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

1. Introduction

A keyword normally indicates a significant word from a title or document used as an index to content. In corpus-based linguistic studies, however, the notion is defined as a word “whose frequency is unusually high in comparison with some norm” (Scott 2008: 135). In particular, although they may not be the most important words in the given text or corpus (mainly because their importance or keyness is

* I am grateful to the anonymous referees for their helpful comments and suggestions.
determined purely statistically), keywords often provide a useful way to characterize a text or a genre and can be used to analyze lexi-co-grammatical features in a corpus. Thus, keyword analysis has great potential for application in linguistics and other relevant fields, including language teaching, forensic linguistics, stylistics, content analysis, and text retrieval (Scott 2008, O’Keeffe et al. 2007).

In general, keywords are computed using two word lists, one from the text or study corpus (SC) that one wants to investigate and the other from a normally larger, reference corpus (RC) that acts as a benchmark corpus or provides background data for keyword calculation. Since keyword calculation is performed basically on the basis of the comparison with the word list of the RC, its results are highly likely to be influenced by the RC chosen by a researcher. This makes us wonder what effects the size and composition of the RC have on keyword calculation and its results and what constitutes a good or bad RC. However, these important issues do not seem to have been sufficiently addressed in any previous studies, although the notion of keywords has attracted great attention and keyword analysis has been employed in many corpus-based language studies (e.g. Tribble 2000, Kemppanen 2004, Scott & Tribble. 2006, Seale et al. 2006, Goh & Lee 2008, Rayson 2008, Mahlberg 2009, McEnery 2009).

This paper is a quantitative study of the relationship between the RC and keyword calculation results. In particular, it will look into how various factors closely associated with the RC affect the keyword calculation of a given SC and its results. In analyzing keyword results and their differences, our discussion will be limited to the size of the keyword lists obtained from SCs in comparison with different RCs, thereby leaving the discussion of the composition of the keyword lists as a question for a future study.

The organization of this paper is as follows. We will first briefly review keyword calculation process and relevant previous studies. We will then investigate the roles of four major factors, including corpus size, genre, national variety, and diachrony, in keyword calculation, focusing on whether differences in any of these four major factors can cause a statistically significant difference in the number of keywords produced. We will also consider what the results of our analysis suggest about keyword analysis and its interpretation.