Choosing a Reference Corpus for Keyword Calculation

Gwang-Yoon Goh
(Yonsei University)

Goh, Gwang-Yoon. 2011. Choosing a Reference Corpus for Keyword Calculation. Linguistic Research 28(1), 239-256. Keywords, which are known to provide a useful way to characterize a text, are usually calculated using two word lists, one from the study corpus (SC) and the other from the reference corpus (RC). Although this notion of keywords has attracted great attention and been employed in many corpus-based language studies, the issue of what constitutes a good or appropriate RC has been left largely untouched, although an RC is generally expected to be larger in size than the SC. This paper looks into how different factors associated with the RC affect the outcome of the keyword calculation of a given SC. The results indicate that genre and diachrony are more important factors to consider than other factors when choosing an RC, especially in that the differences in these two factors, unlike those in other factors such as corpus size and varietal difference, bring about a statistically significant difference in the number of the keywords. Despite the possible effects that the size and composition of the RCs can have on keyword calculation and resulting differences in keyword results, however, keyword analysis is very robust and keywords can be plausible indicators of aboutness, regardless of the RC one chooses. Thus, the aboutness of a text should be interpreted with its possible diversity caused by the use of different RCs in mind. (Yonsei University)

Key Words reference corpus, study corpus, keyword, word list, corpus size, genre, national variety, diachrony