R E P O R T R E S U M E S

ED 014 255 FL 000 573

COMPUTER-AIDED WORD RESEARCH.

BY- SILIAKUS, H.J.
AUSTRALIAN FED. OF MODERN LANGUAGE TEACHERS ASSNS.

PUB DATE JUL 67

EDRS PRICE MF-$0.25 HC-$0.24 4P.

DESCRIPTORS- *COMPUTERS, *LANGUAGE RESEARCH, *GERMAN, *WORD FREQUENCY, *WORD LISTS, COMPUTER ORIENTED PROGRAMS, READING LEVEL, RESEARCH PROJECTS, VOCABULARY DEVELOPMENT, LEXICOGRAPHY, READABILITY, COLLEGE LANGUAGE PROGRAMS, UNIVERSITY OF ADELAIDE, AUSTRALIA,

IN PREPARATION FOR THE DEVELOPMENT OF A GENERAL FREQUENCY WORD LIST IN GERMAN DESIGNED TO MEET THE NEEDS OF THE INTERMEDIATE AND ADVANCED LEVELS OF READING IN THE GERMAN CURRICULUM, A COMPUTER-BASED WORD COUNT WAS BEGUN IN AUSTRALIA'S UNIVERSITY OF ADELAIDE. USING MAGNETIC TAPES CONTAINING (1) A TEXT OF OVER 100,000 RUNNING WORDS, (2) 1,000 MOST USEFUL WORDS, WITH ALL THEIR MORPHOLOGICAL POSSIBILITIES, TAKEN FROM A PREVIOUSLY DEVELOPED ELEMENTARY-LEVEL LIST, AND (3) 500 SPECIALIZED MUSIC TERMS, THE COMPUTER PROCEDURES DETERMINED THAT JUST OVER 30 PERCENT OF THE ORIGINAL SAMPLE TEXT HAD NOT BEEN COVERED BY THE BASIC OR SPECIALIZED WORD LISTS. THE BULK OF THE REJECTED WORDS PROVED TO BE VERY LOW FREQUENCY ITEMS, BUT THERE WERE MASSES OF HIGH FREQUENCY PROPER NOUNS AND COGNATES FOUND IN THE UNCOVERED PORTIONS OF THE TEXT. THIS ARTICLE APPEARED IN "BABEL," VOLUME 3 (NEW SER.) NUMBER 2, JULY 1967, PAGES 19-21. (AB)

Babel

New Series

Vol. 3 — No. 2

Journal of the Australian Federation of Modern Language Teachers' Associations

Contents:

REGISTERED AT G.P.O., MELBOURNE, FOR TRANSMISSION BY POST AS A PERIODICAL

80c

COMPUTER-AIDED WORD RESEARCH

H. J. SILIAKUS*

IN a previous issue of *Babel* (October, 1964) I reported on a word-frequency project carried out at the University of Adelaide. The main reason for undertaking this work was the fact that almost all existing word lists for German were based, directly or indirectly, on F. W. Kaeding's *Häufigkeitswörterbuch*[1] which is unsuitable for pedagogic purposes. Since then a great deal of original work has been done in the field of lexicology[2], most of it with the help of computers. At Adelaide we have just completed the first stage of a computer-based word count for German. This article reports on its methods and the first tentative findings.

Our existing list of the thousand most useful words[3] has served us well. We use it as a basis for the evaluation of the vocabulary content of elementary readers and it has been helpful when setting examinations. More recently, with the advent of language laboratories, the need for suitable materials has greatly increased, and most of these are produced locally. Here too the list has proved itself a useful aid. It is generally agreed that when the student is practising certain structures the whole of his attention should be focussed on the structure itself. The inclusion of unfamiliar vocabulary would create problems of meaning. We feel that we ought not to distract from the morphological and syntactical aspects by creating extra difficulties. Vocabulary building must of course be carried on side by side with the structural work, but this can be done just as effectively outside the laboratory.

Having used our list regularly over a number of years we are now not altogether confident that the last few hundred words really have a claim to being included. It is quite possible that another word count conducted along exactly the same lines would produce a number of changes in the last quarter of the list. Words at present included might not qualify for inclusion again and be replaced by other words at present not listed. Our original sample was only 50,000 running words and in the lower frequencies the error-margin is bound to be large. It has been argued[4] that one would need a sample at least ten times as large to obtain a fully dependable basic vocabulary. This does not mean, of course, that the last quarter of the Adelaide list is altogether useless, since it contains words well worth teaching. However, we cannot say with any degree of certainty that the word listed in position 950 is more frequently used than that in position 990. In a word list based on another sample these two words might well occur in the reverse order.

We are no longer prepared to count large samples by hand and yet we have felt the need for greater precision. Moreover, we need a word list that goes beyond 1,000 entries, as the bulk of our work is concerned with the post-elementary stages. A 1,000-word basic list can easily achieve a text-coverage of 85-90 per cent of elementary readers. But how much needs to be added to this list to deal effectively with the intermediate and advanced stages of our work?

The following table gives an answer to this question. Stejnfeldt[5], a Russian scholar, found that the expansion of a vocabulary list from 1,300 to 2,000 basic words added very little text coverage.

Sample taken from	Text-coverage obtained with	
	1,300 words	2,000 words
Russian prose	74%	78%
drama	75%	79%
newspapers	73%	78%
poetry	65%	68%

This means that a student of Russian who increases his vocabulary from 1,300 to 2,000 words will increase his text recognition by at most 5 words per hundred. He will still have to use his dictionary twenty-odd times (easily recognized cognates excepted) per hundred words. It is clear that after mastering a basic vocabulary students will have to learn vast amounts of additional words to achieve real reading fluency.

---

*Mr. Siliakus is senior lecturer in German and director of the language laboratory at the University of Adelaide.

Our own findings confirm this in a startling way. One can cover 20% of a German text with a knowledge of only ten words. Twenty words cover 28%, fifty words 40%, one hundred words 46%. If these figures are plotted on graph paper one sees that the curve flattens out very quickly. Hence, the further one goes along the curve, increasingly more vocabulary has to be absorbed in order to increase the text-coverage.

Since we were not prepared to analyse a large text by the old-fashioned method, we used the university's computer[6]. We had over 100,000 words typed out on paper tape and transferred on to magnetic tape. We chose texts from musicological writings. For some time, several colleagues have been unhappy about their students' lack of reading ability. We intend to help them by providing specialized word lists for music, history, geography, psychology and literary criticism. In this way we hope to achieve two things at the same time: firstly, word-lists for several Arts disciplines, and secondly a large sample of at least half a million words, from which we can compile a dependable general list of up to 2,000 words. Present indications are that a list of 2,000 general words and a special subject list of say 500 specialized terms would give a very good text-coverage, probably somewhere near 90%.

In conclusion I should like to describe very briefly the actual method employed. As mentioned above, our first stage was the typing of a text of over 100,000 running words. Then we compiled two dictionaries which were also transferred on to tape. The first consisted of our 1,000 most useful words together with all their morphological possibilities. Our list gives stems only. However, in the text a verb like *geben* may occur as *gebe, gibt, gab, gegeben* etc. We used a code of indices to instruct the machine to charge all those possible occurrences to the stem word, but to keep a tally of the individual frequencies. It was necessary to make similar arrangements for adjectives and nouns. This first store finally came to about 3,300 entries.

Then we made a list of music terms in the same way. We abstracted some 500 words from articles and by listing the instruments of the orchestra etc. These lists were also transferred on to magnetic tape.

The computer then produced first of all a list of all the words in the sample that were covered by our basic 1,000 and their forms. This accounted for 64% of the sample. Then a list of words was printed that were covered by our music dictionary; the coverage was about 5%.

This meant that just over 30% of the original text had not been covered by our two dictionaries. Of course we were anxious to have a look at these rejects. We found that the bulk of them consisted of very low-frequency words (1-3 occurrences in 100,000 words), but that there were masses of high-frequency proper nouns (Beethoven, Beethovens), as well as high-frequency cognates (Komposition, Instrument, Text). A preliminary calculation points to a total coverage of about 85%.

With these results in front of us all sorts of exciting projects can be undertaken. Our first priority is a manageable word-list for music students. The general frequency list will have to wait until another four batches are analysed, and this will take time and money. However, preparations are under way for the analysis of the vocabulary content of literary criticism. Once this has been done we shall know much more about the sort of vocabulary a second-year university student should know if he is to write an essay on the lyrical poems of the Romantic period. It is with such concrete and practical issues in mind that we have embarked on this, our second stage of vocabulary research.

Notes

[1]F. W. Kaeding, *Häufigkeitswörterbuch der deutschen Sprache,* Berlin, 1898.

[2]Two of the best-known institutions where computer-based research in lexicology is carried out are *Centro per l' Automazione dell'Analisi Letteraria* in Gallarate near Milan, where under the direction of Pater Roberto Busa S.J. concordances of medieval philosophers ar· ·ng compiled; and the *Centre d' Etude du ·ılaire Français* in Besançon, headed by Professor B. Quemada. It was at Besançon that the most recent word count of Basic Spoken German was analysed. The findings have been published by Prentice Hall, New Jersey, 1964; J. Alan Pfeffer: *Grunddeutsch. Basic (Spoken) German Word List.* For a general account of modern work in vocabulary research, see article in *Beiträge zur Sprachkunde und Informationsverarbeitung,* Heft 1, p. 33 ff, by Dr. F. de Tollenaere.

[3]*The Adelaide List of the 1,000 most useful words in German;* (ed. H. J. Siliakus); Department of German, University of Adelaide, October 1964.

[4]Rolf-Dietrich Keil, *Einheitliche Methoden in der Lexikometrie,* IRAL III/2, 1965, p. 102 ff.

[5]Table based on IRAL II/4, 1964, pp. 245, 246.

[6]We wish to thank the Computing Centre of the University of Adelaide and especially Mr. K. C. Lee, a Ph.D. student from Malaysia, who designed the computer programme for us.

Bibliography of some basic word lists

French

Cheydleur, F. D. *French Idiom List.* N.Y., 1929.
Henmon, V. A. C. *A French Word Book.* Madison, 1934.
Vander Beke, G. E. *French Word Book.* N.Y., 1924.
Gougenheim, G. *et al. L'élaboration du français élémentaire.* Paris, 1956.

German

Hauch, E. F. *German Idiom List.* N.Y., 1929.
Kaeding, F. W. *Häufigkeitswörterbuch der deutschen Sprache,* Berlin, 1898.
Michea, R. *Vocabulaire allemand progressif.* Paris, 1959.
Morgan, B. Q. *A German Frequency Word Book* (based on Kaeding). N.Y., 1928.
Wadepuhl, M. *Minimum Standard German Vocabulary.* N.Y., 1934.

Spanish

Hoz, V. G. *Vocabulario usual, comun y fundamental.* Madrid, 1952.
Kenniston, H. *Spanish Idiom List.* N.Y., 1929.

Russian

Josselson, H. H. *The Russian Word Count.* Detroit, Mich., 1953.
Schilling, I. *Unser russischer Wortschatz.* Berlin, 1963.
Stejnfeldt, E. A. *Castotnyj slovar' sovremennogo russkogo literaturnogo jazyka.* Tallin, 1963.