Clustering Keywords to Identify Concepts in Texts: An Analysis of Research Articles in Applied Linguistics

Punjaporn Pojanapunya
King Mongkut's University of Technology Thonburi, Thailand

Abstract

Keyword analysis is one of the most widely used methods in corpus linguistics. The method is used to generate keywords which provide an indication of concepts in texts or a corpus. Keyword analysis tools commonly produce resulting keywords presented as a list which rather poorly indicates what the corpus is about since it typically requires analysts' knowledge on conceptual associations between keywords. Therefore, common follow-up methods of keyword analysis are to examine concordances, collocational patterns, and some other patterns of associations between keywords and contexts. This study focuses on the association within a group of keywords by constructing a representation of a keyword list as keyword clusters. The keywords for an analysis were generated from two corpora; the target corpus was collected from research articles in applied linguistics and the comparative corpus was a collection of research in pure and applied sciences. The relationship between the top 30 keywords was identified using mutual information scores of all possible pairs of the keywords within a span of 20 and these scores were used as input for creating keyword clusters. The representations of the 30 keywords as a list and clusters are presented and discussed.

Keyword: *analysis, extended tree, collocates, cluster*

1. Introduction

Keyword analysis is one of the most widely used methods in the field of corpus linguistics. Keywords are words which occur with significantly high frequency in one corpus when compared to some appropriate normative corpus (Scott, 1997; Scott & Tribble, 2006). A keyword analysis function available in corpus analysis tools automatically compares frequencies of a word in the two corpora and typically outputs a list of keywords that are statistically more frequent in a target corpus than in a comparative corpus. These keywords are commonly used to provide an indication of what the corpus is about. An interpretation of the results heavily relies on intuition and requires knowledge of analysts on associations between the keywords to conceptualise the corpus description, since the keyword on its own is not distinctive enough to convey a meaning unless it is in particular combinations with other words or keywords (Cheng, 2009; Sinclair, 2005).

To characterize the corpus, examining a relationship between keywords with other words will help create meaning and concept. However, presenting keywords as a series of items or in a list format rather poorly indicates the relationship both among the keywords and between these keywords and the texts. Consequently, keywords as a list format normally perform as a starting point to signpost some interesting items for further detailed investigations.

Concepts in a corpus can be usually revealed by a closer look at keywords and their contexts via concordances (Mahlberg & McIntyre, 2011; Poole, 2016; Sealey, 2010), keywords and collocates of the keywords and their collocational patterns (Cheng, 2009; Loudermilk, 2007; Menon & Mukundan, 2010; Sinclair, 2005), keywords in clusters or a sequence of words (O' Donnell, Scott, Mahlberg, & Hoey, 2012), bond between keywords (Watson Todd, 2013), and a combination of these analyses (Römer & Wulff, 2010; Taylor, 2013). These follow-up investigations can be conducted by using existing functions of corpus tools, namely, concordance, collocation, and plots of keyword positions in a corpus to provide a broader context in which keywords are used. These investigations help facilitate understanding of what the texts are about by means of two broad patterns of the relationships between words: relationship between the keywords and other words in contexts (e.g. concordances, collocations, sequences of words) and between the keywords themselves (e.g. networks of keywords, categories or keywords).

2. Literature review

2.1 Keywords and contexts

Researchers normally have options for conducting a subsequence analysis to identify concepts in a corpus, for example, concordancing the keywords, observing collocational patterns of keywords, or observing the keyword in clusters of word sequences. First, the concordancing option gives all the instances of a keyword in their immediate linguistic context. We can sort the context to the left and to the right of the target keyword which allows us to observe regular patterns that emerge (Granger, 2012). We can also identify combinations of keywords and their collocates that co-occur frequently in the texts by using collocation option. The corpus tool usually provides the table of collocates of target words which are statistically important for the target word by statistical measures, e.g. z-score, mutual information (MI), and log-likelihood (LL) (see Barnbrook, Mason, and Krishnamurthy, 2013). Third, researchers frequently look at keywords in clusters retrieved from the concordance lines, e.g. two-word, three-word sequences which repeatedly occur in the corpus (Granger, 2012). These analyses of word relationships are useful because they treat words as a larger unit and provide visualizing patterns of lexical association and show how a word can acquire meaning in context (Xiao, 2015) which also reflect its actual use.

These further identifications of word associations provide insights into the use of keywords in contexts to reveal concepts. However, these methods are heavily qualitative in a way that we need to conceptualize the patterns of each keyword at a time. Consequently, several keyword studies ended up dealing with a few words which are usually intuitively selected from the keyword list. This pattern of the relationship may also lead us to paying attention to words in contexts which are possibly less important to the keywords.

This first pattern of associations probably includes the relationships between keywords. For example, examining concordances and collocational patterns of the keywords may reveal a presence of words which are also identified as key in the keyword list. However, the following section presents the other pattern of association which uncovers the relationships between keywords themselves more explicitly.

2.2 Relationship between keywords

Investigating relatedness between keywords is more closely related to the focus of this study. This relationship shows a link between words which are identified to be important to texts. In previous research, the relationship between keywords has been presented as categories or themes of keywords which express concepts in the texts or the corpus according to their research questions or some theoretical framework which is applied to their research (Culpeper, 2009; Fidler & Cvrcek, 2015; Gabrielatos & Baker, 2008; Gooberman-Hill, French, Dieppe, & Hawker, 2009; Lukac, 2015).

For example, top 11 keywords of the letters corpus in Lukac (2015) are *apostrophes*, *apostrophe*, *grammar*, *punctuation*, *spelling*, *possessive*, *plural*, *language*, *sir*, and *I*. While *sir* and *I* were considered as stylistic features of the letters in common, the rest were grouped together as the words related to grammar and punctuation which is the topic discussed in the letters corpus. In another study, keywords were classified into semantically related categories which are the keywords connected to cold war, collective possession, and ideological markers according to the analytical framework for socialist discourse analysis (Fidler & Cvrcek, 2015).

This thematizing method may be considered subjective because it has frequently been conducted by hand classification. In other words, the connections between keywords and other keywords heavily rely on researchers' intuition and experience in relatedness of concepts, and this analysis is probably not easy for novices. While the classification of the 11 keywords in the example given previously may seem straightforward, it requires the researcher's background knowledge of a letter genre to understand relationship between words. It can be even more difficult for novices to classify words when the corpus is specialized, for example, the corpus of socialist discourse in the other example (Fidler & Cvrcek, 2015). However, it is worth noting that intuition and background knowledge of texts and genres will always be useful in spotting words that denote key concepts even though they do not occur as frequently as other words.

This pattern of relationship, a keyword and other keywords, can be revealed by using more objective process. Scott and Tribble (2006) presented a contextual relationship between words that are key in the same texts and across texts in the corpus and proposed the idea of co-keys and associates which can be identified by word occurrence and dispersion in the corpus. These co-keys and associates can be grouped into clumps and represented as network of keywords. However, creating the network is likely to focus on associations of the particular words across different individual texts than the associations among different keywords in the corpus. Moreover, in terms of the methodology used to identify

this keyword linkage, researchers need to identify keywords, create a keyword database, and identify sets of keywords which are key in a range of texts. The whole process is methodologically demanding and still lacks corpus tools to assist, except WordSmith Tools (Scott, 2015). Therefore, the method has not been widely adopted in keyword analysis yet and the majority of keyword-based studies are likely to rely on more accessible concordancing and collocation options in subsequence analyses.

In another study, connections between keywords were identified to provide conceptual associations between keywords that reveal knowledge and concepts in texts (Watson Todd, 2013). The relationships between keywords were identified based on information of a collocation statistic and this information suggests pairs of keywords which form conceptual associations, for example, *autographical* and *memories* (pair 1) and *bias* and *biases* (pair 2). These pairs could also be intuitively considered as concepts expressed through multi-word units and concepts expressed through morphological variants. However, the collocation statistic is clearly useful in identifying concepts which appear less likely to be associated, such as *brain* and *milliseconds*.

The methodology for creating bonds between words also requires an extra stage for data preparation. However, it is more practical than a keyword network of Scott and Tribble because it requires only collocation scores which can suggest relationships between keywords and uses the scores as input for generating keyword clusters. Therefore, the method of identifying conceptual associations was adopted in this study to construct a representation of a list of keywords as clusters.

3. Research purpose

This study produced a keyword list and took a more objective approach which is used to identify links between pairs of the keywords to create keyword clusters. Shifting from form to meaning, keywords represented as a list and as clusters will be examined on how they provide indications to what the corpus is about.

4. Research methodology

4.1 Data

Previous research has used different types of corpus data as a target corpus and a comparative corpus. While a target corpus is commonly specialized rather than general, a comparative corpus could be a general corpus, e.g. British National Corpus, or a corpus similar to the target corpus but different in topics, e.g. history and marketing research articles (Malavasi & Mazzi, 2010).

To produce a list of keywords and identify their relationship to highlight concepts in a target corpus, this study requires sample data or keywords which fit the following criteria. Texts should be familiar to the researcher to allow classifying keywords either based on background knowledge or collocation scores. Keywords should express some certain topics or provide indications for concepts in the corpus clearly. They should be informative in

themselves so that they can indicate concepts in the corpus when they are represented either as a list or as a group.

Data used to produce keywords were research articles from two distinct academic domains, one of which is a corpus of a specific discipline: applied linguistics (AL) and pure and applied sciences (SCI). From a broader view, they were from humanities and science which have used different linguistic resources in the creation of specialised knowledge (Boulton, Carter-Thomas, & Rowley-Jolivet, 2012; Hyland, 2012).

Research articles in applied linguistics, as discipline-specific texts, were selected because concepts in a corpus are much more clearly attained when a target corpus is domain-specific (Scott & Tribble, 2006), or even topic-specific. The AL corpus consisted of research articles from second language acquisition, lexicography, discourse analysis, translation, and corpus linguistics covering qualitative and quantitative research from data sources which are available to the researcher. Although the AL corpus was collected from various sub-disciplines, keywords reported in this paper largely depended on topics of the selected sub-disciplines which may be different from a wider range of topics in applied linguistics. The SCI corpus was from major journals in mechanical engineering, microbiology, biotechnology, and nursing representing applied sciences, pure science, technological science, and clinical science, respectively.

Since population size of research articles in these two disciplines is inaccessible, this study took a statistical perspective which returned 385 samples required for unknown population or a very large population size to have 95% confidence level with a 5% error (Smith, 2013). Therefore, each of the corpora in this study consisted of 400 research articles, 2.7 million words in the AL and 2.1 million words in the SCI. Keywords of the AL compared to the SCI were generated in two forms: a list of keywords and their clusters.

4.2 Procedures and analysis

A keyword lists and clusters of keywords were created by the following procedures. First, a list of keywords of the comparison between the AL as the target corpus and the SCI as the comparative corpus was generated by using a Keyword List function available in AntConc (Anthony, 2014). A default output presents a list of keywords ranked by a log-likelihood statistic (LL) (see Dunning, 1993; Rayson, 2008).

Second, a graphical representation of associations between the top 30 keywords was constructed by using Extree software which clusters keywords based on relatedness between words (Corter & Tversky, 1986) using a graphical representation feature which is not available in existing corpus analysis tools. In this study, information about the relatedness between keywords is provided by a collocation statistic, a mutual information (MI) statistic which indicates strength of relationship between each pair of the keywords (see 4.3 Collocation statistics).

A threshold level above which words are considered to be key is arbitrary. This study used a certain number of the highest ranking keywords, i.e. the top n words, the most

common method to set the threshold. This method is less problematic than other criteria, such as setting a minimum LL value or a minimum p-value which are affected by corpus sizes (Pojanapunya & Watson Todd, 2016) and could still return too many keywords for detailed analysis. The top 30 was set as the cut-off point by identifying themes of different numbers of keywords and seeing that keywords outside this threshold did not add any new theme. These 30 keywords also allow the keyword classification to have many clusters of keywords which show different forms of their relationship.

The relationship of each of the keywords was identified by the MI scores for every other words located on either side of the target keywords with a span of 10. This span of 10 words to the left and to the right should be appropriate for identifying association and concepts because it covers the output which tends to present multiword keywords (e.g. a span of 2 or 3). The span should also cover concepts which can be held by working memory of approximately 7 lexical concepts at a time (Watson Todd, 2013). Examples of the MI scores of all pairs between the top 10 keywords are presented in Table 1.

Table 1 MI scores of each of all pairs of the 10 keywords

No	keywords	1	2	3	4	5	6	7	8	9
		students	language	english	learners	writing	learning	words	teachers	task
1	students									
2	language	4.55								
3	english	4.84	5.34							
4	learners	2.08	5.31	4.85						
5	writing	5.24	4.13	4.63	3.59					
6	learning	4.58	6.27	4.63	5.38	3.24				
7	words	3.87	3.17	4.39	4.42	3.74	4.05			
8	teachers	5.63	4.70	5.04	4.18	4.60	4.05	2.24		
9	task	3.83	3.88	2.94	4.95	5.05	4.28	3.92	2.66	
10	vocabulary	4.25	4.20	4.03	5.13	4.06	6.46	5.62	3.21	4.40

Finally, the keywords represented as a list and as clusters of keywords were examined for how they indicate what the AL is about.

4.3 Collocation statistics

Different collocational statistics highlight different characteristics of results. The most commonly used collocation statistics are z-test and mutual information (MI). Z-test measures the confidence which claims that there is some association between words. This means that the more occurrence of each keyword in the text, the more likely other keywords will also be found collocationally linked. So, their high values tend to highlight high-frequency pairs (Scott & Tribble, 2006; Xiao, 2015). It is possible that words with collocational relationship identified by these tests may be because of the high number of co-occurrence or the high frequency of each word.

Strength of relationship between each pair of the keywords can be identified by calculating a mutual information score. The higher the MI, the stronger the association between two items is (Xiao, 2015). MI can substantially overestimate the significance of infrequent words or tend to include low frequency words.

This study only uses the statistical values to identify relatedness between the 30 selected keywords and focuses on the strength of the associations between them. It is not the issue whether there are many different types of words competing for the position to become collocates of the target keywords. In other words, the study focuses on the strength of relatedness between keywords rather than the relatedness identified because of the high frequency of collocates or the frequent collocations. Therefore, the relationship between the top 30 keywords in this study was identified by MI scores with the higher MI of each pair of keywords suggesting a strong association between the two items.

5. Results

This section presents concepts of what the AL is about which are uncovered by the keywords represented in two main patterns: the list of keywords and clusters of keywords based on collocation scores. The keyword clusters provided by the Extree are given in two output formats, namely, the primary and the secondary classifications.

5.1 A list of keywords

Figure 1 shows a series of the top 25 keywords, the default output, captured from an interface of AntConc (Anthony, 2014). They are words which occur significantly more frequent in applied linguistics than in pure and applied sciences.

All 30 keywords investigated in this study are:

1 students	*13 teacher*	*25 her*
2 language	*14 text*	*26 student*
3 English	*15 their*	*27 academic*
4 learners¡	*16 proficiency*	*28 discourse*
5 writing	*17 they*	*29 speakers*
6 learning	*18 learner*	*30 what*
7 words	*19 linguistic*	
8 teachers	*20 tasks*	
9 task	*21 feedback*	
10 vocabulary	*22 lexical*	
11 word	*23 l*	
12 reading	*24 texts*	

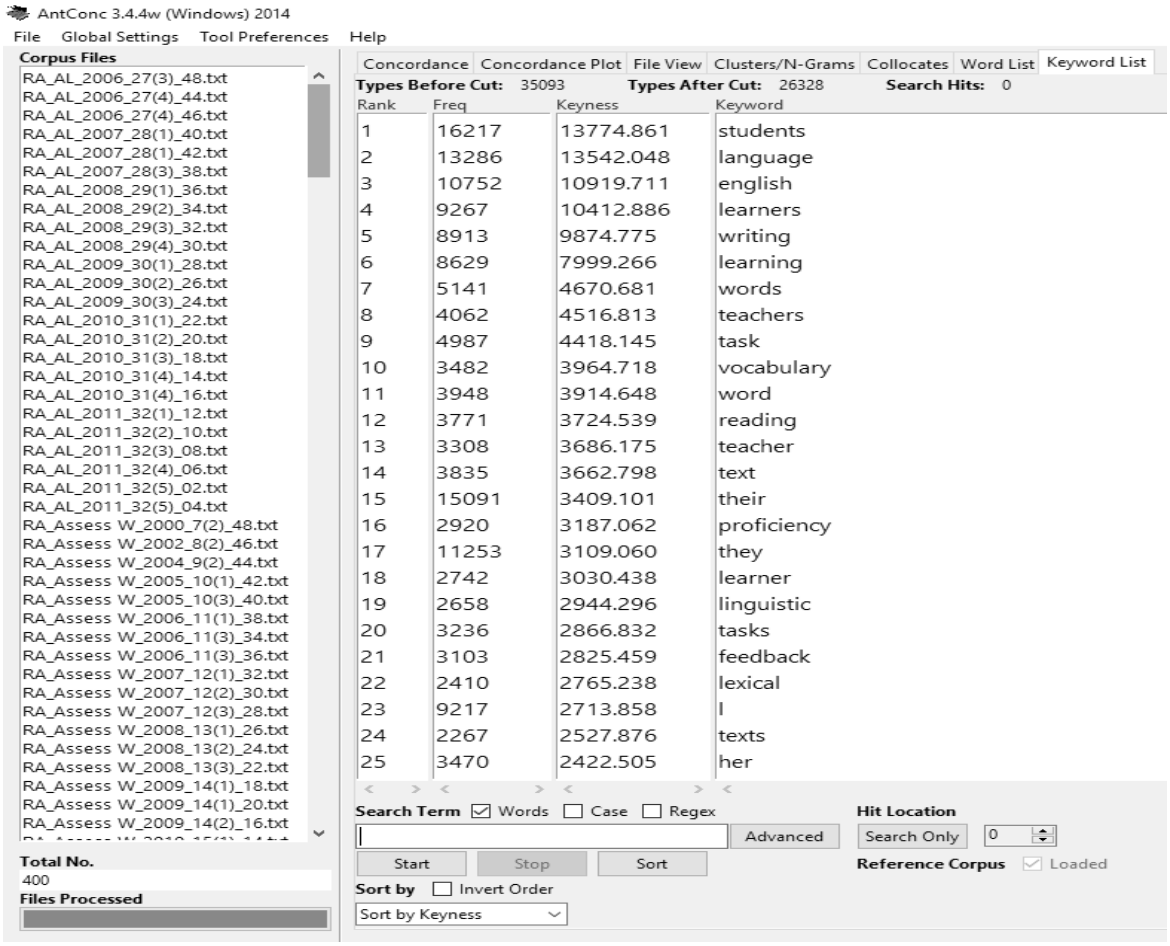

Figure 1 Top 25 keywords of AL vs. SCI

As can be seen in Figure 1, keywords presented as a list do not provide information about their associations which is required to support processing in interpretation, especially for the purpose of identifying what the target corpus is about. These keywords can be further examined for concordances, collocational patterns, or clusters to reveal conceptual associations between these words and other words in the corpus. Their relationship could be identified, e.g. by categorising them into groups.

There are many possibilities for categorising these 30 keywords into groups depending on type of texts collected in a corpus, corpus content, a theoretical framework, and most importantly, research purposes or research questions. Different categories used for categorising words typically produce different groups of keywords. Two of several possibilities for word classification are given as examples. First, to describe the concepts included in the corpus of applied linguistics research, we could categorise some of the top 30 keywords intuitively into four sub-categories.

- Words related to language learning: *language*, *English*, *linguistic*, *lexical*, *learning*, *proficiency*, *L*(1,2)
- Words related to objects of research: *text(s)*, *discourse*, *word(s)*, *vocabulary*
- Words related to classroom activities: *writing*, *reading*, *task(s)*, *feedback*
- Words related to research participants: *student(s)*, *learner(s)*, *teacher(s)*, *speakers*, *her*, *they*, *their*

Second, the keywords also correspond to the 3 dimensions of register (Collins & Hollo, 2010).

- Field or the text's subject matter: *language*, *English*, *linguistic*, *lexical*, *learning*, *proficiency*, *L*(1,2), *text(s)*, *discourse*, *word(s)*, *vocabulary*, *task(s)*, *feedback*
- Tenor or the participants in a given situation: *student(s)*, *learner(s)*, *teacher(s)*, *speakers*, *her*, *they*, *their*

We can see that the keywords can be categorized differently by using the different framework into different numbers of sub-categories according to the researcher's consideration.

5.2 Clusters of Keywords

Having assigned the MI-scores for each pair of keywords within a span of ten to the left and to the right to the Extree program, the relationship between the top 30 keywords of applied linguistics research compared to pure and applied sciences are constructed as a tree diagram (Figure 2) and clusters with uppercase letters used to indicate the clusters (Figure 3).

The two output formats which are clearly different from the list (in Figure 1) seem readily interpretable and, unsurprisingly perhaps, reflect the topics discussed in research articles in applied linguistics field. In Figure 2, many branches from the root show different levels of classifications. It is not immediately obvious how many clusters in the tree of keywords are represented. Here, I will focus on the relationships which represent the most similarities between pair or groups of keywords.

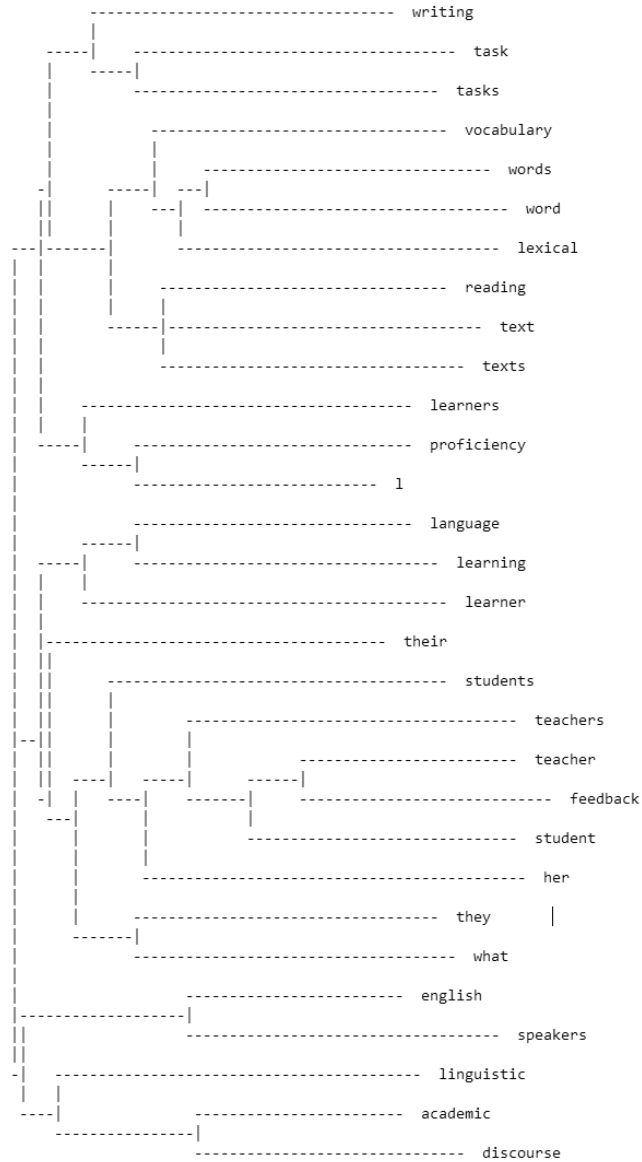

Figure 2 Extended tree representation of the top 30 keywords

5.3 Primary classification

In general, the primary classification tends to reflect intuitive associations between words. Natural relations found in the tree are akin to classifications of meaning which are meaning-based, position-based, and form-based associations (Fitzpatrick, 2009).

First, the form-based association is unsurprising relation of singular and plural forms, for example, *task – tasks*, *word – words*, *text – texts*, *teacher – teachers*, *and student – students*. Secondly, the position-based association signals conceptual associations or the type of relationship which is a lexical linkage by the co-occurrences suggesting the relatedness of topics in the corpus. For example, the primary classification matches *writing – task – tasks*, *reading – text – texts*, *l – proficiency*, *language – learning*, *teacher – teachers – feedback*, *English – speakers*, *academic – discourse*. Moreover, some of these are the co-occurrences of two or more keywords within immediate or near context, e.g. within a span of 2. These adjacent keywords signal multi-word keyword units which represent established sub-fields under applied linguistics area. Third, the other keywords, *vocabulary* – (*word – words*) – *lexical* were clustered together because they are synonymous words associated based on their meanings.

5.4 Secondary classification

In the previous section, primary classification includes meaning-based, form-based, and position-based associations. These associations include keyword collocates which co-occurred within immediate or near context which are likely to be multi-term collocations. Compared to the primary classification, the secondary classification seems to reflect the meaning-based pattern which requires closer consideration to interpret.

```
final set of marked features:

feature  objects sharing feature
-------  -----------------------

   C     [ language, teachers, english, speakers, ]

   D     [ teacher, student, academic, ]

   E     [ lexical, linguistic, discourse, ]

   H     [ student, academic, words, word, ]

   I     [ discourse, linguistic, speakers, ]

   N     [ feedback, learner, ]

   O     [ learning, vocabulary, ]

   U     [ teacher, feedback, learner, ]

   X     [ feedback, learners, ]

   Z     [ feedback, linguistic, ]

   A     [ teacher, feedback, student, learner, ]

   D     [ speakers, l, ]

   G     [ lexical, proficiency, ]

   J     [ academic, writing, ]

   K     [ teachers, teacher, feedback, her, student, learner, ]
```

Figure 3 Secondary classification of the top 30 keywords

The primary classification matches *english* and *speakers* (in Figure 2), whereas the secondary classification in Figure 3 matches *speakers* with the words *linguistic* and *discourse* (cluster I) with closer links than *english* as the topical association.

The other example, *proficiency* was matched with *L* based on primary classification as shown in Figure 2, while *proficiency* conceptually links to *lexical* in cluster G. Next, primary classification of *academic* and *discourse*. On the other hand, the secondary classification combines *discourse* with *linguistic* and *speakers* in cluster I.

Also, there are clusters of keywords with a close relationship in many clusters. While primary classification matches *teacher* and *feedback*, secondary grouping provides more complex relationship of *teacher* and *feedback* with other key concepts, i.e. in clusters D, N, U, X, Z, A, K. The secondary categories show the link between *teacher* and *student*, *academic*, *learner*, *teachers*, *her* (clusters D, U, A, K) and between *feedback* and *learner*, *learners*, *linguistic*, *student*, *her* (clusters N, U, X, Z, A, K).

Without this graphical representation, we only have individual keywords ordered as a list and normally work with a long list of concordances of each of the keywords for further details. For example, there were 3,308 concordances for *teacher* in the AL corpus. Suggested by cluster D [teacher, student, academic], we may pay attention to concordances of *teacher* with *academic* and *student*. Out of 3,380 concordance lines of *teacher*, we can now focus on, e.g. 40 concordances of *teacher* occurred with *academic* and 268 concordances of *teacher* occurred with *student* to further investigate these terms. Focusing on relationship between *teacher* and *student*, the following concordances show that the teacher and student were key participants in research in this context.

, a semi-structured background interview with the	*teacher*	and each *student* was conducted. The interv
data in this study were collected using	*teacher*	and *student* interviews and classroom obser
feedback with the scores given on the	*teacher*	and *student* questionnaires, statistical ad
lly, the scheduled individual conferences between	*teacher*	and *student* were observed, audiotaped, and
Peer Feedback Through Blogs:	*Student*	and *teacher* perceptions in an advanced Ger
. Teacher and student questionnaires For both the	*student*	and *teacher* questionnaires, numerical valu

The concordances also show relationship between *teacher* and *student* when each of the terms takes key roles, such as *teacher* 'answer', 'ask', 'assign', 'edit and proofread', as shown below, while *student* is concerned with 'beliefs on teacher behavior', 'teacher assessment', and 'perceptions of teacher feedback'.

asked a question about it, when the	*teacher*	answered the *student*'s question it was
by an external agent such as the	*teacher*	asking the *student* to use the word
second session with the following steps. The	*teacher*	1. Assigned each *student* with a number bet
), serves as a vehicle through which the	*teacher*	calculates *student* performance and makes j
ssment criteria when assigning writing tasks, and	*teacher*	editing and proofreading of *student* writin
to compare the value of peer and	*teacher*	feedback on *student* writing by simply coun
period. In addition, once a month the	*teacher*	met with each student in a private
as a device best suited to help	*teacher*	evaluation of student performance rather t

6. Discussion

Although a keyword list, the default output of keyword analysis carried out by commonly used keyword analysis tools provides no information about association between keywords, it allows a researcher to choose and classify the keywords according to different factors, such as theoretical framework, research questions, researchers' consideration supported by detailed investigations of concordances and collocational information of the keywords. This typical practice requires the researcher's expertise in the subsequence analyses and interpretation.

The graphical representation of keyword associations provides more rigorous method of classifications of the keywords. Although it requires an additional stage for preparing input of information about keyword association, both primary and secondary clusters of the keywords bring about links between concepts of the corpus more explicitly. Therefore, the output seems to be more easily understandable and readily interpretable. Classifying keywords in this way helps minimize hand classification of keywords and may help direct researchers to some certain theoretical framework which can facilitate interpretation. This is especially helpful for novice researchers.

This method can be objectively established. Its results are informative about the corpus organized in different levels of classification. While there is no clear separation of an application of the primary and secondary clusters in research contexts, researchers have options to choose which of those are simpler and more interpretable (Corter & Tversky, 1986). However, the primary and secondary classifications of the keywords in this study are likely to have clear distinction of application. The results show that the primary clusters direct the researchers to the topical-related keywords reflecting fairly specific concepts, mainly form-based and position-based associations. On the other hand, the concepts provided by secondary classification are likely to reflect broader conceptual associations between words.

Some methodological issues are worth discussing. Apart from a graphical representation of word associations provided by the Extree, some corpus tools have already integrated visual displays of the corpus results, for example, word dispersion plot in AntConc (Anthony, 2014) and WordSmith Tools (Scott, 2015), word cloud and key word cloud in Wmatrix (Rayson, 2009). These representations can be created from the primary source of information of a specific word, e.g. word position in texts, frequency, and relative frequency. However, there is no tool so far which automatically produces graphical representation of words from the secondary source of information such as MI or z-scores except Graphcoll (Brezina, McEnery, & Wattam, 2015).

Recently, GraphColl has been developed to create collocation networks presenting words and other words with a collocational relationship. The program is greatly flexible in that users can identify collocations with criteria they desire. For example, they can choose type collocation statistics, statistic cut-off value, distance between a target word and its collocates, and minimum frequency of collocations and others. The program presents

a graphical view of the target words assigned by the users in the same way that the 30 keywords were examined in this study. However, the graphical view not only shows the 30 target keywords, but also any other collocates of them. So, the users cannot see the links between the selected target words very easily.

The results of this study suggest that having a program to create diagrams or figures showing the relationship between keywords visually or even adding visual features to the existing corpus tools would be useful.

In conclusion, different representation of keywords benefits different applications. The method presented in this study is not the replacement, but provides more options for researchers to have different representation that facilitates interpretation. Similar to this study, the method could be used as a platform for a better understanding of concept formation and representation in a corpus of texts from other disciplines or genres through keywords.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer software]. Retrieved from http://www.laurenceanthony.net/software/antconc/

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). *Collocation: Applications and implications*. London, England: Palgrave Macmillan.

Boulton, A., Carter-Thomas, S., & Rowley-Jolivet, E. (2012). Issues in corpus-informed research and learning in ESP. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and applications* (pp. 1-14). Amsterdam, the Netherlands: John Benjamins.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics, 20*(2), 139-173.

Cheng, W. (2009). Income/interest/net: Using internal criteria to determine the aboutness of a text In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 157-178). Amsterdam, the Netherlands: John Benjamins.

Collins, P., & Hollo, C. (2009). *English grammar: An introduction.* London, England: Palgrave Macmillan.

Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika, 15*(3), 429-451.

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.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Fidler, M., & Cvrcek, V. (2015). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics, 23*(2), 197-239.

Fitzpatrick, T. (2009). Word association profiles in a first and second language: Puzzles and problems. In T. Fitzpatrick & A. Barfield (Eds.), *Lexical processing in second language learners* (pp. 38-52). Clevedon, England: Multilingual Matters.

Gabrielatos, C., & Baker, P. (2008). Fleeing, sneaking, flooding: A corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press, 1996-2005. *Journal of English linguistics, 36*(1), 5-38.

Gooberman-Hill, R., French, M., Dieppe, P., & Hawker, G. (2009). Expressing pain and fatigue: A new method of analysis to explore differences in osteoarthritis experience. *Arthritis Care & Research, 61*(3), 353-360.

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 7-29). London, England: Basil Blackwell.

Hyland, K. (2012). *Disciplinary identities: Individuality and community in academic discourse*. Cambridge, England: Cambridge University Press.

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

Lukac, M. (2015). Linguistic prescriptivism in letters to the editor. *Journal of Multilingual and Multicultural Development, 37*(3), 321-333.

Mahlberg, M., & McIntyre, D. (2011). A case for corpus stylistics: Ian Fleming's Casino Royale. *English Text Construction, 4*(2), 204-227.

Malavasi, D., & Mazzi, D. (2010). History v. marketing: Keywords as a clue to disciplinary epistemology. In M. Bondi & M. Scott (Eds.), *Keyness in texts* (pp. 169-184). Amsterdam, the Netherlands: John Benjamins.

Menon, S., & Mukundan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. *Pertanika Journal of Social Sciences & Humanities, 18*(2), 241-258.

O' Donnell, M. B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory, 8*(1), 73-101.

Pojanapunya, P., & Watson Todd, R. (2016). Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*. Retrieved from https://www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2015-0030/cllt-2015-0030.xml?format=INT

Poole, R. (2016). Good times, bad times a keyword analysis of letters to shareholders of two fortune 500 banking institutions. *International Journal of Business Communication, 53*(1), 55-73.

Rayson, P. (2008). Log-likelihood and effect size calculator. Retrieved from http://ucrel.lancs.ac.uk/llwizard.html

Rayson, P. (2009). Wmatrix: a web-based corpus processing environment [Computer Software]. Retrieved from http://ucrel.lancs.ac.uk/wmatrix/

Römer, U., & Wulff, S. (2010). Applying corpus methods to written academic texts: Explorations of MICUSP. *Journal of Writing Research, 2*(2), 99-127.

Scott, M. (1997). PC analysis of key words – and key key words. *System, 25*(2). 233-245.
Scott, M. (2015). WordSmith Tools (Version 6.0) [Computer Software]. Oxford, England: Oxford University Press.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam, the Netherlands: John Benjamins.
Sealey, A. (2010). Probabilities and surprises: A realist approach to identifying linguistic and social patterns, with reference to an oral history corpus. *Applied Linguistics, 31*(2), 215-235.
Sinclair, J. M. (2005). *Document Relativity*. Tuscany, Italy: Tuscan Word Centre.
Smith, S. (2013). Determining sample size: How to ensure you get the correct sample size. Retrieved from https://www.qualtrics.com/blog/determining-sample-size/
Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora, 8*(1), 81-113.
Watson Todd, R. (2013). Identifying new knowledge in texts through corpus analysis. *International Journal of Language Studies, 7*(4), 57-76.
Xiao, R. (2015). Collocation. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 106-124). Cambridge, England: Cambridge University Press.

Author

Punjaporn Pojanapunya is a researcher in Applied Linguistics at King Mongkut's University of Technology Thonburi. Her research interests include keyword analysis, corpus linguistics, and corpus-based discourse analysis. Her recent publication (with Richard Watson Todd) is entitled "Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis".
punjaporn.poj@kmutt.ac.th