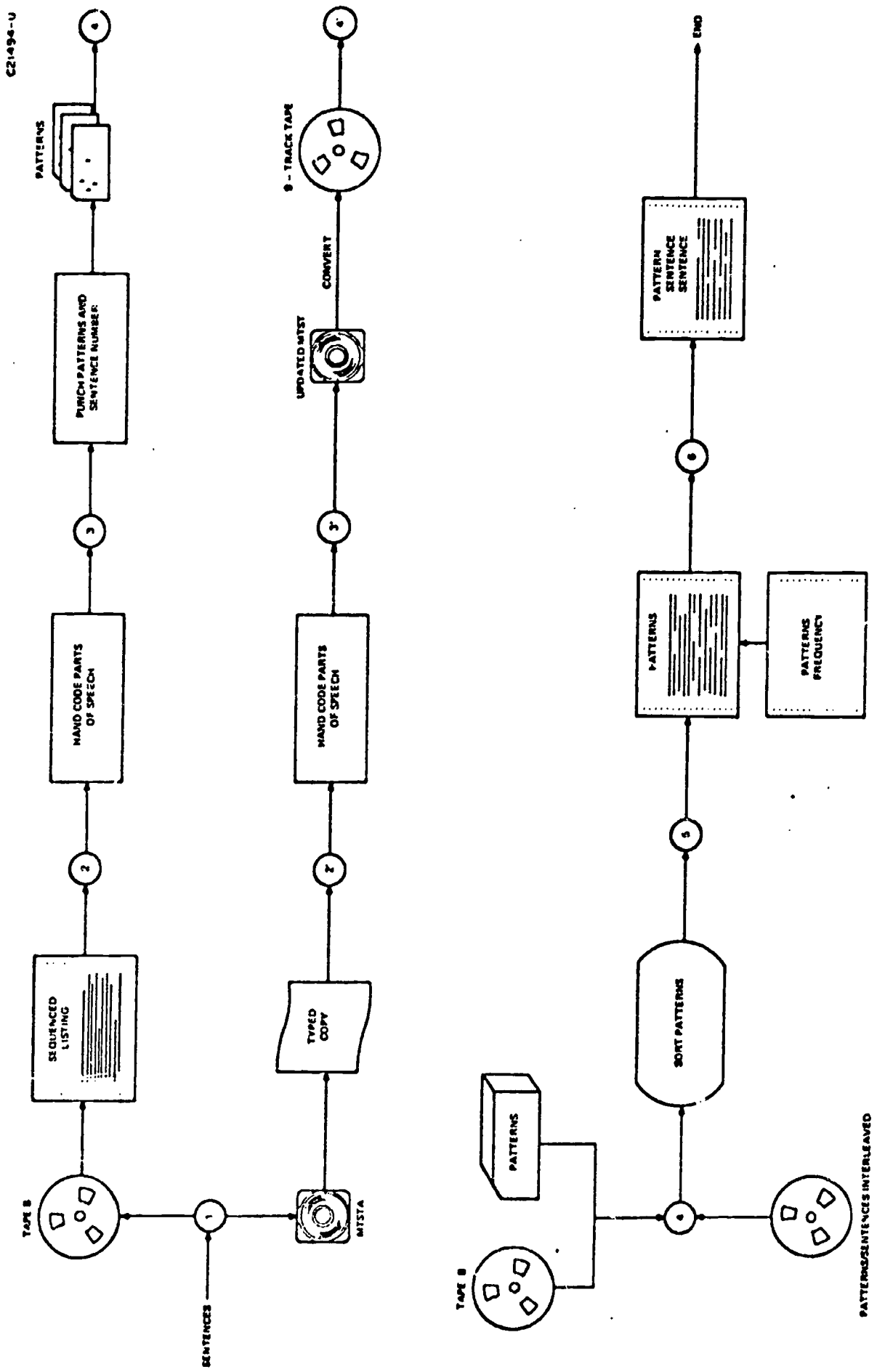

FIGURE IIC. SENTENCE PATTERNS

Figure IIc References

Depending on the storage media, magentic computer tape or MTST tape, the parallel paths perform the same function and any special remarks appear under paragraphs a and b.

Note 1.

First list the elicited sentences, preferably double-spaced.

a. Sentences on tape should be logical records of 80 characters each. On the listing each line or sentence should be numbered consecutively leaving space above each print line for the codes for parts of speech. It is important to realize that some sentences may be longer than one logical record. If each line is sequenced, then the codes punched in the following step should have consistent numbers, so that there is a one-to-one correspondence between tape records and punched cards. It also means the input processor in step 4 must find a punctuation mark at the end of each sentence such as period, question mark or exclamation mark, and the patterns generated by hand must always end in the corresponding letter code.

b. A printed listing from MTST A is already available but an additional copy could be made for use as a work sheet in step 3. Since the pattern will be inserted and copied to a new tape rather than to punched cards a sequence number is not pertinent at this point.

Note 2.

Coding each sentence for parts of speech is done by writing the letter code above each word, and ending each pattern with the code for sentence

type. The input program for cards or tapes will ignore spaces so the typist can just type the string of letters on a line, or card.

Note 3.

a. It is recommended that sentence numbers be included on the punched cards. In the next step the program can then check order of cards and match with the sentence record from tape.

b. Although the chart shows an extra step to convert MTST tape to 9-track tape at this point, the hand coding could have been done back in step 2 in Figure IIb. The scan program could then have skipped every other record on Tape B and used just the sentences for creating the dictionary.

Note 4.

Two read subroutines for this program were written and, depending on the choice of input, one or the other was compiled at run time.

a. For patterns on cards, sentences on tape, the subroutine simply read a card into core, got the next logical record from tape, combined the data and returned the array to the main program.

b. Interleaved pattern/sentence on tape just required reading two logical records into the same array as above.

Note 5.

The sort of pattern codes is first done on increasing length of patterns, with a subsort on alphabetical order. This order seemed to be more useful than a straight alpha sort. The second listing is a frequency-ordered printout.

Note 6.

The final summary of sentences with the same code pattern was done separately and, due to memory restrictions, a system sort was used to create a tape with the patterns arranged in increasing length. This output format makes the first listing in step 5 somewhat redundant, but it provides a more compact listing for easy reference.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)

Syracuse University Research Corporation
Merrill Lane, University Heights
Syracuse, New York 13210

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3. REPORT TITLE

Manual for the Development of Language Frequency Counts.

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Special Report 1 July 1972 - 30 June 1973

5. AUTHOR(S) (First name, middle initial, last name)

Murray S. Miron, Charles C. Pratt

6. REPORT DATE	7a. TOTAL NO OF PAGES	7b. NO OF REFS
June 1973	44	0

8a. CONTRACT OR GRANT NO

DAAG-05-72-C-0574

b. PROJECT NO

c.

d

9a. ORIGINATOR'S REPORT NUMBER(S)

SURC TR 73-235

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

None

10. DISTRIBUTION STATEMENT

Approved for public release, distribution unlimited

11. SUPPLEMENTARY NOTES

None

12. SPONSORING MILITARY ACTIVITY

Defense Language Institute

13. ABSTRACT

As part of a continuing project of language analysis, SURC presents its final manual under the terms of the above mencioned contract. This manual is an explanation of the procedures used to collect and analyse data for this project. After explaining the theory and application of the methodology, the manual discusses specific problems encountered in the design, administration and analysis of the language data collected.

59

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6801

UNCLASSIFIED
Security Classification

KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Structural Analysis						
Languages						
Mathematical Linguistics						
Vocabulary						
Language Research						
Descriptive Linguistics						
Contrastive Linguistics						
Etymology						