Calculating and Displaying Key Labels: The texts, sections, authors and neighbourhoods where words and collocations are likely to be prominent

Abstract

Corpora are usually not only made up of words, sentences and plain texts; they usually also have metadata, background information and structural features which can be used to filter searches or provide additional information about the context of specific concordance lines. This paper presents a new approach which uses the information about the texts in which words and collocations occur, generating clouds and tables of what are called Key Labels. The procedure can be likened to looking at key words (Scott, 1997; Scott & Tribble, 2006) from the opposite starting point: beginning with a word of interest and exploring the features of texts and the parts of text in which it occurs. The paper explains the background to the procedure, how it is carried out, and how these Key Labels are integrated into *The Prime Machine* corpus tool for English language learning.

Key words: *keyness, metadata, dispersion, semantic associations*

1. Introduction

As well as the words and sentences making up the language sample, corpus texts usually contain other information about the text or sections of the text. These metadata often provide details of the source including information about the authors of the texts or other bibliographical information, as well as how each file fits into the corpus design: for instance which sub-genre it is intended to represent. They may be part of the header of the text file, or they may be tags in an XML tree. When looking at a concordance line a user may wish to know more information about its source and concordancers usually offer some means of displaying this information. A researcher who has a specific sub-set of documents or text types in mind may also want to use these metadata as a means of filtering the results; the query could specify, for example, that only concordance lines where the text type is identified as spoken data should be included. This kind of information could also be used to split a larger corpus into sub-corpora, facilitating comparisons between the genres it contains. Comparisons may be based on searches within filtered results, or could be made through

other procedures such as key word analysis which measure relative differences between corpora or sub-corpora. The starting point for these procedures is a corpus or a collection of texts from a corpus, and these procedures give concordance lines or lists of key words as a result. This paper introduces a new procedure which begins with interest in a specific word or collocation in a corpus, and draws on the metadata and tags to give the user what are called Key Labels[1]. *The Prime Machine*[2] is a corpus tool for English language learners, and provides novel ways to present source information for each concordance line on concordance cards (Jeaco, 2017a, 2017b). The card for the currently selected line on the Key Words in Context display appears on a panel to the right, showing collocations of the node within a 4 word window, the subsection of the corpus, source information and an extended context including one sentence before and after the sentence containing the node. The software also incorporates Key Labels as described here for its online corpora[3]. This paper presents the Key Label approach which uses log-likelihood contingency tables with Bayes Factors to create a list of key metadata, section and neighbourhood labels which are then displayed using a tag cloud or table. After introducing related work including key words and dispersion, the paper presents the method for calculating and displaying Key Label data at four levels: Text, Section, Producer and Neighbourhood. It ends with some ways these data could help language learners and their teachers, as well as possible applications of the method for linguistic research.

2. Literature Review and Related Work

The fundamental corpus method on which the procedure described in this paper relies is that of key word analysis (Scott, 1997; Scott & Tribble, 2006). The key word procedure is designed to examine the relative frequency of an item in a study corpus and to compare it to the relative frequency in a reference corpus, with a view to providing a list of words which are likely to be prominent in a text or collection of texts. It is a procedure which has been the mainstay of a range of corpus tools, allowing users to start with a text or collection of texts and obtain words which could be of interest for further exploration. Software such as *WordSmith Tools* (Scott, 2010), *AntConc* (Anthony, 2004), *WMatrix* (Rayson, 2008) and

[1] In earlier versions of *The Prime Machine* these were called Key Tags, but they have been renamed Key Labels to avoid confusion with tags such as Part-of-Speech tags.
[2] See www.theprimemachine.com for details.
[3] The procedure is currently only available for the pre-processed corpora on the server, not for DIY corpora.

LancsBox (Brezina, McEnery, & Wattam, 2015) include this function, as do specialized web-based corpus tools such as *CLiC* (Mahlberg, Stockwell, Joode, Smith, & O'Donnell, 2016). The procedure described in this paper is essentially looking at keyness from the other direction: starting with a word and considering how a corpus could be re-organised in order for this to become a keyword.

Although connecting metadata to specific instances by using keyness is fairly innovative, there are other measures and processes which look at related features of language. The Key Domains feature of *WMatrix* (Rayson, 2008) can be used to show semantic tags which are key in a corpus, and by splitting a corpus into sub-corpora, key words and tags can be calculated at the text level. Some work been done using equally sized strips of text and comparing relative frequencies of words within one strip against the others (Liang, 2015). Liang's software is able to automatically divide each of the texts in a corpus into strips and then to use key word statistics to show which words are key in each section of the text. The idea of looking at where words tend to occur is also related closely to the well-established concept of dispersion (Gries, 2010; Oakes, 1998). One way of showing the user how words or phrases are spread throughout texts and a corpus is through dispersion plots (Scott & Tribble, 2006). *LancsBox* (Brezina, McEnery, & Wattam, 2015) shows dispersion information at the text level through its Whelk and dispersion tools, and shows collocation networks of words or tags. Other studies have explored the centrality and connectivity of specific nodes across chapters of book (Phillips, 1985), how repetition of lexis forms part of cohesion (Hoey, 1991), and the way in which vocabulary across wide text windows can help identify topic divisions in texts (Biber, Connor, & Upton, 2007). Attempts have also been made to search key word databases for words with a specific pragmatic function in order to see whether their role in the text can be automatically identified (Scott, 2000). There have also been recent developments in topic modelling of corpora through machine learning (Murakami, Thompson, Hunston, & Vajn, 2017). While dispersion calculations and key words on strips do provide some insights, corpora often have tags and metadata which could provide much more detail. It seems that in other concordancers these metadata are currently only used to filter searches rather than to examine the distribution of specific words and phrases under investigation.

This paper opens up the potential for a new method for corpus consultation which provides users with information about the typical contexts in which a word or collocation occurs. Using the frequencies inside and outside XML nodes, the system pre-calculates the typical environments so that a user searching for a word or phrase can instantly see what are called

Key Labels: the XML tags and some other tags which are statistically significant for the search term.

3. Method

The log-likelihood contingency table which is used to rank and test the significance of the relationships [4] is given in Table 1. This contingency table is formed by comparing the number of instances of a word within a text or section which is mapped to a metadata tag against the number of times the word occurs outside this context. The log-likelihood formula also balances this against the overall number of other words within the same context. A similar procedure is used to calculate Key Labels for collocations, where the frequencies are multiplied by the length in words of the collocations, since each instance of a two word collocation occurring within a metadata tag would account for two words from the total word count for that tag.

[Table 1]

As with log-likelihood collocations and priming features in *The Prime Machine*, following Wilson's (2013) application of Bayes Factors to key item analysis, the log-likelihood formula and Bayes Factors are used in combination to calculate scores and degrees of evidence (Author, 2015), and only items which occur proportionally more often inside the tags than outside the tags are stored[5].

In *The Prime Machine*, the overall aim of this new concordance software feature is to provide additional information about the distribution of words and collocations to unsophisticated users of the system. The clouds and tables of results are to be displayed in tabs alongside concordance lines and other summary data as a means of enriching the contextual clues available. If corpora include metadata that give indications of the function of specific

[4] While there have been some discussions about how best to rank results from key word analyses (Gabrielatos & Marchi, 2012), the intuition here is that the measure should take into account the relationships between the sizes of the sub-corpora, the overall frequency of the word or collocation in the corpus, and the overall size of the corpus, with larger corpora requiring a higher BIC (Bayesian Information Criterion) to qualify for inclusion in the list. The Log-likelihood and Bayes Factor combination are sensitive to all these relationships in ways which %DIFF and other normalized frequency based measures are not.

[5] During the early development of this approach, tendencies for words or collocations not to occur inside tags were also measured. However, the updated *SQL* scripts which generate these results no longer include these negative relationships. Although tendencies for words not to occur inside tags may be of interest to a linguist, results showing both positive and negative relationships could be confusing for a language learner and the focus in the software is on positive relationships.

sections of text, it could also be considered as a possible way to approach the automatic identification of what Hoey (2005) calls pragmatic association.

During the development of this procedure, some consideration had to be given as to how to help users interpret the scope of the Key Labels, and also how they should interpret 'thin' or 'empty' clouds. One way in which support for such interpretation is provided is through a range indicator which appears at the top of each cloud panel. The range indicator has a start and end arrow head showing the proportion of instances which are accounted for by the tag with highest frequency, leading up to the proportion accounted for by the combined frequencies of all the tags visible in the cloud. These values also appear as percentages above the range indicator, with the frequency of the search term also provided. The lower value is intended to provide the most cautious interpretation of the cloud, showing the smallest possible coverage of the environments in which the search term would be occurring if all the other tags in the cloud were representing exactly the same set of concordance lines as the most frequent tag. In such an extreme case, the labels would perhaps essentially fit into a single hierarchy or the multiple labels might map homogeneously to the same concordance lines. At the other extreme, the upper indicator shows the proportion of concordance lines represented if all the labels account for unique instances of the search term in different environments with no overlapping.

4. Examples of Key Labels

Key Labels are calculated and displayed in four groups: Text Level, Section Level, Producer Level and Neighbourhood Level.

4.1. Text Level

The Text Level Key Labels can provide some indication of the tendency of a word or collocation to be used in texts from a particular set of sources in a corpus, or of texts of a particular type. Since the metadata mappings rely primarily on the tags which are provided in the corpus file headers, and also on decisions made during the refactoring process, it is not possible to stipulate whether these will be indicative of genre, register, style or the corpus

sampling process[6]. Essentially, as with the other Key Labels, the user should try to keep the following two questions in mind:

- Do these results suggest that the word or collocation is associated with a particular kind of text type?
- Do these results suggest that the texts which were chosen for the corpus are suitable for my purposes?

From Version 3 of *The Prime Machine*, as well as metadata drawn from the raw corpus texts and corpus documentation, Text Level Key Labels may also include indications regarding tendencies for the word or collocation to occur in texts with a relatively high or low score on the six dimensions from Biber's original work on Multidimensional Analysis (Biber, 1991). Multi-Dimensional Analysis is a field of corpus linguistics of itself, with many studies using Biber's original methodology and applying it to collections of texts of different genres, as well as to collections of similar texts in one domain or genre (Biber & Conrad, 2009; Friginal, 2013). With the release of *MAT* (Nini, 2014), it is now possible to tag plain text files using an automated tagging system which, as Nini has shown, has very similar results to those of Biber's original work. In *The Prime Machine*, full MD analysis is not attempted, but during the refactoring process, results from *MAT* are imported into the database and additional tags are added to texts which have relatively high or low scores on each of the six dimensions. While researchers following full Multidimensional Analysis procedures tend to interpret dimension scores in different ways, in *The Prime Machine*, ±3.71899 and ±5.199336 were used to determine whether a dimension label would be added to the text and whether additional strength would also be indicated in this label. These figures were chosen using the NORMDIST function and Goal Seek operations in Microsoft Excel, to give expected z-score standard deviations required for 99.99% and 99.99999% of the data.

A pair of examples for Text Level Key Labels is shown in Figure 1 and Figure 2. The Key Label cloud of text metadata for *therefore* in the *BNC* provides (in descending order of keyness) 'ACADEMIC', 'NON-ACADEMIC', 'W ac:polit law edu', 'Written Text', 'W commerce' and some general publishing or sampling information. The MAT tags include 'Abstract Information', 'Context-Independent' and 'Informational'. As expected, this suggests strongly an association with written texts. The same search for *thus* gives

[6] See Jeaco (2015) for more details about the corpus refactoring processes

‘ACADEMIC’, ‘Written Text’, ‘NON-ACADEMIC’, ‘W ac:soc science’, ‘W commerce’ and the publishing information, showing an even stronger tendency for use in Written Text. In both these sets of results, the range indicator at the top is relatively narrow and it is located at the right-most side of the display, showing there could be some overlap, but that these labels definitely cover the vast majority of the results. This can easily explained for *thus* because while ‘W ac:soc science’ is a subset of ‘ACADEMIC’ texts, and ‘ACADEMIC’ is a subset of ‘Written Text’, taken separately or together these labels account for all or most of the data.

[Figure 1]

[Figure 2]

The Text level results for KeyTags can also show how a word may have different meanings across different text types. Figure 3 shows the Text level results for *goal* in two sub-corpora of the *BNC*. The range indicators for these two sets of results are much lower than those for *therefore* and *thus*, meaning that while these results are significant, *goal* also occurs in many other contexts. For the BNC: Academic sub-corpus, "Technology and Engineering" only represents 188 out of 754 instances. Checking the BNC: Newspapers sub-corpus files directly reveals that both "CEP.xml" and "CBG.xml" are also labelled "Sports" , so through examining the files it can be found that in this case the lower range indicator showing 1,256 out of 2,620 instances (48%) is correct.

[Figure 3]

4.2. Section Level

Section Level Key Labels are based on the sub-headings used in different sections of a text, with all the words in or under section headings being mapped to these sub-headings. They can give insights into aspects of text structure and the actual topics of the parts of the texts containing the word or collocation. They can reveal how academic texts use words with a similar meaning in different sections of text, indicating a particular sense. In the *Hindawi Biological Sciences* corpus, the clouds shown in Figure 4 show how *important* differs from *significant*, with the latter clearly identified with its use in statistics.

[Figure 4]

4.3. Producer Level

Items for the Producer panel can include metadata about the authors or speakers of complete texts and also metadata about the authors or speakers for each section of texts (which may, for example, differentiate between two or more speakers in a spoken text). For single texts where a word or collocation appears very frequently, the author's name and other metadata about the author may appear in this panel, complementing the information provided about the

text which appears in the Text panel. For example, the Text panel for *marginal cost* in the *BNC* includes prominent tags for the title of the book 'Economics', 'W Commerce' and 'NON-ACADEMIC' as well as publishing information, while the Producer panel shows the three names of the authors of this book: 'Begg, David', 'Fischer, Stanley' and 'Dornbusch, Rudiger'. Corpora of spoken texts tend to have more metadata available about the speakers and Key Label for these are also shown in the same Producer panel.

4.4. Neighbourhood Key Labels

An important way of viewing concordance lines in *Author_Software* is the cards view, which shows the sentence containing the node word as well as the context up to one sentence before and one sentence after (Author, 2017a, 2017b). The Neighbourhood Key Labels represent repeated semantic tags which occur within these extended co-texts. When corpora are refactored, USAS (Rayson, et al., 2004) is used to tag individual words and multi-word units. Sentences in the corpus are then tagged for the calculation of Neighbourhood Key Labels, retaining only those semantic tags which meet a threshold of repetition within the extended co-text. The database stores semantic tags for the sentences based on links with thresholds of between two and eight repetitions, allowing for some fine-tuning in the concordancer itself. As individual words when tagged using USAS may have multiple semantic tags, the requirement for a minimum number of links across the extended co-text, provides a straight-forward (albeit limited) means of 'disambiguation'; the same semantic tag has to occur between 2 and 8 times in order for the semantic tag of the sentence to be counted. The log-likelihood contingency table for Neighbourhood Key Labels, essentially creates a sub-corpus of sentences marked with each specific semantic tag, comparing it to the other sentences in the corpus which have not been tagged with this semantic tag. Figure 5 shows screenshots of the clouds for *heart* in the BNC: Academic sub-corpus and the Fiction Collection 12x7[7].

[Figure 5]

5. Conclusion

As well as providing new kinds of data for corpus users, this approach also tries to bridge the gap between the sophisticated mark-up of modern XML corpora and visual presentation of

[7] For information about these corpora, see the help pages of www.theprimemachine.com

Key Labels which might aid users in interpreting typical contexts for search terms. The tool was initially developed for language learning and teaching. In essence, the Key Label display should provide useful information for both language teachers and language users by helping both of them to understand the composition of the corpus and the kinds of examples which will be displayed. For a language teacher the examples that a corpus can provide need to be judged not only in terms of the lexico-grammatical range, but also in terms of the appropriateness of the registers and text types represented in the corpus. A teacher using the Key Label function would be able to quickly see which kinds of examples were most prominent in the set of concordance lines for the currently selected corpus and the teacher should be able to get a clear sense of whether it is balanced and whether it fits their intended target group. For language learners, Key Labels provide information about typical uses in terms of the major text categories, section headings and language producers. The Neighbourhood Key Labels can provide a snapshot of the local contexts in which a word or collocation occurs, helping learners to see some of the differences between synonyms in terms of their semantic associations. By looking at the Key Labels a student could also be alerted to potential restrictions on usage, as these labels can help show how a word or collocation may be restricted to specific genres/registers or in connection with certain kinds of topic. For example, a word typically only used in casual, informal situations may show a Key Label for "Conversation", while an alternative synonym may be more suitable for academic writing. This is one of the ways in which *The Prime Machine's* ability to display results for similar words and terms side by side aims to help learners select an appropriate term from a choice of near synonyms or different word forms.

The Key Label method could also provide complementary data for other kinds of corpus research. It is not intended to replace key word methods or other measures of dispersion. It should also be noted that while key word techniques are corpus-driven, Key Labels lose the initial objectivity from the perspective of specific lexical items since search terms for analysis need to be specified by the researcher. However, from the perspective of metadata or division of corpus texts into groups, the Key Labels method actually provides corpus-driven results. As such, it could be used as a follow-up to traditional key word analyses, allowing a researcher to see whether other text labels, section labels, author information or semantic tags may provide a more fine-grained explanation of the data within the study corpus. After discovering through a key word tool that a word is key in a corpus, Key Labels could reveal whether sub-corpora or specific sections of the study corpus could be used to narrow down

further. For research where the initial decision to select specific words is predetermined this approach also provides more ways to summarise data and provides hints for further analysis of concordance lines, perhaps in terms of the specific sources, dates, typical section headings or semantic senses. Research of this type might include analysis of proper names or words strongly connected with specific themes for corpus stylistics, historical linguistics and sociolinguistics.

Acknowledgements

The author is grateful to the supervisor and external examiner of the thesis (Jeaco, 2015) and to the two anoynmous reviewers for their comments and suggestions.

References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. Paper presented at the Interactive Workshop on Language e-Learning, Waseda University, Tokyo.

Biber, D. (1991). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D., Connor, U., & Upton, T. A. (2007). *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam: John Benjamins.

Biber, D., & Conrad, S. M. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

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

Friginal, E. (2013). Twenty-five years of Biber's multi-dimensional analysis: Introduction to the special issue and an interview with Douglas Biber. *Corpora, 8*(2), 137-152.

Gabrielatos, C., & Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. Paper presented at the CADS International Conference 2012, University of Bologna, Italy. http://repository.edgehill.ac.uk/4196/1/Gabrielatos%26Marchi-Keyness-CADS2012.pdf

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. In S. T. Gries, S. Wulff & M. Davies (Eds.), *Corpus-Linguistic Applications: Current Studies, New Directions* (pp. 197-212). Amsterdam: Rodopi.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Jeaco, S. (2015). *The Prime Machine: a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language.* Unpublished Ph.D. dissertation. University of Liverpool. Available from https://livrepository.liverpool.ac.uk/2014579/

Jeaco, S. (2017a). Concordancing Lexical Primings. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 273-296). Amsterdam: John Benjamins.

Jeaco, S. (2017b). Helping Language Learners Put Concordance Data in Context: Concordance Cards in The Prime Machine. *International Journal of Computer-Assisted Language Learning and Teaching, 7*(2), 22-39.

Liang, M. (2015). Patterned Distribution of Phraseologies within Text: The Case of Research Articles. In B. Zou, M. Hoey & S. Smith (Eds.), *Corpus linguistics in Chinese contexts* (pp. 74-97). Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Mahlberg, M., Stockwell, P., Joode, J. d., Smith, C., & O'Donnell, M. B. (2016). CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora, 11*(3), 433-463.

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is This Corpus About?': Using Topic Modelling to Explore a Specialised Corpus. *Corpora, 12*(2), 243-277.

Nini, A. (2014). *Multidimensional Analysis Tagger 1.1 - Manual*. Retrieved from http://sites.google.com/site/multidimensionaltagger

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Phillips, M. A. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam: North-Holland.

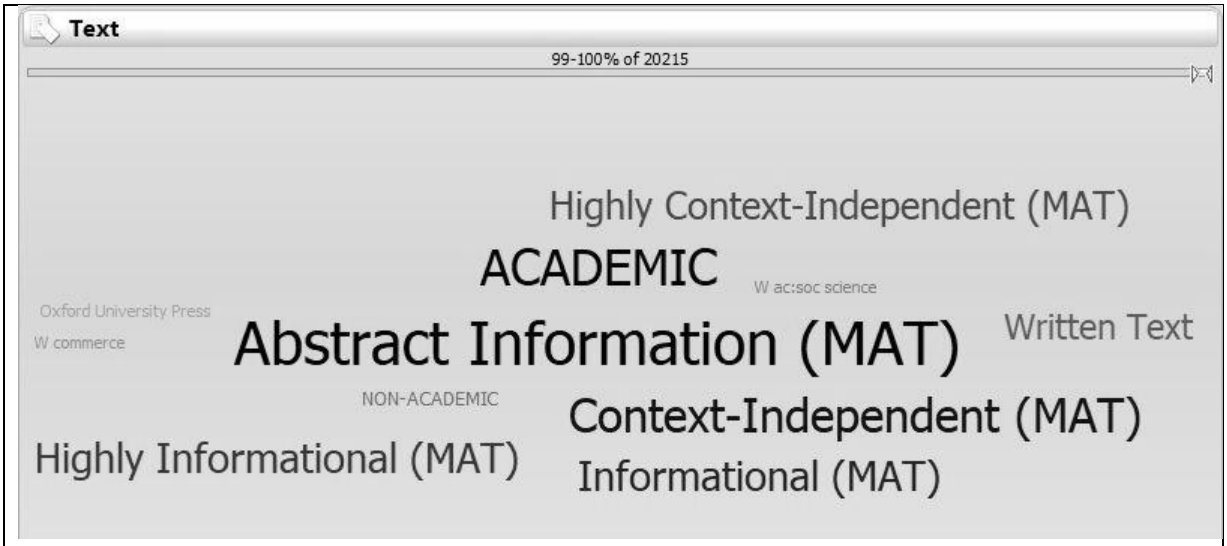
Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519-549.
Scott, M. (1997). PC analysis of key words -- and key key words. *System*, *25*(2), 233-245.
Scott, M. (2000). Mapping key words to problem and solution. In M. Scott & G. Thompson (Eds.), *Patterns of Text: In Honour of Michael Hoey* (pp. 109-127). Amsterdam: John Benjamins.
Scott, M. (2010). *WordSmith Tools* (Version 5.0). Oxford: Oxford University Press.
Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
Wilson, A. (2013). Embracing Bayes Factors for key item analysis in corpus linguistics. In M. Bieswanger & A. Koll-Stobbe (Eds.), *New Approaches to the Study of Linguistic Variability*. (pp. 3-12). Frankfurt: Peter Lang.

Table 1: Key Labels contingency table.

	Sub-Corpus 1	**Sub-Corpus 2**	**Total**
Node Word	Node word inside XML node	Node word outside XML node	Frequency of node word
Other Words	Other words inside XML node	Other words outside XML node	Frequency of other words
Total	Word count inside XML node	Word count outside XML node	Whole Corpus

Tag	Frequency	LL	Bayes Factor
Abstract Information (MAT)	11,518	12381.3	Very strong evidence
Highly Abstract Information (MAT)	6,849	8037.99	Very strong evidence
Context-Independent (MAT)	16,815	6711.1	Very strong evidence
ACADEMIC	8,455	6061.7	Very strong evidence
Highly Context-Independent (MAT)	12,474	4309.01	Very strong evidence
Highly Informational (MAT)	18,651	2112.84	Very strong evidence
Informational (MAT)	19,454	2085.85	Very strong evidence
Overtly Persuasive (MAT)	4,445	1686.09	Very strong evidence
NON-ACADEMIC	7,564	929.85	Very strong evidence
W ac:polit law edu	1,272	897.47	Very strong evidence

Figure 1: Tag clouds and tables for *therefore* in the *BNC: Complete* Corpus.



Tag	Frequency	LL	Bayes Factor
Abstract Information (MAT)	10,946	13499.7	Very strong evidence
Highly Abstract Information (MAT)	6,848	9616.43	Very strong evidence
ACADEMIC	8,875	9031.92	Very strong evidence
Context-Independent (MAT)	15,548	7751.24	Very strong evidence
Informational (MAT)	18,818	5763.58	Very strong evidence
Highly Informational (MAT)	18,232	5724.7	Very strong evidence
Highly Context-Independent (MAT)	11,825	5405.84	Very strong evidence
Written Text	19,957	4520.43	Very strong evidence
NON-ACADEMIC	7,422	1617.85	Very strong evidence
W ac:soc science	929	992.4	Very strong evidence

Figure 2: Tag clouds and tables for *thus* in the *BNC: Complete* corpus

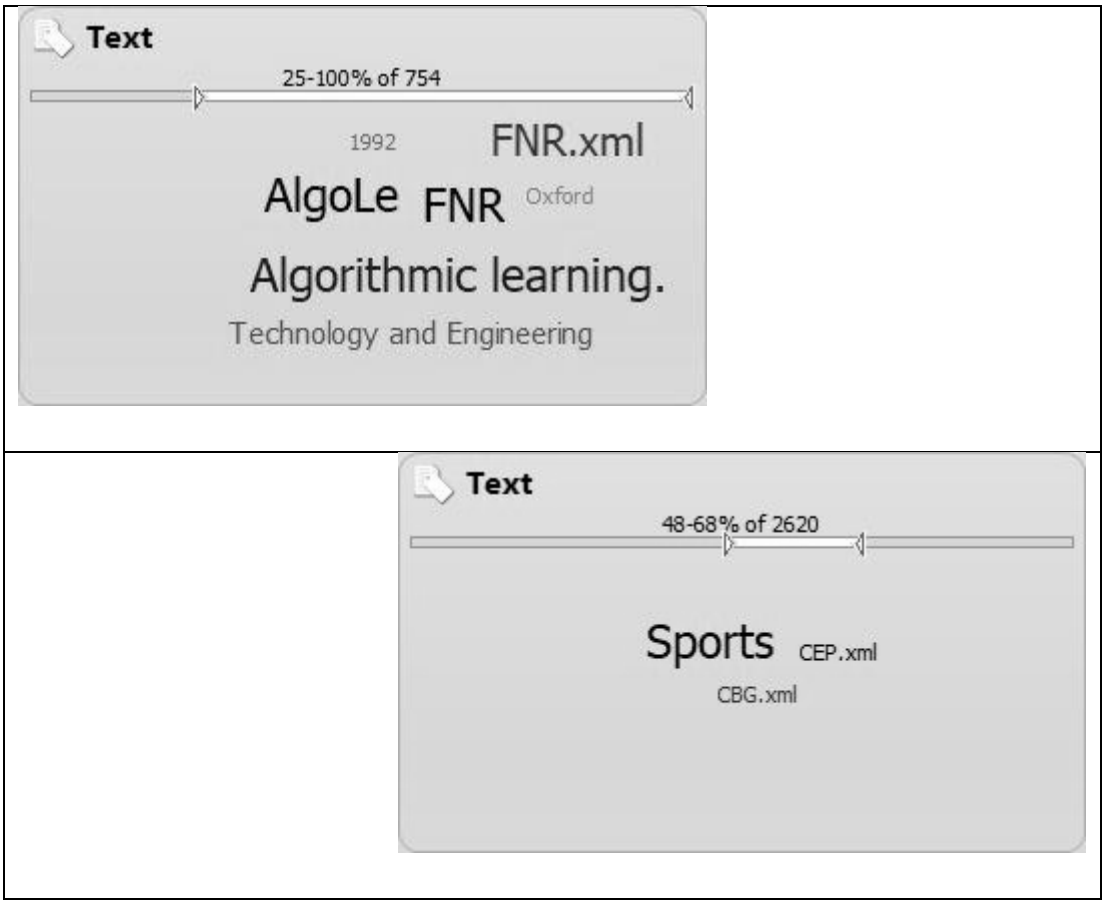


Figure 3: Clouds for *goal* in the *BNC: Academic* sub-corpus (top) and the *BNC: Newspapers* sub-corpus (bottom).

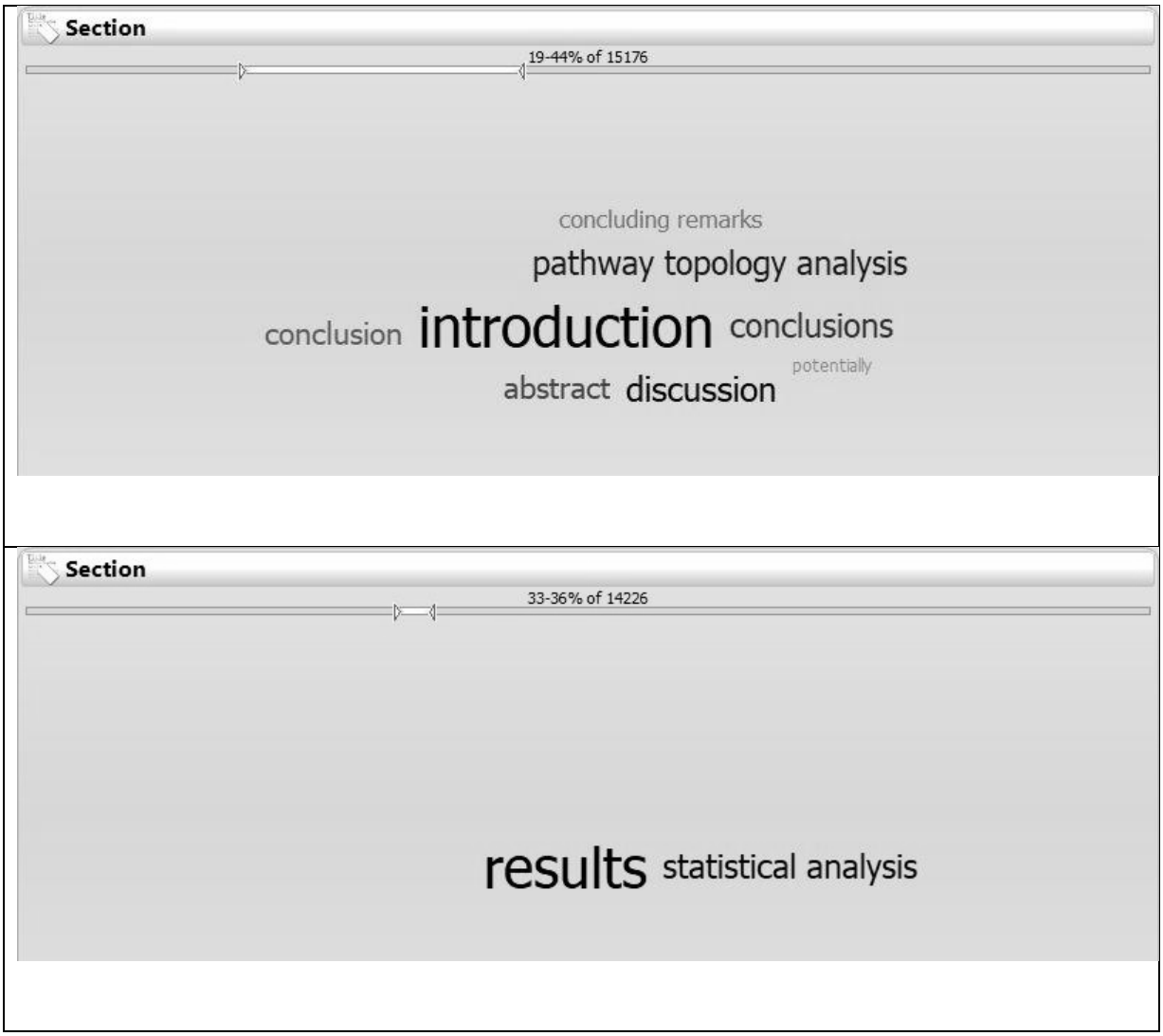


Figure 4: Clouds for important (top) and significant (bottom) in the Hindawi Biological Sciences corpus.

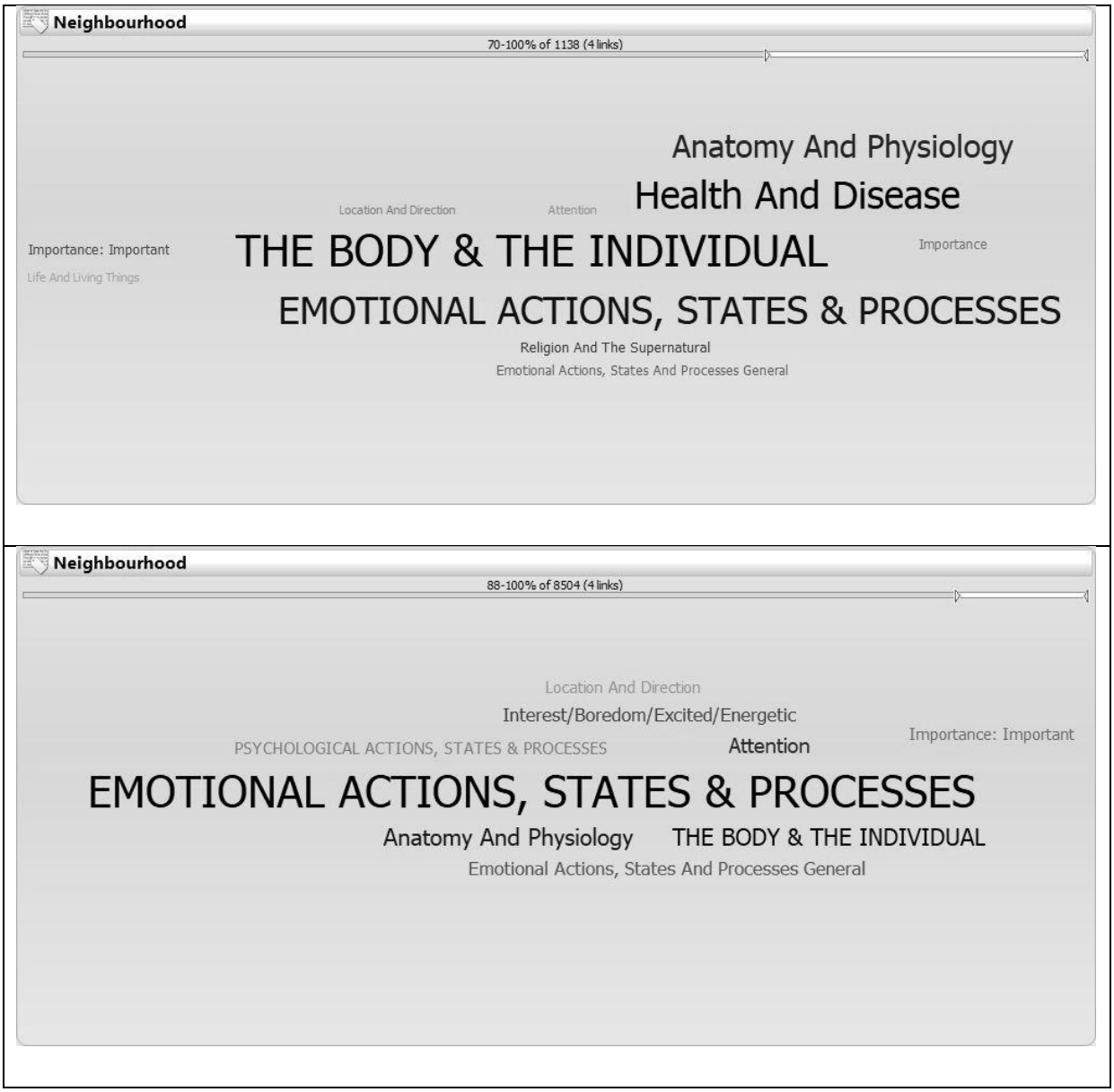


Figure 5: Clouds for *heart* in the BNC:Academic (top) and Fiction Collection 12x7 (bottom) corpora.