A Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria

Kenji Kita,† Yasuhiko Kato,†† Takashi Omoto† and Yoneo Yano†

While corpus-based studies are now becoming a new methodology in natural language processing, second language learning offers one interesting potential application. In this paper, we are primarily concerned with the acquisition of collocational knowledge from corpora for use in language learning. First we discuss the importance of collocational knowledge in second language learning, and then take up two measures, mutual information and cost criteria, for automatically identifying or extracting collocations from corpora. Comparative experiments are made between the two measures using both Japanese and English corpora. In our experiments, the cost criteria measure proved more effective in extracting interesting collocations such as fundamental idiomatic expressions and phrases.

KeyWords: *Corpus, Collocations, Mutual Information, Cost Criteria, Second Language Learning*

† Tokushima University, Faculty of Engineering
†† The National Language Research Institute, Section for Dictionary Research