An Analysis of Lexical Text Coverage in Contemporary German

Randall L. Jones
Department of Germanic & Slavic Languages
Brigham Young University

One of the many practical applications of corpus studies is the generation of word frequency information, which can provide useful data to language teaching professionals as well as lexicographers. It makes sense that for the teaching of vocabulary in a second language, lexical frequency should play a significant role in the selection of vocabulary to be included in pedagogical materials. Likewise, a dictionary intended for language learning should be based at least in part on words that occur most frequently in the language being studied. As Aston has stated, "Insofar as frequency data from corpora can indicate whether a particular feature is likely to be worth learning, this fact makes it relevant to teachers and learners as well as to syllabus and material designers" (P. 8).

In this arena of corpus-based frequency studies for pedagogical applications, Nation and others have been successful in determining what has come to be known as "text coverage", or the relationship between a given number of high frequency words in a text and the percentage of the text that these words account for. As Nation points out, "... studies of native speakers' vocabulary growth see all words as being of equal value to the learner. Frequency based studies show very strikingly that this is not so, and that some words are more useful than others" (P. 9). He goes on to show that in a specific text of English fiction the most frequent 1000 words in the text accounted for 82.3% of the total text. The second 1000 most frequent words increase the coverage by only 5.1%, i.e. a total of 87.4%. To achieve coverage more than 90% requires 2000-3000 additional words.

This paper will report on the results of a similar study for contemporary German. The corpus from which the frequency information is generated is the BYU/Leipzig Corpus of Contemporary German and contains more than 4 million words, including a spoken component (ca. 1,000,000 words) and a written component (ca. 3,300,000 words). The spoken corpus is based on 400 interviews conducted in 60 communities in Germany, Austria and Switzerland. The written corpus consists of four genres: fiction, journalism, special purpose texts (e.g. operating instructions, advertisements, etc.), and academic prose. Using the RANGE program available from the School of Linguistics and Applied Language Studies at Victoria University of Wellington, New Zealand, calculations are being made on subsets of the BYU/Leipzig corpus. Information will be given for the total corpus, the spoken corpus as well as for each of the four genres of the written corpus. The paper will also discuss some of the difficulties of German frequency studies, e.g. complex inflectional and derivational morphology, polysemy and ambiguity. Applications to the generation of a word frequency dictionary for German will also be discussed.

REFERENCES

Aston, Guy (Ed.) . 2001. Learning with Corpora. Houston: Athelstan.

Nation, I.S.P. 2001. Learning Vocabulary in Another Language. Cambridge:
University Press.