PROBLEMS OF VOCABULARY FREQUENCY AND DISTRIBUTION

by

William E. Bull

PART I: Introduction: I assumed in preparing this report that this group would be more interested in conclusions and operational facts than in the procedures by which such information was obtained. To save valuable time for discussion, I shall make a few introductory and rather categorical statements especially pertinent to frequency problems in linguistics and to mechanical translation. If I sound dogmatic, the impression should be attributed to haste rather than intention.

I shall begin by exposing four major fallacies which are current in most discussions of word frequencies:

(I) The traditional vocabulary frequency studies, with which you are all familiar, are not primarily linguistic investigations. Neither the number of nouns in the Oxford Dictionary nor the frequency with which any English noun is used are basic linguistic facts. The total number of English nouns is a manifestation of the technological and cultural advancement of speakers of English (the language would be neither more English nor less English with an increase or decrease in the number of nouns), and the frequency of a noun like aspirin, for example, is simply a reflection of the headache rate of these speakers. Frequency studies of vocabulary are, consequently, not primarily language studies; they are investigations of human activities.

Our conceptualization of the entire frequency problem is one thing if we ask "What is the frequency of 'vector' in the English language?" (a false linguistic frame of reference), and something quite different if we ask "What proportion of the total population, at what time intervals, has use for the word 'vector'?" Man if he talks at all, always talks about something specific and what word counters are trying 'to find out is what he talks about most, that is, how he distributes his time among all the possible things he might talk about.

This leads us to the second fallacy.

(II) If we are actually investigating, in frequency counts, the specific verbal activities of real people every utterance has space-time coordinates, that is, every speaker talks somewhere at some time. (Printed material is a fossil of this activity and becomes, as a result, ambiguous in space-time.) Now space and time, as elements of objective reality, determine human activity and, consequently, the frequency of word usage. The frequency potential of all words, then, depends upon the distribution of population in space and time. For example, the frequency of the word "rain" is undoubtedly much higher here in New England than in Southern California, first, because there are more people to use the word and, second, because it rains here more often. The total frequency of "rain" for these two regions represents neither area, no reality, and is obviously not a significant linguistic fact.

There exists, if this principle is extended, no uniform vocabulary frequency potential for the language and an average is meaningless for any specific purpose.

(III) There does not exist, nor can there be devised, a scientific method of sampling which will reveal anything reliable about word frequencies in a language as a whole. Actual speech (writing, etc.) has a linear structure. What I am saying now is coming at you word by word, serially, on a time line. A number of random segments of such linear speech cannot be welded together into a composite line which will represent any reality. A set of such examples is not even a satisfactory report on the specific material sampled. On an absurdly simple level, a linear weld of this kind produces something like the following three piece composite: "The special significance of vector analysis—in all Congregational Church socials—causes most hens to produce twice the normal number of eggs."

The distribution of lexical items along this linear compound cannot possibly provide useful information about any extended segment of each compounding sample. Distribution and frequency have meaning only in terms of a homogeneous whole which is, theoretically, a non-existent entity in actual speech.

(IV) The 80 per cent fallacy deserves special attention. It has been demonstrated by numerous

word counts that a few hundred words make up some 80 per cent of all the running words found in the counts. It has been concluded, as a result, that you can say almost everything you want to say with a very small vocabulary. It has even been said that the average American uses only some 500 words per day.

Both the facts and the conclusions drawn from these counts are non-scientific and relatively meaningless. Let us take a simple example. James Joyce's Ulysses has 260,430 running words. Approximately 1000 words make up 80 per cent of this total. There are, however, 29,899 different words in the novel. Consequently, while 1000 words take care of 80 per cent by volume of what Joyce wrote, they actually represent, if we assume that he intended every different word to be meaningful, only 3.3 per cent of what he said. The word counters have been, obviously, misled into the belief that quantity and quality are identical. By such logic we should have to contend that once we have bought the nails we practically own a house since, after all, there are more nails in a house than anything else.

Preliminary conclusions relevant to MT: There exists no scientific method of establishing a limited vocabulary which will translate any predictable percentage of the content (not the volume) of heterogeneous material. An all-purpose mechanical memory will have to contain something approaching the total available vocabulary of both the foreign language and the target language. In order to cover most semantic variations several million of items would be needed. At the present time we have no machine which can manage such a number at a profitable speed.

PART II: A statistical analysis of arbitrarily selected samples (segments) of various types of discourse does reveal a number of facts which are important to problems of mechanical translation.

I should like now to show you a number of slides which represent the results of an analysis of 60 samples of about 500 words each taken from contemporary Spanish.

(A) Internal ratios. Language as a structural system is statistically closed, that is, any finite sample will exhibit a ratio pattern of the various parts of speech. Some notion of the volume of each part of speech will be essential to the spatial arrangement of vocabulary items in a mechanical memory, that is, the contact rate of each part of speech will determine whether its vocabulary should be in the high or slow speed memory of any multiple speed machine.

Slide 1: parts of speech in which form and function are identical; not all of total vocabulary covered. Observations:

(a) ratios are not determined by content, that is, subject matter, but by types of discourse: dialogue, expository prose, narration, description, etc.

(b) the frequency of specific items in any part of speech is strongly conditioned by the total ratio of the part of speech and the total available items of that part of speech in the lexicon of the language. Although nouns make up the largest percentage of items in most of the samples, the fact that some 200,000 are potentially available means that no single noun can build up a high frequency. This suggests that even when dealing with highly specialized topics that no predictable limit can be established for the content-bearing vocabulary simply because the low incidence of individual items would require an exhaustive search which is highly impractical.

(c) three major patterns are to be observed: noun, definite article, and preposition behave alike. Pronouns, verbs, and adverbs are in reverse complementary distribution. The conjunctions, indefinite article, and adjectives are independent of the ratios of the contrasting constellations.

Slide 2: all major parts of speech, items classified entirely by function. Conjunction is now only individualistic class. Two major constellations which reveal patterns determined by types of discourse.

General conclusion: Treatment of vocabulary will probably be most satisfactory for mechanical translation if dealt with in terms of types of discourse.

(B) Frequency ratios: The speed and efficiency of mechanical translation depends, vocabulary-wise, on the number of times specific lexical items are repeated in the text. To demonstrate the problem I shall show you three slides giving the ratio distribution of the noun, verb, and adjective (instances where form and function are identical) for the 60 samples. Observe the following points:

(1) Words which appear twice or three times in the samples are relatively stable. The singlettes and high frequency words are in complementary distribution and make up the major portions of most of the samples.

(2) Ratios are not identical for the different parts of speech.

(3) Several factors appear to determine ratios: potential vocabulary of the writer (children's short stories), restricted topic of discourse (physics), type of discourse (dialogue, etc.).

General observations:

(1) The suggestion that words which appear only once may be ignored in MT does not appear profitable. The singlettes make up too large a portion of discourse.

(2) Planned omissions should be analyzed in terms of the part of speech.

(3) High frequency words are dominant and may be presumed to determine the main topic or theme of discourse. The low frequency words, in contrast, are most critical since they represent what the writer is saying about the core topic.

(C) Frequency and distribution of the various parts of speech, by function.

Preliminary observations:

Aside from the general fact that all linguistic data plot out in the form of a parabolic curve, the various parts of speech behave statistically in quite different fashions:

(1) the degree of incline of the parabolic curve is (except for proper names, formulae, etc.) in proportion to the number of lexical items available in the language, that is, the frequency and distribution reach 1 (or the lowest possible minimum) sooner, for example, in the case of adverbs than nouns since there are fewer of the former than the latter. Slides: Articles, Pronouns, Adverb 1, Verb 1, Adjective 1, Noun 1.

By extending this principle to the problem of micro-vocabularies we may predict, with reasonable assurance, that there should be found in all highly technical and restricted fields a high frequency of a few words and an extremely large number of single words in typical samples. This explains, in part at least, why Oswald found that a relatively few nouns cover nearly 90 per cent of the vocabulary in brain surgery in German. Whenever the total available number of lexical items is small, repetition must increase or discourse cannot be sustained.

The facts just demonstrated also provide a general answer to the question of the feasibility of micro-vocabularies and throw some light on the problem of not translating rare words. The less specialized the field, the larger the number of available words, the lower the frequency of the most common and, consequently, the greater the semantic importance of rare words. The present paper, for example, deals with a highly technical field in linguistics but draws upon such a wide and unpredictable vocabulary that no micro-vocabulary based on previous articles on word frequency would provide a satisfactory translation. I shall cite a few critical words (those that cannot be omitted without serious distortion of the sense) to demonstrate the point. Notice, also, that you cannot determine the topic of discourse from this list: coordinate, aspirin, technological, headache, rain, population, linear, weld, slide, church, machine, hen, constellation, singlette, egg, children, physics, core.

A micro-vocabulary appears feasible only if one is dealing with a micro-subject, a field in which the number of objective entities and the number of possible actions are extremely limited. The number of such fields is, probably, insignificant.

(2) The non-content-bearing parts of speech (articles, prepositions, conjunctions, adverb 2, relative, demonstrative, and indefinite pronouns) exhibit an extremely high degree of correlation between frequency and distribution. 3 slides: articles, pronouns, adverbs.

Within the same part of speech the correlation between frequency and distribution increases as the referential value decreases. 2 slides: Adjective 1, Adjective 2, 3, etc. The possessive, demonstrative, pronoun adjectives show a much higher correlation than adjectives which refer to the nature of reality. These adjectives are, so to speak, indifferent to the subject of discourse while fat, lateral, huge, etc. are restricted to certain subjects.

High correlation appears in other parts of speech only at the tail of the parabola where it is, for obvious reasons, insignificant, and at the head where it indicates that all high frequency words are

non-specific, operational, and non-indicative of the subject of discourse. 2 slides: Noun 1, Verb 1.

This principle may be expected to show some significant variations when applied to restricted fields, what I have called micro-subjects. In specific examples of discourse high frequency content-bearing words commonly outnumber the parallel high frequency words (of the same part of speech) in the general language. They define the main topic, for example, the way frequency, word, speech, distribution, noun, etc. define the subject of this paper, but, curiously enough, these words do not tell us what is being said about the topic. This fact establishes a principle which cannot be over-stressed in dealing with vocabulary problems in mechanical translation, namely, that the rarer words carry the significant and critical message in most extended communications. The tail of the parabola is what makes one article on brain surgery different from another.

(3) The middle range of the parabola for all content-bearing words exhibits a low correlation between frequency and distribution. The actual degree of divergence depends upon the general semantic function of the various parts of speech and their potential descriptive range or combinatory power. Modifiers have a wider distribution potential than head words, and verbs more than nouns. 4 slides: Noun 1, Adjective 1, Verb l, Adverb 1. This confirms a principle which has been much debated in structural linguistics, namely, that the noun is the core word in communication.

We have now established a hierarchy of the parts of speech which should provide an operational principle in the preparation of vocabulary lists for a machine memory. The value of micro-vocabularies depends directly upon the function of the part of speech and the total number of available words. It may be predicted that as the degree of correlation between frequency and distribution increases the larger the percentage of the total available vocabulary for such parts of speech which will have to be included in any micro-vocabulary. Thus, to pick up Oswald's problem again, a micro-vocabulary for brain surgery will require less than 1 per cent of the available nouns in German but probably100 per cent of the secondary adverbs. If it seem valuable, further research could presumably define rather accurately the percentage of the total vocabulary for each part of speech normally required to carry on discourse in any well-defined field.

Conclusion: The present data also points to another division of vocabulary which I should like to discuss in the way of conclusion. Vocabulary appears to fall into three major classes:

(1) words which are primarily indifferent to the subject of discourse and which will be indispensable for any type of translation.

(2) words which define the theme or topic of discourse and which cluster in somewhat predictable constellations and appear with especially high frequency within specialized fields.

(3) words which provide the running commentaries upon the theme or topic of discourse and which appear with very low frequencies and do not tend to cluster. These words make up the tail of the parabola and are not amenable to precise prediction since they represent the potential associations which every speaker may establish between his topic and the infinite universe. They are the bridge between the closed system of structural vocabulary, the restricted vocabulary of the specialties, and the cosmic reality within which the language system and the speciality exist and operate. To presuppose that such a vocabulary can be defined and limited requires the assumption that knowledge has reached its maximum potential and that man will discover no new and hitherto unknown associations between departments of knowledge.

The limitations of machine translation which we must face are, vocabulary-wise, the inadequacy of a closed and rigid system operating as the medium of translation within an ever-expanding, open continuum.

Special mention must be made of the contributions of Harry Huskey and Charles Africa to the preparation of this paper. Mr.Huskey has made many valuable suggestions and his staff has done the graphs and slides. Mr.Africa has done almost all of the actual counting and the preparation of the raw data.