Pojanapunya, P., & Watson Todd, R. (2021). The influence of the benchmark corpus on keyword analysis. Register Studies, 3(1), 88-114.

The definitive version of this article was published as Pojanapunya, P., & Watson Todd, R. (2021). The influence of the benchmark corpus on keyword analysis. Register Studies, 3(1), 88-114.
DOI: https://doi.org/10.1075/rs.19017.poj
It is available at https://www.jbe-platform.com/content/journals/10.1075/rs.19017.poj

The Influence of the Benchmark Corpus on Keyword Analysis

The growing popularity of keyword analysis as an applied linguistics methodology has not been matched by an increase in the rigour with which the method is applied. While several studies have investigated the impact of choices made at certain stages of the keyword analysis process, the impact of the choice of benchmark corpus has largely been overlooked. In this paper, we compare a target corpus with several benchmark corpora and show that the keywords generated are different. We also show that certain characteristics of the keyword list and of the keywords themselves vary in relatively predictable ways depending on the benchmark corpus. These variations have implications for the choice of benchmark corpus and how the results of a keyword analysis should be interpreted. Analyzing the keywords from a comparison with a large general corpus or the keyword lists from multiple comparisons may be most appropriate for register studies.

Keywords: keyword analysis, reference corpus, aboutness, register

1. Introduction

Keyword analysis has become an increasingly used research technique in applied linguistics. A Google Scholar search for *'keyword analysis'* and *corpus* returns over 2,500 articles (as of mid-2019), about half of which have been published since 2015. Many of these studies aim to provide

a lexical characterization of a genre or a register, and thus keyword analysis has become an important tool for register studies. The recent popularity of keyword analysis may reflect how quickly and easily an analysis can be conducted, but this "recent explosion in studies involving keyword analysis has not been matched by sensitivity to the procedures they involve" (Culpeper 2009: 53). Conducting a keyword analysis requires the researcher to make impactful decisions at several points. In this paper, we will focus on one of these decision points that has been identified as a key issue that can affect the results of the analysis, namely, the choice of the benchmark corpus (Scott 2006). Selecting an appropriate benchmark corpus could strengthen a study, while it may be possible to produce biased results by selecting an inappropriate corpus. To see how the benchmark corpus affects keyword analysis results, we will examine the characteristics of the keyword lists (such as the mean keyness statistic value of words identified as keywords) and the keywords produced when the same target corpus is compared to five different benchmark corpora.

1.1 Uses of keyword analysis

Keywords are words which exhibit keyness, "a quality words may have in a given text or set of texts, suggesting that they are important, they reflect what the text is really about" (Scott & Tribble 2006: 55-56). Deciding which words are 'important' is based on frequency, but "this does not mean high frequency but unusual frequency, by comparison with a reference corpus" (Scott 1997: 236). Keywords, then, are words whose frequency is notably higher in the text(s) under investigation than in comparative text(s), and this relatively high frequency is taken as showing their importance to the text(s) under investigation.

Identifying which words have such unusual frequency is a deceptively straightforward procedure, facilitated by several automated corpus analysis tools. The ease with which keywords can be identified may underlie the recent explosion in the popularity of keyword analysis. From a survey of 84 previous research studies using keyword analysis (Pojanapunya 2017), the procedure is most popular in investigations of media discourse and academic discourse (21 articles each), but the range of possible applications is wide and also includes health communication, political discourse, and literature.

The majority of these previous studies have taken one of two broad approaches: an interpretive analysis of some keywords in the text(s), or a lexical characterisation of the genre or

register that the text(s) represent. The former usually involves an analysis of concordance lines of the selected keywords and is often associated with critical paradigms. The latter usually aims to identify the lexis that distinguishes the genre or register under investigation from others. In both cases, the keywords are taken as providing information concerning the aboutness or style of the text(s) under investigation (Scharl & Weichselbraun 2008; Gerbig 2010).

1.2 Conducting a keyword analysis

A keyword analysis aims to provide insights into a corpus under investigation, variously termed the target corpus, the node corpus, and the study corpus. This target corpus may be a text, part of a text, or a collection of texts. The first stage in the analysis is to conduct a word frequency count of this corpus producing a list of word types together with the frequency with which each appears in the corpus. Although there is a range of possible choices concerning what counts as a word (e.g. lemma, word family) (Gardner 2007), the vast majority of studies use surface word forms as the basis for counting since this is the most practical option.

Since keyness concerns relative frequency, a corpus against which to compare the word frequencies in the target corpus is needed. This second corpus is variously termed the benchmark corpus, the comparative corpus, and the reference corpus (we will use the term *benchmark corpus* since some authors use *reference corpus* to refer specifically to a large general corpus). The benchmark corpus may comprise the rest of the text (where the target corpus is part of a text), another text, another collection of texts, or a general corpus such as the British National Corpus (BNC). Although it is not actually necessary for statistical reasons, in nearly all studies the benchmark corpus is at least as large as the target corpus. A word frequency count of the benchmark corpus is conducted.

The two word frequency lists from the target corpus and the benchmark corpus can then be compared statistically. A large number of possible statistics can be used but generally they fall into two categories: probability statistics (e.g. log likelihood (LL), chi square), and effect size statistics (e.g. odds ratio, Damerau's relative frequency ratio). Choosing whether to use a probability statistic or an effect size statistic can have a major impact on what words will be identified as keywords (Pojanapunya & Watson Todd 2018) with effect size statistics highlighting more unusual words suitable for critical purposes and probability statistics producing keywords more useful for register studies. The choice between the various possible

statistics within each category results in little variation in the keywords identified. Probability statistics are far more commonly used than effect size statistics. In the survey of 84 previous keyword analysis studies, a probability statistic was used in 81% of studies which clearly identified the statistic used, and LL was used in 84% of these (Pojanapunya 2017), perhaps because it is readily available in all of the corpus analysis programs used for conducting a keyword analysis.

From the statistical comparison of the two word frequency lists, a list of all of the word forms in the target corpus together with their keyness values is produced. In most cases, this list is ranked by keyness value, and the researcher then needs to decide a cutoff point above which the words will be considered to be keywords. Again, a range of choices is available. Many studies apply filters such as a required minimum frequency or minimum dispersion through the corpus as an initial step. Dispersion, following Scott and Tribble (2006), is typically measured simply as the proportion of texts in a corpus in which a word appears with those studies using a dispersion filter most commonly setting the proportion at 5% of texts (e.g. Gilmore & Miller 2018). Cutoff points are usually identified as either a minimum statistical or probability value (these can be problematic as the values are often heavily influenced by the size of the corpora, see Pojanapunya & Watson Todd 2018) or a threshold number of keywords (such as the top 10, the top 50, or the top 100). The decisions at this stage are dependent on the purpose of the research, but ultimately the specific choice made is usually arbitrary.

Once a keyword list has been compiled, the final decision, dependent on the purpose of the study, concerns how to deal with the keywords. Choices at this stage include choosing a few keywords to focus on and investigate in depth, categorizing the keywords, and looking for patterns in the keywords related to the research purpose.

Although conducting a keyword analysis appears straightforward, we can see that there are numerous points at which the researcher needs to make a decision, and these decisions can have major impacts on the results. The decisions include choosing what to count as a word, choosing the benchmark corpus, choosing the statistic(s) to use, choosing a cutoff for keyness, and choosing how to deal with the keyword list. There has been research that has examined the impacts of the choices at each stage. For example, for the choice of keyness statistic, Gabrielatos and Marchi (2012) conducted several keyword analyses using different statistics for various corpora to see if similar keywords were produced and found almost no overlap between

keywords produced using probability statistics and effect size statistics; and Pojanapunya and Watson Todd (2018), also comparing keywords produced using probability statistics and effect size statistics, found that different statistics served different purposes (also see Kilgarriff & Berber Sardinha 2000 for further methodological concerns relating to comparing corpora). The stage which can have a major impact but for which the existing research provides the least guidance for researchers is choosing the benchmark corpus.

1.3 Research into the effects of different benchmark corpora

Five previous studies have investigated the impact of different benchmark corpora on the keywords produced. Xiao and McEnery (2005) investigated the effects of the size of the benchmark corpus by comparing three genre corpora against two benchmark corpora of different sizes. The three target corpora comprised conversations, professional spoken language, and academic prose; the two benchmark corpora were the 100-million-word BNC and the one-million-word Freiburg-LOB Corpus. Examining the top 10 keywords produced by each comparison, they found that the keywords were very similar for the two benchmark corpora. They therefore concluded that the size of the benchmark corpus was not important.

Scott and Tribble (2006) investigated the keywords in *Romeo and Juliet* produced by comparisons with three benchmark corpora: Shakespeare's other tragedies, Shakespeare's complete works, and the BNC. Although they found some variation in the keywords produced, the analyses generated a common core of keywords irrespective of the benchmark corpus, allowing them to argue that keyword analysis is a robust procedure producing similar results when a large benchmark corpus is used.

The third study (Scott 2006) specifically set out to find an inappropriate benchmark corpus. Two target corpora, a book profile and an extract of doctor-patient communication, were compared to various benchmark corpora. The quality of the keyword lists produced were evaluated based on the extent to which the keywords matched the keywords produced by other comparisons, termed comparative precision by Scott. In the first stage, the two target corpora were compared against 22 benchmark corpora of different sizes created by randomly selecting different numbers of texts from the BNC. As with Xiao and McEnery (2005), the findings showed that the size of the benchmark corpus had very little impact on the keywords produced. In the second stage, the two target corpora were compared against a supposedly absurd

benchmark corpus, Shakespeare's works, but again the keywords produced were judged reasonable. Finally, the first target corpus was compared against nine benchmark corpora representing different genre categories in the BNC (e.g. prose fiction, spoken discussions). In this case, there was less agreement between comparisons (the comparative precision values ranged from 20% to 70%), allowing Scott to conclude that "the keywords generated do differ if genre-different RCs [reference corpora] are used" (p. 10). However, the nature of these differences was not explored beyond suggesting that they reflected different types of aboutness.

The final two studies are the most rigorous. Goh (2011) compared a target corpus consisting of Sherlock Holmes short stories to several benchmark corpora designed to vary on four dimensions: size, genre, variety (i.e. British or American English), and diachrony. Goh shows that benchmark corpora based on differences in genre and diachrony produced significantly different numbers of keywords when a probability value threshold was used. However, it is unclear whether these benchmark corpora result in different actual keywords, since Goh only investigated numbers of keywords produced.

Geluso and Hirch (2019) compared two very similar target corpora of applied linguistics research to three benchmark corpora: a much larger corpus of published applied linguistics research, a corpus of newspaper and magazine articles, and a corpus of fiction. Their analysis shows that the choice of benchmark corpus influences the keywords produced. Comparison with a very similar benchmark corpus highlights "content unique to each target corpus" (p. 209), such as technical terms in applied linguistics and researchers' names. We will term such content specific aboutness. Comparison with benchmark corpora of other registers generates keywords that are more general to the register (which we will term general aboutness), such as broad concerns in applied linguistics, or style.

These studies give mixed results about the potential for the choice of benchmark corpus to influence the keyword results. The size of the benchmark corpus does not affect the keywords produced very much, but the genre, register and diachrony of the benchmark corpus has some influence. Because of the methods used to evaluate the keywords generated in the different comparisons in these four studies, the findings only show whether there is an impact from the choice of benchmark corpus but give little information about how the benchmark might influence the results of a keyword analysis. The nature of the influence of the choice of the benchmark corpus is, then, still open to debate, and the literature is replete with quotations

suggesting that the impact can be large. For example, "the choice of the reference corpus will affect whether you acquire keyword results that are all relevant to the particular aspect of the text(s) you are researching" (Culpeper 2009: 35); and "the choice of reference corpus does, of course, affect the reliability of comparisons" (Jones, Byrne & Halenko 2018: 37).

If the choice of benchmark corpus in some cases has a large impact on the keywords produced, we need to know the nature of the impact. With the exception of Geluso and Hirch's (2019) finding that the amount of similarity between the target and benchmark corpora influences the specificity of the keywords produced, there is little in the previous studies that could be used by researchers to guide their choice of benchmark corpus. For example, if a researcher aimed to generate a word list of technical terms using keyword analysis, using a similar benchmark corpus is likely to be more productive. This is potentially useful guidance, and conducting further comparisons between the results produced by other types of benchmark corpora could generate further guidelines that researchers could follow for other purposes. These potential guidelines could concern the types of keywords produced, the number and amount of variation in the keywords produced, and the nature of the keyword list as a whole (for example, how compact the list is, or the average level of keyness of the list).

Since the extent to which and the ways in which the choice of the benchmark corpus affects the keywords generated is still unclear, the purpose of this study is to investigate the extent to which and how the choice of benchmark corpus affects the keyword analysis results. In doing this, we will look at the impact on two different aspects of the results. First, we will investigate the impact on the keyword list as a whole by examining certain quantitative measures that characterize a whole keyword list. Second, we will investigate the impact on the actual keywords by assigning values to the keywords generated in numerous different ways. Therefore, we will attempt to answer three research questions:

1. To what extent does the choice of the benchmark corpus influence the results of a keyword analysis?
2. How does the choice of the benchmark corpus influence the characteristics of the keyword list generated as a whole?
3. How does the choice of the benchmark corpus influence the keywords generated?

2. Methodology

In setting up the methodology to investigate the influence of different benchmark corpora, there are numerous decisions to be made. In some cases, the decision is dictated by our research purpose; in other cases, there is a range of alternatives available, all of which meet the research purpose. In these latter cases, we will choose the alternative which has been most commonly used in previous keyness research as this will allow our findings the greatest applicability.

2.1 The target corpus

As we saw above, a target corpus may be a text, part of a text, or a collection of texts. Since the previous research has shown that using a reasonably large corpus leads to consistency in keyword results, we will use a collection of texts as the target corpus. In this study, we are focusing primarily on the purpose of conducting a keyword analysis to characterize a genre or register (rather than the more critical interpretative purpose), and so the texts comprising the target corpus should all be from the same genre or register. To facilitate the interpretation of the results, we decided to use a genre familiar to us, namely, applied linguistics research articles.

The target corpus, then, consists of 400 research articles (to ensure a reasonably large corpus size and to fulfill Cochran's (1977) sampling criterion) from the field of applied linguistics. To create the corpus, a list of highly-ranked journals (see Appendix A) which cover a range of sub-topics in applied linguistics and which include both quantitative and qualitative research was made. The most recent articles from these journals which follow an Introduction-Methodology-Results-Discussion format and which are between 4,000 and 8,000 words in length were selected. Before inclusion in the corpus, tables, figures, reference lists, and appendices were removed. The final target corpus of applied linguistics research articles is termed AL1 and consists of about 2.8 million words. Although 400 articles may not seem very large, we believe it is more important to ensure that all articles in the corpus fit the criteria. With the limited numbers of high-level journals with limited number of articles in them, 400 articles seem reasonable.

2.2 The benchmark corpora

When the target corpus is a collection of texts, we can use other collections of texts or a general corpus as the benchmark as these would be reasonably large corpora. To decide what other collections of texts to use, we have seen that genre-different benchmark corpora influence the

keyword lists and that the nature of this influence needs further exploration. It should be noted that Scott's (2006) use of *genre* is more akin to Biber's (1988) registers than, say, Swales' (e.g. 1990) genre. Registers can vary in their content, setting, interactiveness, and mode (Biber, Conrad & Reppen 1998). To investigate the effects of different benchmark corpora on the results of a keyword analysis, in this study we will look at a range of variation in content and a variation in mode as these can be linked to aboutness and style.

To create a range of variation in content, given that the target corpus is a collection of research articles from a specific discipline, we can operationalize content concern as academic discipline, since Hyland (e.g. 2004) has shown that language features, including the vocabulary likely to be identified as keywords, varies across disciplines. Related disciplines, such as applied linguistics and education, are likely to have less variation than unrelated disciplines, such as applied linguistics and mechanical engineering. A continuum of disciplinary relatedness represents the range of variation in content, and allows us to create three benchmark corpora representing a continuum of similarity to the target corpus. To represent the highest degree of similarity, the first benchmark corpus (termed AL2) is a corpus of applied linguistics research articles created following the same procedures as were used for the target corpus but with different articles. Since it has the same characteristics in terms of content, genre and mode as the target corpus, the AL2 corpus is akin to a control comparison. The second benchmark corpus (termed SSH) is somewhat similar to the target corpus, since it is a corpus of research articles from disciplines in the social sciences and humanities which are related to applied linguistics (e.g. psychology, education). The third and most distant benchmark corpus in terms of content relatedness to the target corpus (termed SCI) is a corpus of research articles from unrelated disciplines in the sciences (e.g. microbiology, mechanical engineering). The construction of SSH and SCI followed the same procedures as the other research article corpora. These three corpora are all from the same genre as the target corpus (research articles) but fall on a continuum of similarity in terms of their content. Since corpus size can affect keyness values (for log likelihood), these three benchmark corpora are roughly the same size as the target corpus (see Table 1).

For variation in mode, since the target corpus is written academic language, we need a benchmark corpus of spoken academic language. For this we used 152 transcripts from the

Michigan Corpus of Spoken Academic English (MICASE). Tags in the transcripts were removed.

As a final benchmark corpus, we will use a general corpus since this is a common practice in previous keyness research (see Pojanapunya 2017). The most commonly used general benchmark corpus (used in 30 of 50 previous studies using a general benchmark corpus) is the BNC, which we will use as the fifth benchmark corpus in this study. These previous studies typically use the whole BNC as the benchmark. Although doing this means that this fifth benchmark corpus is vastly different in size from the other corpora (an issue that can be solved through normalization), using the whole BNC allows comparison with previous studies.

To investigate the effects of benchmark corpora on the results of a keyword analysis, we will compare the same target corpus against five benchmark corpora designed to manifest a range of variables of difference. The corpora used in this study are summarized in Table 1.

[TABLE 1 NEAR HERE]

2.3 Generating the keyword lists

To generate the keyword lists, the main decisions concern the keyness statistic to use and the cutoff point for deciding which words are keywords. To allow this study to be widely applicable, the keyness statistic used is the frequently used log likelihood (LL) statistic. Applying a log likelihood keyness analysis to each pair of corpora produces long lists of words (each 35,094 long – the number of different words in the target corpus) sequenced by LL value. The next stage is to remove words which are unlikely to be important to the AL1 corpus as a whole since the words have low frequencies (but may still have high LL values if they do not occur in the benchmark corpus) or are restricted to a few articles in the AL1 corpus. We therefore decided to set two initial filters for words to be considered keywords (a frequency filter of occurring at least 200 times in AL1, and a dispersion filter of occurring in at least 40 of the 400 texts in AL1). These filters aim to ensure that the keywords are representative of the whole corpus and were set intuitively following examination of initial results. After that, we need to decide which of the words should be considered keywords, a process which involves starting with the highest ranked keyword and working down the list until a cutoff point is reached. Since this study focuses on using keywords for register characterization purposes, we will need at least, say, 50 keywords to

ensure that the salient features we use in research question 3 are each illustrated through more than one keyword and so provide a reasonable characterization of the target corpus, but the choice of how to set a cutoff point depends on the research question.

For research question 1 asking if the choice of benchmark corpus influences the keywords produced, following Scott (2006) we will use comparative precision as a measure of similarity between keyword lists. This measure is most effective when keyword lists are the same length, so for research question 1 we will use a cutoff point of the top 100 words (the most commonly used cutoff point in previous keyword research, see Pojanapunya (2017)) to produce keyword lists.

For research questions 2 and 3, we need justifiable keyword lists which are somewhat comparable in terms of both keyness values and the number of keywords. The two most commonly used approaches are to use the top n keywords (as in our approach for research question 1) or to set a minimum LL value or associated probability value. An LL or probability value threshold was used in 25% of previous keyness studies (see Pojanapunya 2017) and allows us to compare our results with Goh's (2011) study. One criticism of using a statistic value threshold is that these values are greatly influenced by the size of the corpora. Since most of our corpora are of comparable size and we can normalize the values for the comparison against the BNC, this should not be a problem.

Having decided to use a statistic value as the cutoff threshold, we need to decide what this value should be. Traditionally, research has used a probability value of 0.05, equivalent to an LL value of 3.84, to identify significance (although most corpus research uses much lower probability values), so we will start with this value and see how many keywords are produced in each of the comparisons. We can then recursively increase the LL value by an order of magnitude until we reach a point where the number of keywords produced is manageable for further analyses. The numbers of keywords produced using the various LL values as cutoffs are shown in Table 2.

[TABLE 2 NEAR HERE]

From Table 2, it appears to be impossible to set a single cutoff statistic value that could be applied to all comparisons to produce manageable keyword lists, since the numbers of

keywords produced vary so much between lists. This confirms Goh's (2011) finding that the benchmark corpus affects the number of keywords produced when statistic values are used as cutoffs, but means that we need to search for a different method of determining cutoff thresholds to be able to answer research questions 2 and 3. Using the top n keywords also produces results differing by at least an order of magnitude. For example, the minimum LL value of the top 100 words in AL1 vs. SCI is 709.6, but using this value would produce no keywords for AL1 vs. AL2. We therefore need to find an alternative to the usual practices to allow comparability between the different keyword lists, and so we used a proportion of the LL value range as the cutoff. To do this, we calculated the full range of LL values for all words produced in a comparison (removing the top two and bottom two words as potential outliers), and then used the top 5% of this range of LL values as the cutoff. Using this method, although the LL values of keywords in the AL1 vs. AL2 keyword list are not very high, the distribution of keywords is very narrow. Therefore, there are a large number of keywords within the top 5% of the LL value range of this keyword list. Doing this produces the number of keywords and cutoff LL values for each of the comparisons shown in Table 3 (the keywords in these lists are given in Appendix B). Although the range of the minimum LL values for a word to be considered a keyword is very large, these LL values all represent the 95th percentile of the LL range and so are somewhat comparable; the numbers of keywords in the different keyword lists vary by less than an order of magnitude and so are also somewhat comparable. In addition, using this method allows us to investigate features of keyword lists, such as the kurtosis of the keyness distribution, which would not be apparent using other cutoff measures. Although this is a novel method for setting a cutoff point, we believe that the number of keywords produced by this method are sufficient to allow us to see how the benchmark corpus influences the results of a keyword analysis while ensuring that all of the keywords are potentially important in the target corpus.

[TABLE 3 NEAR HERE]

2.4 Analysing the keyword lists and the keywords

To answer research question 1, whether the benchmark corpus influences the results of a keyword analysis, we need to see if the keywords generated from comparisons to different benchmark corpora are similar or different. This can be done by calculating the proportion of

keywords in one keyword list which also appear in another keyword list. This is Scott's (2006) comparative precision, and is similar to the concepts of precision and recall commonly used to evaluate natural language processing applications (Ferret & Grau 2000). Since precision values are influenced by the number of keywords in a list, interpretation of values is more straightforward if the number of keywords in different keyword lists is the same, so we will use the keyword lists based on the top 100 words ranked by keyness values for research question 1 (this also means that the precision and recall values are the same). With five keyword lists, we will calculate the minimum, maximum and mean precision values from comparisons between all possible pairs of lists, all possible sets of 3 and 4 lists, and all lists.

For research questions 2 and 3 (how the benchmark corpus influences the keyword list as a whole, and how the benchmark corpus influences the keywords generated), the justifiability and validity of the keyword lists is more important than having the same number of words in each list. For these questions, then, we will use the keyword lists generated based on the top 5% of keyness values summarized in Table 3.

The influence of the benchmark corpus on the characteristics of the keyword list taken as a whole has generally been overlooked in previous research. There are, however, a few studies which have examined certain characteristics of keyword lists: the number of keywords generated (e.g. Blaxter 2014), the distribution of the keywords in the target corpus which can be represented by the range of texts keywords appear in (e.g. Paquot & Bestgen 2009), the frequency of the keywords in the target corpus (e.g. Camiciottoli 2016), and the average keyness value for keywords (e.g. Kotze 2010). Since we are defining words as key when their keyness value is in the top 5% of the range, the number of keywords generated indicates the length of the tail in the distribution of keyness values. A longer thinner tail results in fewer words falling within the top 5%. To represent the length of the tail, we will calculate 1/number of keywords and multiply this by a constant (10,000) to produce figures comparable to the other variables. For dispersion, we will calculate the number of keywords that occur in at least 50% of the texts in the target corpus (the minimum dispersion criterion for a word to be considered a keyword is 10%). We will call such words pervasive keywords (in contrast to restricted keywords which only occur in a small proportion of the texts in the target corpus). For the frequency in the target corpus of the words in the keyword lists, since the distribution of keywords is likely to be a Pareto distribution, we will use the median frequency of the keywords, rather than the mean. Finally, we

will calculate the mean LL value for the keywords (normalized for the comparison against the BNC and using the geometric mean). In these calculations, where appropriate, we will normalize the data so that the different sizes of the benchmark corpora do not affect the results.

We have seen that keywords may be indicative of the aboutness or the style of the target corpus and that there may be different types of aboutness. Research question 3 aims to investigate whether the keywords generated by comparison to different benchmark corpora highlight different aspects of the target corpus. To do this, the keywords in each of the lists were evaluated for 254 features (the full set of features and values can be found in Appendix C). Some of these features can be linked with style and some with aboutness. In addition, the aboutness features vary in level of specificity. To illustrate this, the features associated with style include percentages of words associated with Biber's (1988) register dimensions and percentage of words associated with stance and engagement. For example, the percentages of the keywords associated with Biber's involved and informational dimensions were calculated, as were the percentages of adverbs which act as downtoners, hedges, amplifiers and emphatics. In total, there are 48 features directly linked to style. The features associated with aboutness include percentages of words falling into the semantic categories of the USAS semantic tagset (Archer, Wilson & Rayson 2002). For example, the percentages of the keywords linked to the field of education and the percentages of the noun keywords linked to research processes were calculated. For specificity of aboutness, features include mean values for psycholinguistic properties using the MRC psycholinguistic database (http://websites.psychology.uwa.edu.au/school/MRCDataBase/uwa_mrc.htm) and percentages of words categorized into various existing word lists. These allow us to identify levels of lexical sophistication and familiarity which are associated with specificity. For each of the 254 features, a value was calculated for each comparison against a benchmark corpus. Most of these values are percentages (e.g. the percentage of adverbs which act as downtoners), but a few features, such as the psycholinguistic properties, are calculated as mean values.

To see which of the features were salient for a given comparison, we conducted two analyses. First, taking the AL1 vs. AL2 keyword list as a control comparison, we investigated the extent to which the nature of the keywords in the other four comparisons differed from this using an effect size statistic. Second, we compared the values generated for each of the four

comparisons representing a type of difference from the target corpus with the values from the other three comparisons.

For the first method involving analysis against the control comparison, different types of values require different effect size statistics. The most commonly used effect size statistic for comparing means is Cohen's d. For frequency percentages, there are several possible effect size statistics. For each of the 254 features, we could therefore generate a Cohen's d value for each of the comparisons against the benchmark corpora representing a type of difference. Given the differences in how the values were calculated, positive d values for features measured by frequency and any d values for features measured by means were considered potentially important. To see which features for which comparison could be considered salient, following Meltzer et al. (2016), we used a minimal important difference value of 0.2. In other words, features in a comparison where the d value was 0.2 or greater (or -0.2 or less for features measured by means) were identified as potentially salient.

To investigate the second method examining the values across the four comparisons representing a type of difference, we calculated the z-score or standard score for each feature for each comparison, which indicates how far a score varies from the mean in standard deviation units. In previous work, a z-score of 3 or more has been identified as salient (Nkechinyere et al. 2015). In our case, where only 4 values were used to calculate the z-scores, we need to use a lower cutoff point for saliency. Therefore, any feature in a comparison with a z-score of 1 or more was identified as potentially salient.

To illustrate how these procedures work, for the frequency of prepositions by type (see Appendix C row 12), for the first method we calculated the effect size difference between the value found for each of the four main comparisons and the value for the comparison against AL2. From this, AL1 vs. MICASE is the only comparison showing a clear difference with a d value of 0.5. For the second method, examining how far the scores for each of the four main comparisons differ from the mean, we find the percentage of keywords that are prepositions for AL1 vs. MICASE to be 13.0 whereas the values for the other three main comparisons are 0.0, 1.1 and 2.5, meaning that AL1 vs. MICASE has a z-score of 1.7 with the other comparisons having z-scores of 0. From this, we can see that AL1 vs. MICASE is an outlier suggesting that the use of prepositions is particularly salient in this comparison.

With two methods being used to identify the saliency of features, to be certain in our conclusions, we decided to identify only those features in a comparison which were identified as potentially salient by both methods as truly salient. For the findings on frequency of prepositions by type, the d value of 0.5 is greater than the cutoff of 0.2, and the z-score of 1.7 is greater than the cutoff of 1, showing saliency by both methods.

3. Results

3.1 Does the benchmark corpus influence the keywords?

For research question 1, we are using keyword lists comprising the top 100 words ranked by keyness values, and looking at the amount of precision (or the number of keywords shared between lists). If a large proportion of the keywords are shared, we can conclude that the benchmark corpus is not influencing the keyword results much; if few keywords are shared, then the benchmark is having an influence. The numbers of shared keywords between the various possible combinations of the five keyword lists (AL1 vs. AL2, AL1 vs. SSH, AL1 vs. SCI, AL1 vs. MICASE, AL1 vs. BNC) are given in Table 4.

[TABLE 4 NEAR HERE]

From Table 4, we can see that only 3 keywords are common to all lists, and generally the numbers of shared keywords are fairly low. Pairs of keyword lists, on average, share fewer than half of the keywords suggesting that the choice of benchmark corpus has a large influence on the results of a keyword analysis, contrasting with the findings of Scott and Tribble (2006) but in line with the findings of Goh (2011) and Geluso and Hirch (2019).

3.2 How does the benchmark corpus influence the keyword lists as a whole?

For the second research question, we will use the keyword lists based on the top 5% of the range of keyness values which are summarized in Table 3. The five keyword lists were analysed for four variables: number of keywords representing the length of the tail of the keyword distribution, dispersion of keywords, frequency of keywords, and average keyness value. The results are shown in Table 5.

[TABLE 5 NEAR HERE]

From Table 5, we can see that there appears to be a consistent pattern across keyword lists, with the values for AL1 vs. AL2 being the lowest, and the values for AL1 vs. MICASE the highest. This pattern becomes even clearer if we graph these results as in Figure 1. For the first four keyword lists from comparisons with specific corpora, we can conclude that the more similar the target and reference corpus, the shorter the tail of keyness distributions (in other words, the distribution is leptokurtic), the fewer the number of keywords distributed through the majority of the target corpus, the less frequent the keywords are in the target corpus, and the lower the keyness values. The comparison with a general corpus (AL1 vs. BNC) tends to appear in the middle of the values.

[FIGURE 1 NEAR HERE]

3.3 How does the benchmark corpus influence the keywords?

For the third research question, we will again use the keyword lists based on the top 5% of the range of keyness values. All words in all lists were analysed for 254 features. The full results are given in Appendix C. Features which showed both a Cohen's d value of 0.2 or greater and a z-score of 1 or greater were considered salient. These salient features are summarized in Table 6.

[TABLE 6 NEAR HERE]

Based on Geluso and Hirch's (2019) findings, we can generate expectations concerning the salient features in the different comparisons. Since AL1 and SSH are fairly similar (same genre, related disciplines), we would expect keywords illustrating specific aboutness; for AL1 and SCI (same genre, unrelated disciplines), we would expect a greater focus on general aboutness; for AL1 and MICASE (different modes), we would expect style issues to be most salient; and for AL1 and BNC, both style and general aboutness should be highlighted. Compared to these expectations, the findings in Table 6 are mixed. The keywords for AL1 vs. SSH illustrate aboutness, but concern both specific and general aboutness; and the keywords for AL1 vs. MICASE highlight style issues, but also concern general aboutness. The findings for the

other two comparisons run counter to our expectations. AL1 vs. SCI highlights style issues, not general aboutness; and the comparison of AL1 with the BNC generates no salient features (suggesting that the keywords produced are a mixed bag). The three specific comparative corpora, then, highlight different aspects of the target corpus and these different aspects concern wider issues than the different types of aboutness suggested by Scott (2006) and Goh (2011), while the general corpus comparison does not produce a clear pattern to the keywords. It should also be noted that the more qualitative analysis used to answer research question 3 produces different results to the quantitative analysis used to answer research question 2. For example, when looking at the keyword lists as a whole, AL1 vs. SCI and AL1 vs. BNC were very similar; when focusing on the keywords themselves, there are clear differences between these two comparisons, suggesting that both quantitative and qualitative approaches are needed to provide a full picture of the results of a keyword analysis.

4. Discussion

With little in common between the various keyword lists, the findings contradict the previous research suggesting that as long as the benchmark corpus is large enough it has little effect and vindicate the claims that the choice of benchmark corpus influences the results of a keyword analysis. We acknowledge that, since we used only a single target corpus and compared this to only five different benchmark corpora, the generalisability of the findings is unclear and further research into the impact of benchmark corpora on the keywords generated is warranted. Without such further research, however, in discussing the implications of the findings, we will assume that the results are generalisable. The results from the various analyses are summarized in Figure 2.

[FIGURE 2 NEAR HERE]

The first row in Figure 2 concerns the mean LL values. While it is known that corpus size influences LL values, the influence of the type of benchmark corpus on LL values has not been previously acknowledged. Since the keyness values are influenced by at least two methodological artifacts, if LL is being used as the keyness statistic, we would argue that the cutoff point for identifying words as keywords should not be based on a certain LL value (or its

associated probability value). The top n words or the top n percentage of the keyness range provide a more appropriate basis for the cutoff point.

The ways in which the benchmark corpus influences the keyword results has implications for conducting and interpreting the results of a keyword analysis. The nature of these implications depends on how the benchmark corpus is chosen, and we suggest that there are three main ways of doing this: the benchmark corpus is determined by the data being analysed, the benchmark corpus is chosen to highlight a specific aspect of the target corpus, and the benchmark corpus is chosen to provide a general characterisation of the target corpus.

Research where the benchmark corpus is determined by the data often involves dividing a large data set into two corpora which are then compared. Examples of such research include Baker's (2009) study comparing the discourse of pro-fox hunters and anti-fox hunters, and Meier, Rose and Hölzen's (2017) diachronic study of doping-related articles from two time periods. In such research, the benchmark corpus is predetermined rather than being chosen. In analysing and interpreting the keywords generated by a comparison with a predetermined benchmark corpus, researchers should be aware of how the nature of this corpus could influence the results. For example, if the predetermined corpus is very similar to the target corpus, researchers would need to be careful about using dispersion as a filter since restricted keywords are likely to be generated from the analysis, and they need to be cautious in making claims about the general aboutness of the target corpus since the comparison is more likely to highlight specific aboutness.

In research aiming to highlight a certain aspect of the target corpus, researchers could choose a benchmark corpus most likely to produce keywords associated with that aspect. For example, Willis (2017: 217) aimed to "uncover patterns or styles of speech". Since the research focuses on style rather than aboutness, from our analysis a different specific benchmark corpus, particularly one varying in mode, is most likely to generate keywords providing insights related to the research purpose.

The third category is research aiming to provide a general characterisation of the target corpus. This purpose is most closely related to register studies. For example, Harvey et al. (2008: 306) used keyword analysis as a "means for best defining a particular language variety", and Loudermilk (2007: 191) aimed to identify "the salient linguistic, organizational and socio-communicative features" of the target corpus. In their research, Geluso and Hirch (2019)

advocated using a benchmark corpus taken from a different register for this purpose. In their work, they used three benchmark corpora, each of a different register and found that the more distant the register the better the characterization of the target corpus it provided. However, this 'better' characterization would still highlight differences between registers rather than being a neutral characterization of the register of the target corpus. In using a general corpus as well as corpora from different registers, our study suggests that using a general corpus as the benchmark may be most appropriate as a general corpus appears to be least likely to return keywords focused on a specific aspect of the target corpus and so provides the most neutral characterisation.

Most keyword analysis studies use a single benchmark corpus, but the findings from this study (especially concerning the keyword lists) suggest a possible application using multiple benchmark corpora. Where the target corpus is of an unknown category (most usually this corpus would be a single text), comparing this corpus to several benchmark corpora could allow the text to be categorized as belonging to a certain genre or register based on its keyword list similarities to the other corpora. The text would be categorized into the same category as the benchmark corpus which generates a keyword list with the most keywords, the lowest mean keyness value, and the lowest median frequency of the keywords. Currently, such categorisation can be conducted using the multidimensional analysis of Biber (1988) (e.g. in Xiao & McEnery 2005) or through some natural language text categorisation applications (e.g. Bigi, Brun, Haton, Smaïli & Zitouni 2001; Cselle, Albrecht & Wattenhofer 2007). Using the features of keyword lists provides another tool for achieving this purpose.

In this paper, we have shown that the choice of benchmark corpus influences the results of a keyword analysis, and that this influence follows relatively predictable patterns. These patterns may allow researchers to make better founded decisions in choosing a benchmark corpus and in interpreting the keyword results. With stronger foundations, we hope that keyword analysis can reach its full potential as a research methodology in applied linguistics.

References

Archer, D., Wilson, A., & Rayson, P. (2002). Introduction to the USAS category system. Retrieved from http://ucrel.lancs.ac.uk/usas/usas_guide.pdf

Baker, P. (2009). The question is, how cruel is it? Keywords, fox hunting and the house of commons. In D. Archer (Ed.), *What's in a word-List?: Investigating word frequency and key word extraction* (pp. 125-136). Aldershot: Ashgate.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bigi, B., Brun, A., Haton, J. P., Smaïli, K., & Zitouni, I. (2001). A comparative study of topic identification on newspaper and e-mail. *Proceedings of the 8th International Symposium on String Processing and Information Retrieval* (pp. 238-241). Retrieved from https://hal.inria.fr/inria-00107535/document

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Converting among effect sizes. In M. Borenstein., L. V. Hedges., J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to metaanalysis* (pp. 45-49). Chichester, UK: John Wiley & Sons.

Cochran, W. G. (1977). *Sampling techniques* (3rd Ed.). New York: John Wiley & Sons.

Cselle, G., Albrecht, K., & Wattenhofer, R. (2007). Buzztrack: Topic detection and tracking in email. *Proceedings of the 12th International Conference on Intelligent User Interfaces* (pp. 190-197).

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29-59.

Deeks, J. J., & Higgins, J. P. T. (2010). Statistical algorithms in review manager 5. Retrieved from http://ims.cochrane.org/revman/documentation/Statistical-methods-in-RevMan-5.pdf

Ferret, O., & Grau, B. (2000). A topic segmentation of texts based on semantic domains. *Proceedings of the 14th European Conference on Artificial Intelligence* (pp. 426-430). IOS Press.

Gabrielatos, C., & Marchi, A. (2012). Keyness: appropriate metrics and practical issues. *Critical Approaches to Discourse Studies*. Bologna. Retrieved from http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.

Geluso, J., & Hirch, R. (2019). The reference corpus matters: Comparing the effect of different reference corpora on keyword analysis. *Register Studies, 1*(2), 209-242.

Gerbig, A. (2010). Key words and key phrases in a corpus of travel writing. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 147-168). Amsterdam: John Benjamins.

Goh, G. Y. (2011). Choosing a reference corpus for keyword calculation. *Linguistic Research, 28*(1), 239-256.

Harvey, K., Churchill, D., Crawford, P., Brown, B., Mullany, L., Macfarlane, A., & McPherson, A. (2008). Health communication and adolescents: What do their emails tell us?. *Family Practice, 25*(4), 304-311.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, Michigan: University of Michigan Press.

Jones, C., Byrne, S., & Halenko, N. (2018). *Successful spoken English: Findings from learner corpora*. London: Routledge.

Kilgarriff, A., & Berber Sardinha, T. (2000). *Proceedings of the Workshop on Comparing Corpora*. Hong Kong.

Loudermilk, B. C. (2007). Occluded academic genres: An analysis of the MBA thought essay. *Journal of English for Academic Purposes, 6*(3), 190-205.

Meier, H. E., Rose, A., & Hölzen, M. (2017). Spirals of signification? A corpus linguistic analysis of the German doping discourse. *Communication & Sport, 5*(3), 352-373.

Meltzer, E. O., Wallace, D., Dykewicz, M., & Shneyer, L. (2016). Minimal clinically important difference (MCID) in allergic rhinitis: Agency for healthcare research and quality or anchor-based thresholds?. *The Journal of Allergy and Clinical Immunology: In Practice, 4*(4), 682-688.

Nkechinyere, E. M., Andrew, I., & Idochi, O. (2015). Comparison of different methods of outlier detection in univariate time series data. *International Journal for Research in Mathematics and Statistics, 1*(1), 55-83.

Pojanapunya, P. (2017). A theory of keywords. (Doctoral dissertation). Retrieved from https://opac.lib.kmutt.ac.th/vufind/Record/1370763.

Pojanapunya, P., & Watson Todd, R. (2018). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory, 14*(1), 133-167.

Scharl, A., & Weichselbraun, A. (2008). An automated approach to investigating the online media coverage of US presidential elections. *Journal of Information Technology and Politics, 5*(1), 121-132.

Scott, M. (1997). PC analysis of key words – and key key words. *System, 25*(2), 233-245.

Scott, M. (2006). In search of a bad reference corpus. Paper presented at Word Frequency and Keyword Extraction: AHRC ICT Methods Network Expert Seminar on Linguistics., Lancaster University, UK. Retrieved from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.167.2638&rep=rep1&type=pdf

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.

Swales, J. (1990). *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Willis, R. (2017). Taming the climate? Corpus analysis of politicians' speech on climate change. *Environmental Politics, 26*(2), 212-231.

Xiao, Z., & McEnery, A. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*(1), 62-82.

Table 1. The corpora used in this study

Name	**Description**	**Size (no. of words)**	**Purpose**
AL1	Research articles in applied linguistics	2.8 million	The target corpus
AL2	Research articles in applied linguistics	2.8 million	A benchmark corpus with the same characteristics as the target corpus (same genre, same content, same mode) which can act as a control comparison

SSH	Research articles in the social sciences and humanities	2.7 million	A benchmark corpus from academic disciplines related to the target corpus (same genre, similar content, same mode)
SCI	Research articles in science and engineering	2.1 million	A benchmark corpus for academic disciplines unrelated to the target corpus (same genre, different content, same mode)
MICASE	Spoken academic English	1.8 million	A benchmark corpus where the mode is varied (similar genre, similar and different content, different mode)
BNC	General English	100 million	A general benchmark corpus

Table 2. Numbers of keywords produced using different LL value cutoffs

LL cutoff value	**3.84**	**38.4**	**384**	**3840**
p value	**.05**	**$5.8x10^{-10}$**	**$1.7x10^{-85}$**	**0**
AL1 vs. AL2	406	68	0	0
AL1 vs. SSH	914	599	142	9
AL1 vs. SCI	1097	815	212	10
AL1 vs. MICASE	1296	1177	330	15
AL1 vs. BNC	1231	1035	378	21

Table 3. Basic data about the keyword lists

Keyword list	**Number of keywords**	**Minimum LL value for keywords**
AL1 vs. AL2	144	19.5
AL1 vs. SSH	84	647.8

AL1 vs. SCI	87	846.0
AL1 vs. MICASE	46	2033.7
AL1 vs. BNC	120	3350.1

Table 4. Numbers of shared keywords between different combinations of keyword lists

	No. of keywords shared between pairs of lists (N = 10)	**No. of keywords shared between sets of 3 lists (N = 10)**	**No. of keywords shared between sets of 4 lists (N = 5)**	**No. of keywords shared between all lists (N = 1)**
Average number of shared keywords	38.4	20.5	10.4	3
Minimum number of shared keywords	3	3	3	3
Maximum number of shared keywords	77	48	40	3

Table 5. Characteristics of the five keyword lists

Keyword list	**Number of keywords (1/#KWs)*10000**	**Dispersion of keywords %pervasive KWs**	**Frequency of keywords Median f**	**Average keyness value Geometric mean LL**
AL1 vs. AL2	69.4	24.3	541.5	42.7
AL1 vs. SSH	119.0	60.7	1929.0	1441.9
AL1 vs. SCI	114.9	72.4	2513.0	1833.6
AL1 vs. MICASE	217.4	93.5	5270.5	3632.2

AL1 vs. BNC	93.3	71.6	2358.5	2229.7

Figure 1. The characteristics of the keyword lists

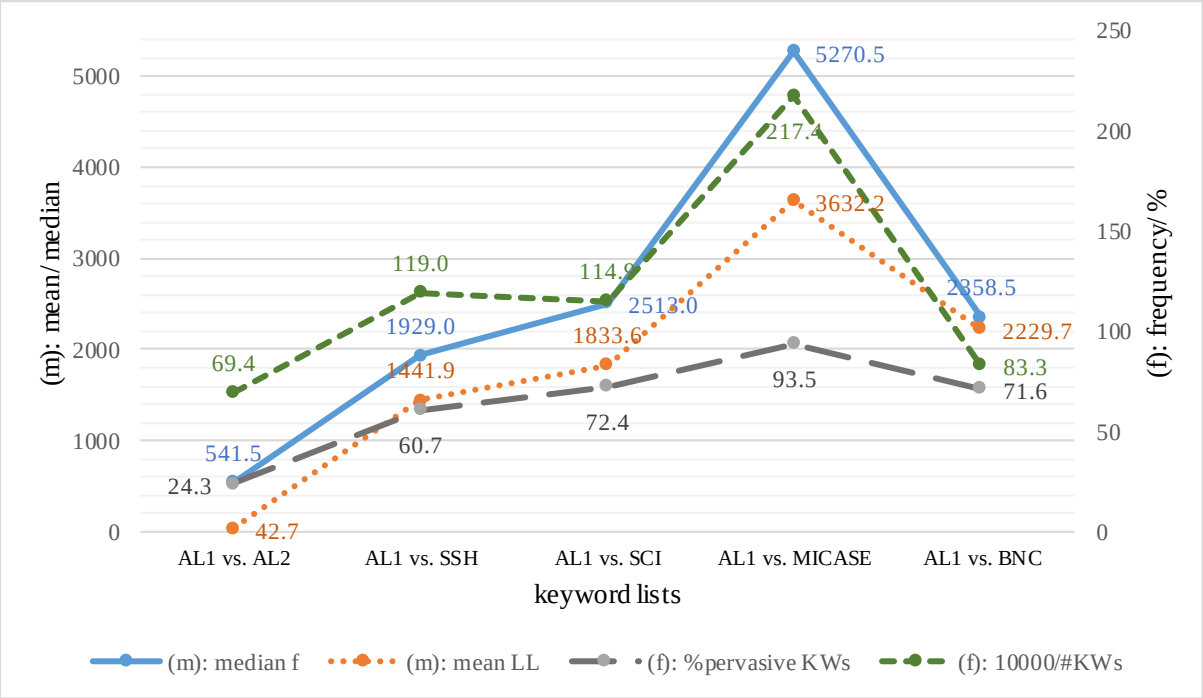


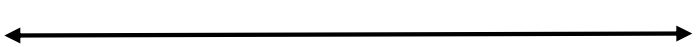
Table 6. Salient features of the keywords in comparisons with different benchmark corpora

Comparison	Salient features	Sample keywords	Aspect manifested	Shared aspect of salient features
AL1 vs. SSH	Semantic tags for Linguistic actions, states & processes (% of types)	*text, discourse, lexical, genre, language, words, vocabulary, sentence*	Specific aboutness General aboutness	Aboutness
	Verbs of communication (% of types) Verbs of activity (% of types) % of content keywords concerning action	*writing, speaking, listening, read, write*	General aboutness	
	Topical adjectives (% of types)	*lexical, grammatical, Chinese*	General aboutness	
	% of keywords not in top 15,000 words in the BNC (off-list words)	*ESL, EAP, raters*	Specific aboutness	
	% of keywords illustrating topics in applied linguistics	*proficiency, ESL, acquisition, EFL, pronunciation*	Specific aboutness General aboutness	

	Extent to which keywords with 1R collocate represent applied linguistics	*foreign* (language, teachers), *target* (language, words), *peer* (review, feedback), *vocabulary* (learning, knowledge)	General aboutness	
AL1 vs. SCI	Personal pronouns (% of types)	*you, they, she, he, it*	Style	Style
	% of keywords associated with Biber's narrative dimension of register % of keywords related to situational characteristic referring to shared personal knowledge % of keywords related to narrative communicative purpose	*their, they, her, she, he, his*	Style	
AL1 vs. MICASE	Prepositions (% of types) Prepositions (% of tokens)	*of, in, by, for, between*	Style	Style
	Function words (% of types) Function words (% of tokens)	*of, the, in, their, as, by, for, between, however*	Style	
	Semantic tags for Education (% of types)	*students, test, teachers*	General aboutness	

	Semantic tags for Names and grammatical words (% of types) Semantic tags for Names and grammatical words (% of tokens)	*their, the, of, in, for, as, were, by, between, such, however*	Style
	% of keywords appearing in more than 50% of the texts in AL1	*the, of, by, in, with, for, as, used, results, were, using, study,*	Style
		analysis, table, fig	General aboutness
AL1 vs. BNC	-		

Figure 2. Characteristics of keyword results from comparisons to different benchmark corpora



Similar specific benchmark corpus		Different specific benchmark corpus
	General benchmark corpus	
Lower mean keyness value		Higher mean keyness value
Shorter tail to keyness distribution		Longer tail to keyness distribution
Restricted keywords		Pervasive keywords
Less frequent words as keywords		More frequent words as keywords
Highlight specific differences between corpora		Highlight general differences between corpora
Highlight specific aboutness (e.g. technical terms)		Highlight general aboutness
Highlight aboutness		Highlight register and style differences (in addition to aboutness)

Appendix A. Journals from which articles in the AL1 target corpus were sourced

1. Applied linguistics
2. Assessing Writing
3. Computer Assisted Language Learning
4. English for Specific Purposes
5. English Language and Linguistics
6. International Journal of Applied Linguistics
7. International Journal of Lexicography
8. Journal of English for Academic Purposes

9. Journal of Second Language Writing
10. Journal of Sociolinguistics
11. Language Awareness
12. Language Learning
13. Language Learning & Technology
14. Language Learning and Development
15. Language Learning Journal
16. Linguistics and Education
17. ReCALL
18. Studies in Second Language Acquisition
19. System
20. TESOL Journal
21. TESOL Quarterly
22. The Modern Language Journal
23. World Englishes

Appendix B. The keyword lists for the five comparisons

AL1 vs. AL2

chat, stance, anxiety, lexical, his, comments, face, revisions, output, peer, translation, robinson, call, me, he, coherence, absent, eap, review, mandarin, focus, roles, class, messages, revision, criticism, model, capacity, negotiation, draft, informal, electronic, size, concepts, graduate, moves, conventional, academic, participation, expertise, expressions, I, generic, input, composing, depth, pronoun, composition, transfer, white, picture, resource, concept, advance, corpus, effective, retrieval, content, local, semester, staff, online, plural, taking, social, interpersonal, technology, o, environment, fit, uk, linked, de, you, rater, was

AL1 vs. SSH

L, language, learners, English, writing, text, vocabulary, words, word, learning, learner, lexical, task, feedback, students, proficiency, use, texts, linguistic, esl, written, speakers, was, tasks, native, reading, comments, grammar, discourse, corpus, writers, grammatical, sentence, oral, the,

meaning, knowledge, essay, comprehension, input, features, computer, test, were, sentences, acquisition, listening, target, essays, speaking, genre, participants, form, efl, her, communication, study, classroom, used, raters, instruction, speaker, content, chat, writer, output, rhetorical, she, peer, he, eap, languages, pronunciation, topic, rater, online, foreign, read, his, write, revisions, Chinese, stance, dictionary

AL1 vs. SCI

students, language, English, learners, writing, learning, words, teachers, task, vocabulary, word, reading, teacher, text, their, proficiency, learner, they, linguistic, tasks, feedback, lexical, l, texts, her, student, academic, discourse, speakers, what, teaching, classroom, instruction, use, comprehension, that, class, she, esl, he, comments, native, test, writers, written, speech, grammar, I, listening, it, participants, corpus, sentence, his, knowledge, grammatical, speaking, meaning, school, course, speaker, essay, sentences, group, question, efl, how, you, input, computer, peer, to, features, essays, explicit, classes, topic, not, languages, more, them, fluency, oral, interaction, courses, genre, questions

AL1 vs. MIC

students, language, of, the, learners, in, English, learning, study, writing, their, were, participants, as, l, task, research, teachers, by, results, vocabulary, knowledge, table, test, for, such, proficiency, feedback, text, however, learner, tasks, analysis, teacher, words, et, group, used, items, lexical, linguistic, scores, studies, between, e, al

AL1 vs. BNC

students, learners, language, learning, l, writing, English, study, participants, vocabulary, proficiency, task, feedback, test, learner, research, tasks, scores, text, lexical, knowledge, e, linguistic, teachers, esl, results, words, interaction, comprehension, reading, texts, student, native, teacher, academic, use, speakers, data, table, analysis, studies, discourse, classroom, et, items, word, instruction, strategies, al, in, group, online, were, peer, differences, findings, responses, raters, g, efl, corpus, their, target, content, focus, grammatical, teaching, comments, oral, significant, based, level, fluency, p, errors, context, grammar, tests, rater, variables, genre, each, explicit, participant, contexts, meaning, used, topic, groups, cognitive, questionnaire, input,

acquisition, between, communication, semester, writers, essay, features, using, frequency, of, essays, these, motivation, eap, listening, researchers, different, posttest, self, written, questions, accuracy, focused, rhetorical, discussion, types, communicative, class

The postal and email addresses of the authors

Punjaporn Pojanapunya
School of Liberal Arts, King Mongkut's University of Technology Thonburi (KMUTT)
126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand
punjaporn.poj@kmutt.ac.th

Richard Watson Todd
School of Liberal Arts, King Mongkut's University of Technology Thonburi (KMUTT)
126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand
irictodd@kmutt.ac.th