Henk Siliakus

TO LIST, OR NOT TO LIST? COMPUTER-AIDED WORD LISTS FOR THE HUMANITIES AND SOCIAL SCIENCES

Introduction

In his introductory essay in the Final Circular for this conference, Reinhard Hartmann suggested that the 'scenarios' of lexicographers are bound to influence their work. I agree wholeheartedly; I would never have entered the field of lexicography were it not for the fact that urgent problems had to be solved. I refer here to the depressing state of German studies in Australia in the late 1950s. Fewer and fewer students enrolled in our first-year courses, and we could foresee the day when we would decide between us who could teach the student the following week. After long negotiations, Faculty allowed us to organize beginners' courses on the condition that they would reach the linguistic standards required at the Matriculation Examination and cover all the topics in Literature and Civilization of German I. Since I had previous experience in teaching beginners, I was asked to design a suitable course.

It is possible to 'cover' the grammar and syntax of German in one academic year. One can deal with the mechanics of the passive in twenty minutes. The students soon see what goes where, and nod their heads: they have 'got' it. That they need copious practice goes without saying.

It is much harder to present, in a systematic way, some 2,500 lexical items in what amounts to no more than 25 weeks of teaching. Obviously, a very careful selection must be made to ensure maximum return for the effort spent. In 1959, 'useful' still meant 'frequent', and so we turned to Kaeding's HÄUFIGKEITSWÖRTERBUCH only to find that in his list the words national-liberal and Butter are about equally frequent. No wonder, for his sample was mainly drawn from parliamentary reports. After all, Kaeding's 'scenario' was to provide a frequency list for stenographers reporting in parliament. However, such a list is of little use in the school classroom.

So we did some counting of our own, mostly from the prescribed texts. After 50,000 words the money ran out; we then developed a formula which took account of our own results, of Kaeding, and of some other basic lists as well as derivational potential. The final list was known as the Adelaide list of 1000 basic words. It turned out to be wildly unreliable in the lowest quarter, and yet high-school teachers praised it, for its text coverage was consistently over 70 per cent.

Bigger samples, smaller error margins

How unreliable it really was we learned after reading Frumkina's (1964) article in IRAL. It was clear that we would

never have the time or money to count a sample big enough for a 1000-word list with acceptable error margins. But in the same year our University acquired a powerful computer. And something else happened to stimulate us into a new start.

In the mid-1960s we had a big increase in the number of M.A. and Ph.D. candidates. Many of them had taken their first degree without knowing any modern language. But now their supervisors insisted on a reading knowledge, and many students turned to us for help. Many of these came from the Musicology Department, so we decided to analyze a sample of 100,000 running words from that discipline.

A sample was punched and transferred to magnetic tape. Then a store of our 1000 basic words, enlarged to include all their paradigmatic potential, was prepared. A comparison of our music corpus with the basic store provided a 'basic coverage' (for this sample just over 70%) and 'the rest'. The latter included high-frequency special terms, as well as general words of lower frequency. Both the basic list and the 'rest' were produced in two versions: alphabetical and in rank order. It was the rank list of the 'rest' that was then used to list the 500 most frequent occurrences.

However, among the 'rest' were many words that did not need listing, since they could easily be understood even with a very elementary knowledge of German. We excluded internationalisms (Musik), proper nouns (Beethoven), dates and numerals (1770), and compounds the constituents of which belonged to the basic vocabulary (Kirchenmusik). These four categories combined covered about half of the rest, i.e. 15 per cent. The 500 listed items cover another 10 per cent, so that in fact only about 5 per cent need to be looked up in the dictionary.

Fig. 1 The breakdown of a typical 100,000 word sample

SAMPLE :	100,000	running words
of which	70,000	are basic
leaving	30,000	of which
	15,000	need not be glossed
This leaves	15,000	of which
	10,000	are covered by the special list
leaving	5,000	to be looked up

The 500 words are provided with some grammatical information, English equivalent(s) and a context. A list of cognates, showing the extent to which both German and English make use of Greek and Latin roots is also included.

After Musicology, we produced, over a number of years, so-called GERMAN WORD LISTS: Literary Criticism, Geography, History, Theology, Sociology, Linguistics and Fine Arts.

In search of the common elements

Whenever we tackled a new discipline we noticed that there was a certain overlap of items, and we became convinced that a large number of terms are common to all the disciplines taught in the Faculty of Arts and Social Sciences. We decided to test this hypothesis, and do a simple range analysis.

The work of such authors as Muller (1972), Carroll (1971) and Corder (1973) had helped to sharpen our awareness of the statistical properties of language, but it was Juilland's FREQUENCY DICTIONARY OF SPANISH WORDS that encouraged us to embark on a distribution analysis. This work is well-known and need not be discussed here. What follows is a report on the results of our analysis.

We took from each of six disciplines four batches of 20,000 running words and added another batch of 20,000 taken from the vocabulary in Wahrig's DEUTSCHES WÖRTERBUCH. In doing this we followed a hint from Mackey and Savard (1967) who argued that a word which is often used to define other words is obviously useful.

This gave us 25 batches of 20,000 running words each. This half million words consisted of about 57,000 different words, a type-token ratio of about 1 : 9. Of these types, about 37,000 were hapax legomena, which is somewhat higher than is normally assumed to be the case. Boot (1975) gives 50 per cent. Our number of words with frequency 1 is 37 in 57, or 65 per cent. We expect that this percentage would decrease as the size of the corpus increases.

This means that we had about 20,000 items of frequency 2 or over. Of these we prepared listings according to distribution, usage, range and frequency (Siliakus 1974). Having lost faith in the supremacy of frequency as an index of usefulness, and full of ardour for the newly discovered distribution index, we arranged the four initials to read DURF.

Our distribution (D) factor measures the evenness of a word's distribution throughout the whole sample. Particles occur with almost equal numbers in all our sub-samples and get close to the possible maximum of 100, whereas specialist terms occurring in a few sub-samples only have a low score. Our range (R) factor indicates in how many of the sub-samples the word in question occurred. In our case, having 25 sub-samples, the R is expressed as X/25. Our frequency (F) factor states the frequency of a word over the entire corpus of 500,000 running words. Finally, our usefulness (U) factor is the product of D and F. This index of usefulness guards against high-frequency specialist terms being included in general lists, because their low D values would counteract their high F values.

Selecting the basic vocabulary

The parameters we set in order to isolate the basic vocabulary were as follows. Having decided that approximately 2500 was a fair figure for the intermediate stages such as Matriculation,

we decided on U ⩾ 10 and D ⩾ 50 per cent. As a result, 2672 words emerged. Of these 2672 there were 864 which had a U ⩾ 30 and D ⩾ 75 per cent. These were starred in the list; they seem to be a fair minimum for beginners' classes.

We decided on setting parameters for both U and D after some trials with U alone. These yielded some examples of words that were obviously specialist terms. Imagine a highly specialized term occurring in two sub-samples only with a total frequency of 55 and a distribution factor of only 20 per cent. The U, being D x F, would still come to 11, and hence be included in our list with U ⩾ 10 as a minimum. We felt that both F and D should be sizeable factors, and hence we set minima for both.

The lean years of the late 1970s brought our work to a virtual stop, but in 1981 we obtained from a colleague a 70,000 word sample taken from Linguistics texts and after having made it up to our usual 100,000 running words we produced our Linguistics volume (No.8). And in the next year, a grant from the Australian Humanities Grants Commission enabled us to produce Number 9 in the series, the vocabulary of the Fine Arts (Kunstgeschichte).

In discussions with colleagues it had been put to us that we should exclude from our lists all those 2672 items from our DURF volume. After all, they argued, these were the basic words which would be met in any kind of expository prose, and should be known. Our own argument that not knowing a 'general' word, such as Entwicklung, would be just as much a stumbling block as not knowing Tonnengewölbe was accepted. But, it was said, students could always consult a dictionary for Entwicklung, and it seemed sensible to limit the list to specific terms. Hence, in Number 9, DURF words were excluded. This meant, amongst other things, that we had to include lower frequencies than we had done hitherto. No doubt this was also caused by the fact that our Fine Arts sample was really composed of three different disciplines: painting, sculpture, and architecture. There is not much overlap in the vocabulary of these three.

The contexts that we provide on the right hand pages are produced by a KWIC (Key-Word-in-Context) routine. Sometimes these have to be adapted somewhat. KWIC indices are also used to solve problems of homonymity.

The work described here has enabled us to produce a general list, as well as a number of special ones. Perhaps this goes some way towards answering Reinhard Hartmann's question about the scenario of the lexicographer. In our case, a real need existed, and all our work has been directed towards an attempt to meet those needs.

References

Boot, M. (1975) "Frekwentie en spreiding, wat doen we ermee?" Levende Talen 311: 131-140

Carroll, J. (1971) "Current issues in psycholinguistics and second language teaching" TESOL Quarterly 5: 101-114

Corder, S.P. (1973) Introducing Applied Linguistics. Harmondsworth: Penguin Books

Frumkina, R. (1964) "Allgemeine Probleme der Häufigkeitswörterbücher" IRAL 2: 235-247

Mackey, W. and Savard, J. (1967) "The indices of coverage: a new dimension in lexicometrics" IRAL 5: 71-121

Muller, Ch. (1972) Einführung in die Sprachstatistik. München: Hueber

Siliakus, H. (1974) Distribution, Usage, Range and Frequency (GERMAN WORD LISTS No. 7). Adelaide University