Abs-155

Kirsten Ackermann (Pearson), Douglas Biber (Northern Arizona University), and Bethany Gray (Northern Arizona University)

An Academic Collocation List

This paper presents a frequency list of the most common and pedagogically relevant collocations in written academic English discourse, derived from the written Academic component of the Pearson International Corpus of Academic English (PICAE). The list can be used, for example, in lexicography, test item writing, and EAP material development.

PICAE is a corpus of over 37 million words, comprising a written component (32.5 million words) as well as a spoken one (4.6 million words). The corpus covers American, Australian, British, Canadian and New Zealand English. PICAE was designed with reference to the question, what English does a non-native speaker need in order to be successful in academic settings where English is the main language. Spoken data include lectures, seminars and broadcasts. Written data comprise textbooks and journal articles reflecting a broad range of academic disciplines, as well as university, student and alumni journals, and study and career information.

For this project only the written Academic component of the Pearson International Corpus of Academic English (PICAE) was used, which comprises over 25 million words covering 28 major academic subjects.

For the compilation of such a collocation list, the concordance program MonoConc was used to first obtain a simple list of words occurring more than 12 times in the corpus.

This list was processed by a computer program written in Pascal to index each word, and the sub-dimensions for each word using variables like frequency, number of texts that the word has occurred in, and frequencies in each of the four academic disciplines (humanities, social science, natural and formal science, professions and applied science).

The output of this program was a reference list of important academic words, which occurred at least 5 times per million words and in at least 5 different texts. A stop- list was created containing the frequent function words that express purely grammatical meaning. This list was used by the collocations and bundles programs written in Perl to exclude those words from subsequent analysis.

The collocation program itself used the reference list of important academic words as input, again using a large multi-dimensional hash or database table. As each potential collocate is located in the texts, the program added the entry to the data structure if it did not exist, or increment frequency and distributional counts if it already existed.

This paper will explain the motivation for the academic collocation project. It will shortly introduce the Pearson International Corpus of Academic English. The paper will then look at the methodology applied and problems encountered. Lastly it will discuss potential usages of the Academic Collocation List.