Word frequency dictionary of the written vocabulary of 10- to 16-year-olds

Erzsébet Cs. Czachesz and János Csirik

[1]Department of Education, University of Szeged, H6722, Szeged, Petőfi sgt. 30-34.
[2]Department of Computer Science, University of Szeged, H6720, Szeged, Árpád tér 2.

Keywords: frequency dictionary, child language

The presentation outlines the preparatory work for a frequency vocabulary and highlights the most important findings of the project. The previous child language word frequency dictionary in Hungarian was compiled by János Cser in 1939, as an outcome of his research on children's vocabulary.

Before the appearance and subsequent widespread need for computerised text processing, the compilation of frequency vocabularies usually served educational objectives besides the purposes of basic linguistic research. The first and most important area of their application is *foreign language teaching*. In recent decades, multifunctional frequency dictionaries based on representative linguistic corpora have appeared. Another area of application, also related to education, is the analysis of *readability*. Researchers of reading and reasoning processes also often rely on frequency dictionaries as resources in developing and executing experimental designs.

In the present dictionary project, student texts were collected in a representative survey of written composition. In grades 4, 6 and 8, sub-samples are representative for settlement types and in grade 10, for school stream. Altogether 8,670 student compositions were analysed. Of these, one fourth, 2,170 were randomly selected and entered into a computerised corpus. Data processing was carried out in the SZTE-MTA Research Group on Artificial Intelligence at the Institute of Informatics of the University of Szeged. In the process of text annotation, every word was assigned the appropriate grammatical and morphological labels. The annotation was based on HuMor, a Hungarian linguistic analysis software developed by MorphoLogic Kft, and on the MSD code system developed for European languages.

The dictionary first gives an alphabetical list of all the word frequencies in the whole corpus. Next the most frequently occurring 1,000 words are given, then lists for parts of speech. In the second half of the dictionary, the same structure is followed in presenting data for age based sub-samples. Frequencies for parts of speech are given for all grades. Data are summarised in tables in the last section of the dictionary.