To cite this document: Bowker, Lynne (2018) "Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research", Library Hi Tech, Vol. 36 Issue: 2, pp. 358-371.
Permanent link to this document: https://doi.org/10.1108/LHT-12-2017-0271

Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research

Lynne Bowker School of Information Studies, University of Ottawa, Ottawa, Canada

Abstract

Purpose – This methodology paper aims to generate awareness of an interest in the techniques used in computer-based corpus linguistics, focusing on their methodological implications for research in library and information science (LIS).

Design/Methodology/Approach – This methodology paper provides an overview of computer-based corpus linguistics, describes the main techniques used in this field, assesses its strengths and weaknesses, and presents examples to illustrate the value of corpus linguistics to LIS research.

Findings – Overall, corpus-based techniques are relatively simple, yet also powerful, and they support both quantitative and qualitative analyses. While computer-based corpus methods alone may not be sufficient for research in LIS, they can be used to complement and to help triangulate the findings of other methods. Corpus linguistics techniques also have the potential to be exploited more fully in LIS research that involves a higher degree of automation (e.g. recommender systems, knowledge discovery systems, text mining).

Originality/value – Over the past quarter century, corpus linguistics has established itself as one of the main methods used in the field of linguistics, buts its potential has not yet been fully realized by researchers in other fields, including LIS. Corpus linguistics tools are readily available and relatively straightforward to apply. By raising awareness about corpus linguistics, we hope to make these techniques available as additional tools in the LIS researcher's methodological toolbox, thus broadening the range of methodological tools applied in this field.

Introduction

Are Library and Information Science (LIS) researchers creatures of habit? After examining the contents of twenty high-profile LIS journals published in 2005, Hider and Pymm (2008) identified ten different research strategies, but found that close to one-third of all the articles published utilized a survey-based research strategy. Likewise, an investigation by Turcios *et al*. (2014) revealed that of the eleven research methods employed in 2013, surveys remained the most popular, being used in 21% of the 307 articles reviewed for their study. Meanwhile, after completing a longitudinal investigation into the research methods used in LIS during the forty-year period between 1970 and 2010, Gauchi Risso (2016) asserts that LIS needs new methodological developments, which should combine qualitative and quantitative approaches.

One method that does appear to be gaining ground in LIS research is content analysis. While Hider and Pymm (2008) indicate that content analysis was used in just 4.8% of research articles published in 2005, this had risen to 13% in 2013 according to the study done by Turcios *et al*. (2014). As described by White and Marsh (2006):

> Content analysis is a highly flexible research method that has been widely used in library and information science (LIS) studies with varying research goals and objectives. The

> research method is applied in qualitative, quantitative, and sometimes mixed modes of research frameworks and employs a wide range of analytical techniques to generate findings and put them into context. (22)

White and Marsh (2006) provide a selective list of 25 examples of studies in LIS that were carried out using content analysis during the 25-year period between 1991 and 2005. Many of these analyses were carried out manually, relying on a close reading of the material under study and deriving prevalent themes from it through interpretative methods. However, more recently, there is evidence that *corpus linguistics techniques* are beginning to emerge as a powerful complement to content analysis. Indeed, Baker *et al*. (2008) describe content analysis and corpus linguistics as a useful methodological synergy.

Corpus linguistics is a methodology that originated in linguistics, but which can be applied in a wider range of fields connected to the Digital Humanities, including LIS. Indeed Digital Humanities (DH) is a key theme identified in the recent *Trend Report* produced by the International Federation of Library Associations and Institutions (IFLA, 2013). The goal of this paper is to introduce corpus-based methods to LIS researchers and to stimulate discussion on the potential of such methods for LIS research. The paper begins with an introduction to corpus linguistics, including a brief history. Next we describe the main quantitative and qualitative techniques used in corpus linguistics, supported by examples of actual and potential applications of such techniques in LIS research. Finally, we discuss some of the implications of corpus linguistics for LIS research, before offering some concluding remarks.

What is corpus linguistics?

A very simple description of corpus linguistics is the study of language based on examples of real-life use (McEnery and Wilson, 1996). When it was first introduced in the 1960s, there was

considerable debate as to whether corpus linguistics constituted a theory, a branch of linguistics, or a methodology. However, in more recent years, the dust has begun to settle and there is now widespread agreement that corpus linguistics is a methodology whose techniques can be applied across nearly all branches of linguistics. Indeed, since the late 1980s, it has gone on to become an increasingly prevalent methodology in the field of linguistics, witnessed by the establishment of the *International Journal of Corpus Linguistics* by the John Benjamins Publishing Company in 1996, followed by a spate of introductory textbooks on the topic (e.g. McEnery and Wilson, 1996; Biber *et al*., 1998; Kennedy, 1998; Oakes, 1998). Now there is a considerable body of literature describing many aspects and applications of corpus-based techniques in all areas of linguistics.

Corpus linguistics is an empirical approach that involves studying authentic examples of what people have actually written or said, rather than hypothesizing about what they might or should say. It is often contrasted with other methods of gathering linguistic evidence, such as *introspection* (sometimes referred to as *armchair linguistics*), where a linguist relies on his or her own intuitions and judgment about what constitutes an acceptable usage, or *casual citation*, where a linguist observes and records the language-related behaviour of family, friends or strangers.

In contrast, and as its name suggests, corpus linguistics requires the use of a corpus. Strictly speaking, a corpus is simply a body of text; however, in the context of corpus linguistics, the definition of a corpus has taken on a more specialized meaning. According to Bowker and Pearson (2002), a corpus can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria. A key observation made by Kennedy (1998) is that rather than initiating corpus research, developments in information

technology changed the way we work with corpora, such that corpus linguistics is thus now inextricably linked to the computer. With the help of a computer, researchers can store huge volumes of text, retrieve particular instances of words or phrases quickly and exhaustively, and sort and display this textual data in different ways, thus facilitating interpretation.

Who uses corpus linguistics?

Lexicographers – the people who compile dictionaries – were the first type of linguist to popularize corpus-based techniques (Sinclair, 1987). As described by Biber *et al*. (1998), lexicographers are interested in finding answers to questions such as what meanings are associated with a particular word, what is the frequency of a word relative to other related words (e.g. synonyms), and what words commonly occur with a particular word (e.g. collocations).

Another group who became early adopters of corpus-based techniques were language teachers. The biennial Teaching and Language Corpora (TaLC) conferences, first held in 1994 and most recently in 2016, serve as a forum for teachers, researchers and software developers to share their ideas on how corpus resources and tools for analysis can assist in language teaching. It is now well-known that corpus examples are valuable in language learning because they expose students at an early stage in the learning process to the kinds of sentences and vocabulary that they are likely to encounter when reading genuine texts or in real communicative situations. Frequency data is also useful in a language learning context because it can help language learners to focus their vocabulary building efforts on those words that appear most frequently in a given language.

Since these early adopters, many other types of linguists – including grammarians, semanticists, sociolinguists, historical linguists, dialectologists, and translators – have used a variety of corpus-based techniques to further their understanding of different aspects of language. And now

corpus-based methods are beginning to attract attention in other domains too, including the broader Digital Humanities as well as LIS. But what exactly are these techniques? As we will discover in the next section, basic corpus linguistics techniques are simple yet powerful, and they can be applied in various types of LIS research.

What techniques are used to process corpora?

Corpus linguistics techniques are essentially based on two things that computers are very good at: *number crunching* and *pattern matching*. Among the most fundamental corpus processing techniques we find measures of frequency and KWIC concordances. These techniques facilitate both quantitative and qualitative analyses, and as such, corpus linguistics can respond to the plea made by Gauchi Risso (2016) for LIS to adopt new methods that support both types of analysis.

Number crunching: Measures of frequency

Consider that a corpus is a text file. It could be made up of tens, hundreds or thousands of documents and may run to hundreds of thousands or even millions of words. Trying to count the number of words, or the number of times each word occurs, would be a time-consuming, labour-intensive and error-prone process if it were done manually. However, this type of work is easily accomplished by a computer, and corpus analysis software can be used to calculate several different measures of frequency, including raw frequency counts (e.g. word lists), measures of disproportionate frequency (e.g. keyness), and measures of relative frequency (e.g. collocations).

Word lists

Firstly, a corpus analysis tool can quickly and easily count the number of words (tokens) and the number of different words (types) in the corpus and then display this information in different ways. As illustrated in Figure 1, the most basic lists show the words in alphabetical order alongside the number of occurrences of each, or in order of descending or ascending frequency. A more sophisticated variation is a lemmatized list, where words that share the same base form are grouped and counted together. For example, in a lemmatized list, the forms *eat*, *eats*, *eating*, *eaten*, and *ate* would be considered to be one form or lemma, and the total number of occurrences for that lemma would be calculated. A stop list, which is a list of words that should be ignored by the computer, can also be introduced. A commonly used stop list is one that instructs the computer to ignore function words, such as articles, conjunctions and prepositions, and to focus on the content words, which are words that have semantic meaning. However, any kind of stop list can be developed depending on the goal of the investigation.

In linguistics, word frequencies can be useful for lexicographers and language teachers, as mentioned above. In LIS, a relevant application is in the context of controlled vocabularies. Often, multiple possible terms could be used to describe a subject, but in order to reduce ambiguity and improve retrieval, authorized terms must be selected by LIS professionals. Authorized terms are chosen based on the principles of *user warrant* (what terms users are likely to use), *literary warrant* (what terms are generally used in the literature and documents), and *structural warrant* (terms chosen by considering the structure and scope of the controlled vocabulary). Although frequency is not the only relevant measure, it can certainly provide an LIS professional with valuable information to help them choose appropriate authorized terms.

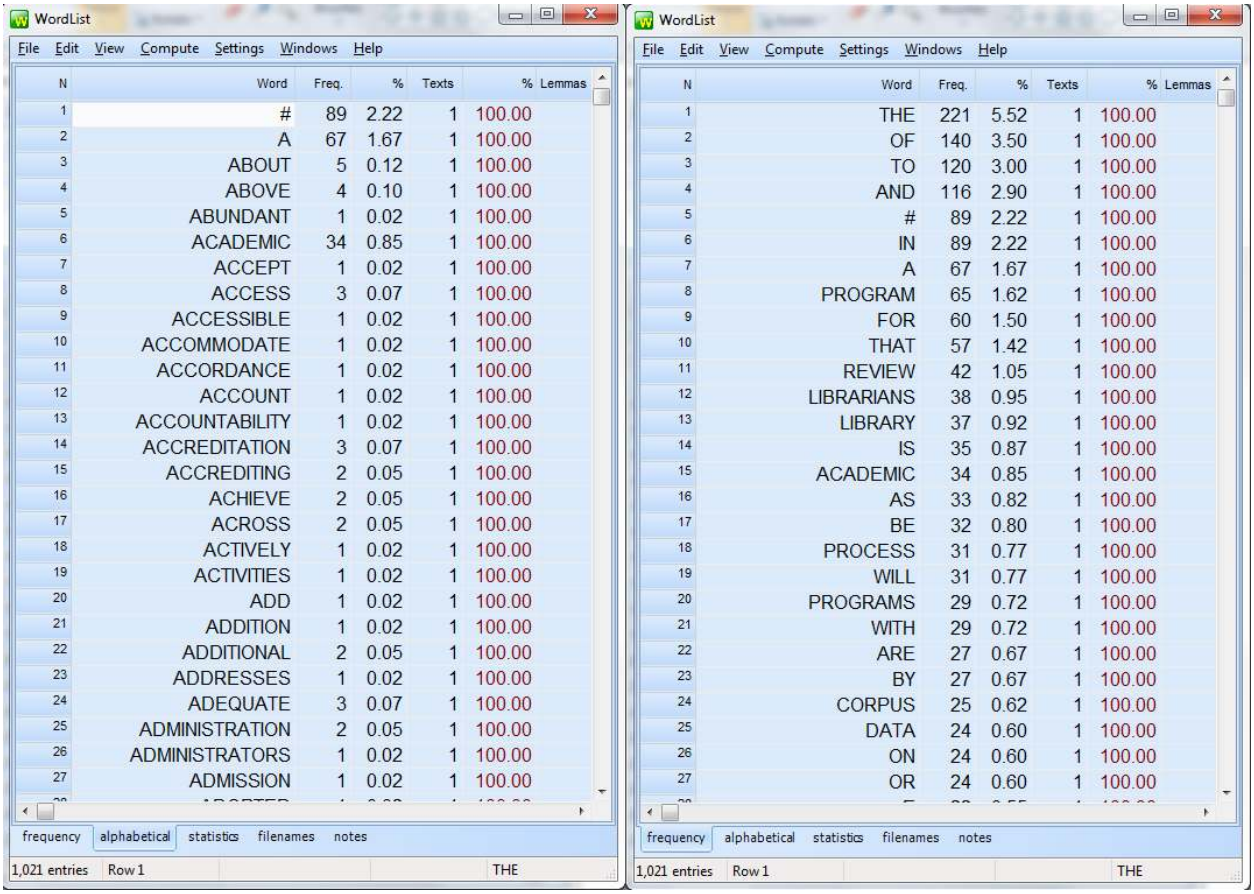


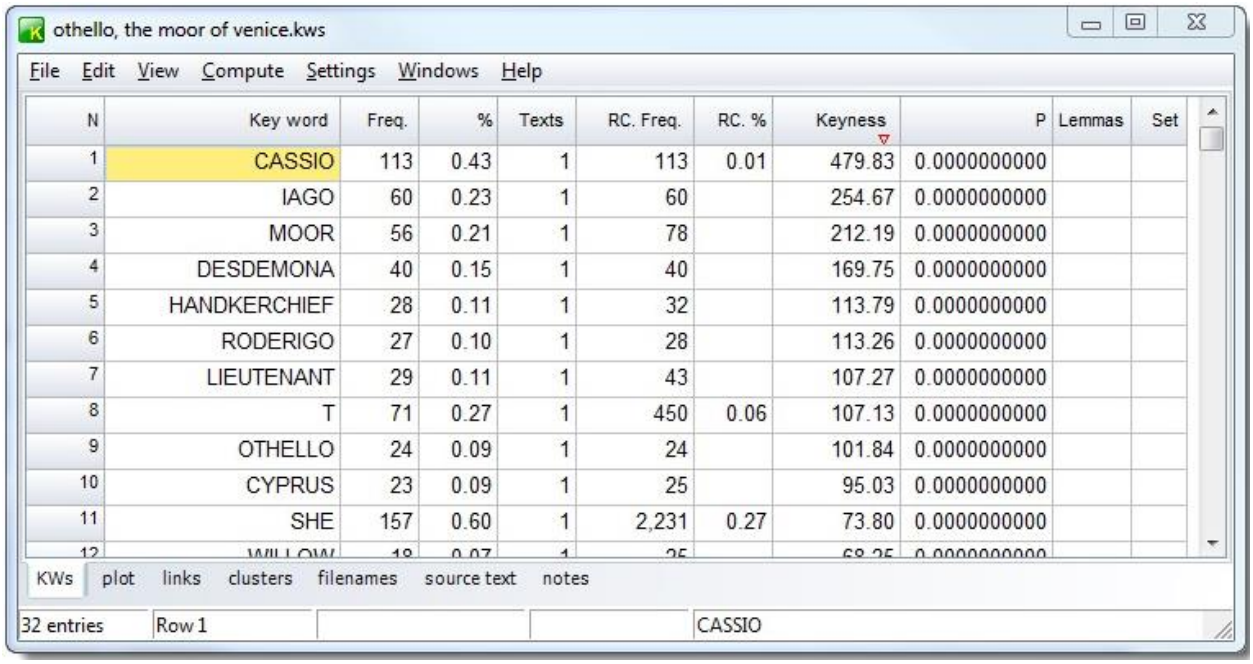
Figure 1. Screenshots of word lists generated by the corpus analysis software WordSmith Tools. The word list on the left-hand side is displayed in alphabetical order with frequency counts to the right, while the list on the right-hand side is displayed in order of descending frequency.

Because corpora are samples of authentic language use, frequency data drawn from a corpus can help an LIS professional to better understand literary warrant of a given text collection. For instance, LIS researchers Anguiano Peña and Naumis Peña (2015) describe how corpus-based techniques can be applied to help establish indexing languages. In particular, they note that frequency measures can be useful when extracting candidate terms from a corpus because those lexical units that appear very infrequently and with scant literary warrant are filtered out (Anguiano Peña and Naumis Peña, 2015).

Keyness

Although raw frequency can be a potential indicator of a word's importance in a text, it is not sufficient as a sole measure of identifying the subject of a text. In addition to a frequency count,

a corpus analysis tool can calculate the *keyness* of a given term. Scott and Tribble (2006), describe keywords as items of unusual frequency in comparison with a reference corpus. To identify keywords, a corpus analysis tool compares the contents of a particular text against a much larger reference corpus to identify those words in the text that are *unusually* frequent as compared to their frequency in the reference corpus. In other words, keyness is a measure of the frequency of disproportionate occurrence. It also important to note that keyness is a textual feature, not a language feature, which means that a word has keyness in a certain textual context, but this same word may not have keyness in other contexts. Scott and Tribble (2006) illustrate keyness using the example of Shakespeare's plays. By taking the play of *Othello* as the specific text to be examined, and comparing this against a reference corpus of all 37 of Shakespeare's plays, we can see which words are key to the play of *Othello* (see Figure 2).



N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	CASSIO	113	0.43	1	113	0.01	479.83	0.0000000000		
2	IAGO	60	0.23	1	60		254.67	0.0000000000		
3	MOOR	56	0.21	1	78		212.19	0.0000000000		
4	DESDEMONA	40	0.15	1	40		169.75	0.0000000000		
5	HANDKERCHIEF	28	0.11	1	32		113.79	0.0000000000		
6	RODERIGO	27	0.10	1	28		113.26	0.0000000000		
7	LIEUTENANT	29	0.11	1	43		107.27	0.0000000000		
8	T	71	0.27	1	450	0.06	107.13	0.0000000000		
9	OTHELLO	24	0.09	1	24		101.84	0.0000000000		
10	CYPRUS	23	0.09	1	25		95.03	0.0000000000		
11	SHE	157	0.60	1	2,231	0.27	73.80	0.0000000000		
12	WILLOW	18	0.07	1	25		68.25	0.0000000000		

Figure 2. A screenshot of the keywords generated by WordSmith Tools for the play *Othello*, when the text of this play is compared against a reference corpus containing all 37 of Shakespeare's plays.

The notion of keyness in corpus linguistics is directly pertinent to the notion of *aboutness* in LIS. In 2007, for example, researchers working at the *Laboratoire de représentation des*

connaissances (knowledge representation lab) at the Université de Paris-13 incorporated corpus linguistics techniques for identifying keyness into a system intended to automatically construct back-of-the-book indexes (Nazarenko and Aït El Mekki, 2007). They summarized their results as promising and encouraged members of the information science and linguistic communities to work more closely to find solutions to issues of common interest.

While Nazarenko and Aït El Mekki's work focused on back-of-the-book indexes, the use of corpus linguistics techniques such as keyness identification also hold promise for other types of indexing (e.g. for research databases). For instance, Bowker *et al*. (2015) conducted a pilot study using a freely available online corpus analysis tool called TermoStat, developed by linguistics professor Patrick Drouin (2003). Bowker et al. used TermoStat to generate keywords for four academic papers, and then compared the usefulness of these keywords for retrieval purposes against the ones generated by professional indexers, authors, and users. They found that there was complementarity among the lists, and suggested that such automatically-generated keywords could be a useful starting point for or supplement to the work of professional indexers.

Meanwhile, Kehoe and Gee (2011) build on the concept of keyness to find the key tags assigned to pages in the social bookmarking site Delicious. According to Kehoe and Gee, key tags are those tags from the social tagging lexicon that best describe the content of a particular web page. In other words, from a retrieval point of view, they are the tags that make a web page stand out from the crowd. With the help of a corpus analysis tool called WordSmith Tools, developed by linguistics professor Mike Scott (2015), Kehoe and Gee compare the tags assigned to an individual web page in Delicious with those assigned to all web pages in Delicious to identify the key tags. In explaining their approach, these researchers argue that, as online textual holdings continue to increase in size, further linguistic insight will be vital to ensure that social tagging

sites continue to function effectively and that the power of social tagging as a window to the views of large numbers of readers can be harnessed effectively. Kehoe and Gee (2011) go on to note that specialists in a variety of fields have begun to examine the aboutness issues in social tagging, and they indicate that their paper describes what is, to the best of their knowledge, the first attempt to apply corpus linguistic methods.

Another potential application of keyness, which does not appear to be attested in the LIS literature, is in the context of recommender systems. For instance, Beel *et al*. (2016) conducted a review of over 200 research articles that were published on research-paper recommender systems, observing that content-based filtering was the most commonly adopted approach. According to Beel *et al*. (2016), the content-based filtering approaches mainly utilized papers that the users had authored, tagged, browsed, or downloaded. None of the approaches seem to have incorporated the notion of keyness as used in corpus linguistics; however, it would be very interesting to determine if this technique could be integrated into recommender systems to see if this could enhance the results.

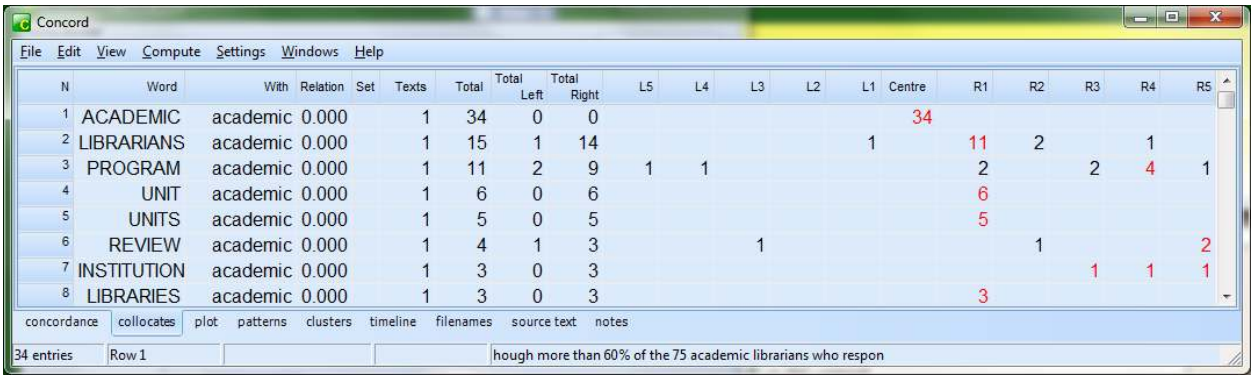
Collocations

In addition to calculating raw frequency counts and measures of unusual frequency, corpus analysis software can also carry out calculations that measure the relationship between different words in the corpus. In linguistics, for example, collocations are words that hang around together or are typically found in each other's company. Some common collocations for the word *book* include *read*, *check out*, *rare*, and *second-hand*. A computer can measure the strength of collocation between two words in a corpus by determining whether these two words appear together with a frequency that is greater than chance. Essentially, the computer can determine

whether the two words appear next to each other more often that we might expect, based on what we know about their individual frequencies. In other words, the computer can tell us whether or not this is an accidental pairing. One commonly used formula to determine whether two words are collocates is mutual information (MI) (Oakes, 1998). To calculate MI, the computer compares:

- the probability of the two words appearing together if they are independent (p(w1)p(w2)); and
- the actual probability of the two words appearing together (p(w1w2)).

Figure 3 illustrates the collocates of the word *academic* as found in a corpus about the participation of librarians in a university's quality assurance process for academic programs. As we can see, words that most typically appear in the company of *academic* in this corpus include *librarians*, *program*, and *unit(s)*, among others.



N	Word	With	Relation	Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	ACADEMIC	academic	0.000		1	34	0	0						34					
2	LIBRARIANS	academic	0.000		1	15	1	14					1		11	2		1	
3	PROGRAM	academic	0.000		1	11	2	9	1	1					2		2	4	1
4	UNIT	academic	0.000		1	6	0	6							6				
5	UNITS	academic	0.000		1	5	0	5							5				
6	REVIEW	academic	0.000		1	4	1	3			1					1			2
7	INSTITUTION	academic	0.000		1	3	0	3									1	1	1
8	LIBRARIES	academic	0.000		1	3	0	3							3				

Figure 3. A screenshot of the collocates generated for the word *academic* in a corpus about librarians' participation in program reviews in postsecondary institutions.

How might collocations be used in the context of LIS research? In a volume on trends in LIS research in Europe, Greene and McMenemy (2012) make a convincing case for the value of triangulating data by using multiple methods, particularly combinations that include both

quantitative and qualitative techniques, in LIS research. Employing both content analysis and corpus linguistics techniques, Greene and McMenemy (2012) use a corpus of 51 policy documents to investigate the impact of neoliberal ideology on public libraries in the United Kingdom. Although collocation analysis can be done manually, without the support of corpus linguistics tools, a purely manual approach presents a number of risks. Firstly, the number of documents that can be analyzed manually tends to be relatively small. In addition, when collocations are identified manually, they are not usually statistically calculated. Therefore, information regarding their statistical significance, the collocation span, or any frequency thresholds, is not usually provided. Such approaches may miss or disregard strong non-adjacent collocates, or include non-significant collocates in the analysis. For their study, Greene and McMenemy (2012) use AntConc, a freeware corpus analysis tool developed by linguistics professor Laurence Anthony (2006), to analyze the policy documents in various ways, including collocation analysis. In one example, Greene and McMenemy (2012) investigate collocates of the term "rights" in the policy documents, which include *human*, *access*, *democratic*, *fundamental*, *cultural*, *borrowing*, *management*, *welfare* and *digital*. According to Greene and McMenemy (2012), their approach reflects the linguistic turn in the social sciences and humanities which emphasises the role of language in the construction of social reality.

Pattern matching

The preceding sections have demonstrated how the superior number crunching abilities of computers can be harnessed by researchers to support various types of quantitative linguistic analyses that are relevant to LIS research. However, quantitative findings need to be interpreted, and corpus linguistics also includes techniques that support qualitative analysis. In this section,

we will consider how another strength of computers – pattern matching – can be incorporated into corpus-based methods for the benefit of LIS researchers.

KWIC concordances

A concordancer is a corpus analysis tool that retrieves all the occurrences of a particular search pattern in its immediate contexts. This information is typically displayed using a format known as key word in context (KWIC), as illustrated in Figure 4. In a KWIC display, all the occurrences of the search pattern are lined up in the centre of the screen with a certain amount of context showing on either side. The amount of context can typically be expanded or contracted as desired. As was the case with the word frequency lists discussed previously, it is possible to sort concordances so that it becomes easier to identify patterns that might otherwise go undetected. Common ways of sorting concordance lines include alphabetically according to the words preceding or following the search pattern.

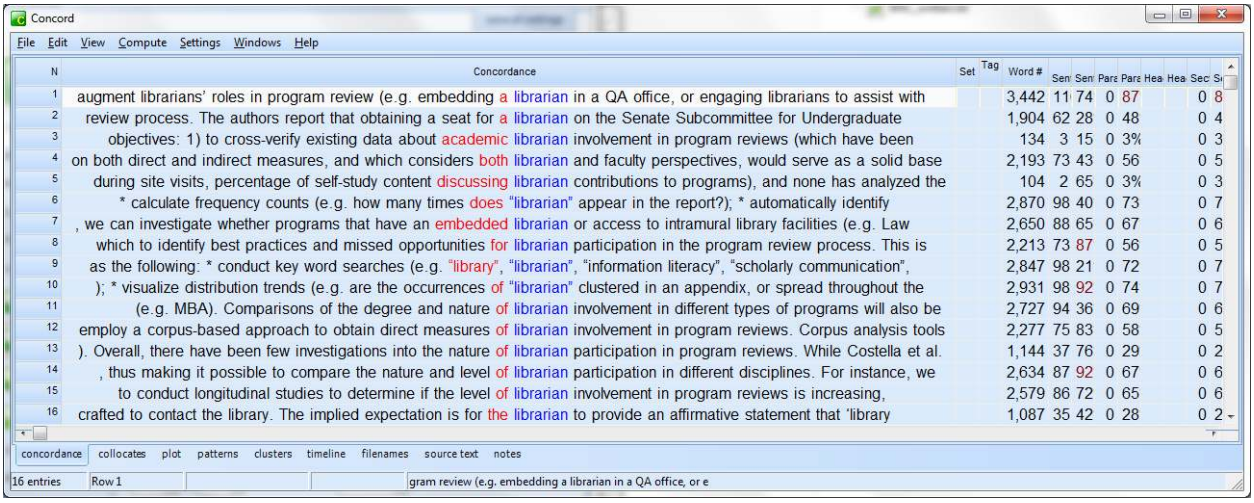


Figure 4. A screenshot of a KWIC concordance for the term *librarian* generated using the WordSmith Tools concordancer.

In addition to exact string searching, concordancers permit more sophisticated search patterns, allowing users to conduct lemmatized searches, case-sensitive searches, wildcard searches, and

searches using boolean operators or other regular expressions. Regardless of the type of search pattern entered, the benefit of using concordance lines as a source of linguistic evidence is that they reveal the context in which individual occurrences of words are found.

In a corpus-based approach, researchers might begin with an examination of relative frequencies and statistically significant lexical patterns in the corpus, and then follow this up with a close qualitative analysis of examples of those patterns in context as presented by a concordancer. Indeed, this approach was taken by Greene and McMenemy (2012) in their study of the public library policy documents. As mentioned above, they first used frequency measures to determine which words and collocations were significant in the corpus, and they then proceeded to use KWIC concordances to study these occurrences in more detail (Greene and McMenemy, 2012). In their closing remarks, Greene and McMenemy (2012) state that their chapter contributes to wider discussions and trends within public librarianship surrounding methods and methodologies. Notably, it illustrates the value of combining quantitative and qualitative methods; moreover, it adds to the growing trend towards the triangulation of methods.

Lexical patterns that reveal semantic relations

Another way in which linguistic researchers have used concordancers is to search for lexical patterns that reveal underlying semantic relations. In lexicography, understanding the relations between concepts is important for constructing definitions. In LIS, understanding the relations between concepts is important for developing classification schemes and ontologies. For instance, in controlled vocabularies, broader terms and narrower terms are used to indicate hierarchical relations. Meanwhile, in the context of information retrieval applications, semantic

relations help the user to browse concept systems for appropriate search terms and enable query expansion. But how can these relations be identified in a corpus?

It has long been suggested by linguists (e.g. Cruse 1986) that certain lexical patterns have the potential to reveal underlying semantic relations. For example:

- an X is a Y that has characteristics A, B and C ("is a" indicates a hypernymic or generic-specific relation)
- an X has a A, B and C ("has a" indicates a meroynymic or part-whole relation)
- an X is used to A, B and C ("is used to" indicates a functional relation)
- X results in A, B and C ("results in" indicates a causal relation)

In recent years, numerous linguists have contributed to the development of detailed inventories in many languages that list lexical knowledge patterns which reveal underlying semantic relations (e.g. Marshman *et al*., 2002, 2008; Soler and Alcina, 2010; Schumann, 2011). Using the KWIC concordance feature of WordSmith Tools, Bowker (2003) illustrates how corpus search techniques that combine both a search term and a lexical pattern can be used to retrieve knowledge-rich contexts that contain useful information about semantic relations. She then posits that incorporating these inventories and associated search techniques into knowledge discovery tools could enhance information retrieval results for users. This notion is taken up and explored more fully by a number of researchers who contributed to the volume *Probing Semantic Relations* (Auger and Barrière 2010), which presents a state of the art of research trends in the area of knowledge extraction from corpora using linguistic patterns.

Semantic prosody

The term *semantic prosody* was coined by linguist Bill Louw (1993), and it refers to a linguistic phenomenon whereby a lexical item that, in and of itself, does not contain any evaluative meaning, takes on a favourable or unfavourable attitudinal meaning by virtue of the lexical

environment in which it is typically found. For instance, the adjective *habitual* is defined in *Webster's* online dictionary simply as "doing, practicing, or acting in some manner by force of habit". It would seem, therefore, that *habitual* is not inherently negative, since there are plenty of good habits in which people could engage. However, a corpus-based examination of *habitual* (Bowker, 2001) has revealed that this word keeps 'bad company', typically being used to modify lexical items such as *criminal*, *drunk*, *drug user*, *gambler*, *liar*, *thief*, and *offender*. Given that *habitual* appears so frequently in such unfavourable environments, it begins to take on an unfavourable semantic prosody itself, to the extent that it might seem strange or unnatural to encounter this lexical item in a favourable environment. (Warning: If we ever hear someone described as a *habitual librarian*, it may not be intended as a compliment!)

Prosodic meanings are part of the language system, and semantic prosody is not a new phenomenon. However, as pointed out by Louw (1993) and echoed by Stubbs (1995), until recently, semantic prosodies remained largely hidden and therefore relatively unexplored because they are not easily accessible through intuition or introspection. In addition, they cannot always be investigated using quantitative measures such as collocation because semantic prosody does not focus on any specific lexical pairing, but rather on a set of different words that appear in the vicinity of the search term. It is possible that individual members of this set may not be statistically significant collocates; however, when the members of the set are viewed together, the semantic prosody becomes apparent. The availability of corpora and concordancers, which make it possible to study a given lexical item (e.g. *habitual*) in multiple contexts, has made it possible to conduct more detailed investigations of semantic prosody – an example of Digital Humanities in action.

With regard to its application in LIS research, Curado Fuentes (2001) undertakes a corpus-based description of the language used in several different specialized domains, including the domain that he describes as Librarianship and Information Management (LIM). The corpus contains research articles, textbooks and technical papers on the subject of LIM, which are analyzed both quantitatively and qualitatively with the help of the WordSmith Tools corpus analysis software. In addition to using the quantitative techniques described in previous sections (e.g. frequency, collocations, keyness), Curado Fuentes (2001) also carries out a qualitative study of the semantic prosody of the term *access* in LIM by examining this term in multiple contexts in the specialized corpus with the help of the concordancer.

Discussion and implications of corpus linguistics for LIS research

As we have seen in the previous sections, corpus linguistics is not a single method. Rather, it employs a set of different techniques which are related by the fact that they are performed on large collections of electronically stored, naturally occurring texts.

An advantage of adopting a corpus-based approach is that it offers researchers a reasonably high degree of objectivity because it enables them to approach the texts (relatively) free from any preconceived or existing notions regarding their linguistic, semantic or pragmatic content. Many corpus linguistics techniques are quantitative and make use of statistical measures, which are performed by corpus analysis software. However, corpus-based analysis does not merely involve getting a computer to objectively count and sort linguistic patterns or apply statistical algorithms onto textual data. Input from the researchers is required too. For instance, they need to determine which corpus-based techniques are to be applied to the data, and what the cut-off points of statistical significance should be. Then, informed by the quantitative findings, researchers must

decide what is to be studied in more detail, which often includes qualitative analysis (e.g. examining concordance lines). In other words, corpus analysis software can carry be used to carry out preliminary processing, but the evidence must be interpreted by the researcher.

Another major advantage that is offered by corpus methods is that is possible to consult a much large collection of texts than would be feasible in a manual study. As noted above, linguistic phenomena such as collocation and semantic prosody are more easily discerned when analyzing large volumes of data, and they are therefore less accessible when manually analyzing a small number of texts. Findings based on a large data set are likely to be more generalizable and reliable than findings based on a small data set. Of course, a concern when working with large volumes of text is that it may be difficult to know where to focus attention. Fortunately, corpus linguistics techniques can help to provide a sort of map of the corpus, identifying areas of interest for a close analysis. For instance, keywords and collocates can be used as search terms in the concordancer, which will then retrieve KWIC examples for inspection. Meanwhile, the combination of search terms and lexical patterns that reveal semantic relations can be used to identify knowledge-rich contexts within the corpus, which can then be studied in concordance lines.

A corpus-based approach can also be iterative in that these qualitative findings can in turn become a source for further quantitative investigation. For instance, since concordance analysis looks at a known number of concordance lines, the findings of the qualitative analysis can be grouped (e.g. themes relating to a specific word) and then quantified in absolute and relative terms to identify possible patterns (e.g. the tendency of a particular words to be associated with particular themes).

However, corpus-based techniques also have some weaknesses. For instance, corpus analysis tends to focus on what has been explicitly written, rather than on what could have been written, or what is implied, inferred, insinuated, etc. Moreover, pragmatic devices cannot be readily analysed through corpus linguistic means. A research project may therefore require a researcher to step outside the corpus in order to consult other types of information or use additional techniques such as text mining (e.g. Seadle, 2017) to arrive at a fuller understanding of the issues under investigation.

Concluding remarks

Overall, corpus-based techniques are relatively simple, yet also powerful. User-friendly corpus analysis software, such as WordSmith Tools or AntConc, can be used to calculate various measures of frequency, search for words or patterns, and sort and display the results in a variety of formats that make it easier for researchers to interpret the results. While corpus methods alone may not be sufficient for research in LIS, they can be used to help triangulate the findings of other methods.

For instance, in the context of LIS research, corpus methods can be employed as a complement or supplement to content analysis. They can be used as an entry point or at a subsequent stage of the research, thus feeding into a virtuous research cycle. Partington (2003) presents a scalar view of the uses of corpus linguistics as a methodology, which points towards a rationale for using corpus-based methods to carry out content analysis:

> At the simplest level, corpus technology helps find other examples of a phenomenon one has already noted. At the other extreme, it reveals patterns of use previously unthought of. In between, it can reinforce, refute or revise a researcher's intuition and show them why and how much their suspicions were grounded. (12)

Corpus linguistics techniques also have the potential to be exploited more fully in LIS research that involves a higher degree of automation. For instance, corpus linguistics techniques for identifying keyness could potentially enhance recommender systems, while corpus-based methods for identifying semantic relations based on lexical patterns might prove useful for knowledge discovery systems. Corpus-based approaches may also contribute to tasks such as text mining, which combines linguistic, statistical and machine learning techniques.

In closing, corpus linguistics can be another tool to add to the LIS researcher's methodological toolbox. Moreover, given that its techniques support both quantitative and qualitative analyses, corpus linguistics responds to Gauchi Risso's (2016) recent call for a methodology that has both qualitative and quantitative aspects. It has been twenty-five years since corpus-based methods took a firm hold in the field of linguistics, and we can now safely say that corpus linguistics is not just for linguists anymore. We hope this paper will stimulate discussion about the potential of corpus techniques for enhancing LIS research in a variety of ways. Don't become a "*habitual* LIS researcher" – give corpus linguistics a try!

Acknowledgements

An unpublished version of this paper was selected as the winner of the 2018 ALISE/ProQuest Methodology Paper Competition, and thanks are due to the competition reviewers for their helpful feedback and encouragement, as well as to the anonymous reviewers of *Library Hi Tech*. We are grateful to Mike Scott, developer of the WordSmith Tools corpus analysis software package, for permission to use screenshots. As stated on the WordSmith Tools site, "For non profit-making academic use: No need to ask. You are hereby granted permission." (http://lexically.net/publications/copyright_permission_for_screenshots.htm)

References

Anthony, L. (2006), "Concordancing with AntConc: An Introduction to Tools and Techniques in Corpus Linguistics", in *Proceedings of the Japan Association of College English Teachers (JACET) 45th Annual Convention*, pp. 218-219.

Anguiano Peña, G. and Naumis Peña, C. (2015), "Extraction of candidate terms from a corpus of non-specialized, general language", *Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información*, Vol. 67, pp. 19-45. DOI: 10.1016/j.ibbai.2016.04.002

Auger, A. and Barrière, C. (Eds). (2010), *Probing Semantic Relations: Exploration and Identification in Specialized Texts*, John Benjamins, Amsterdam/Philadelphia.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T. and Wodak, R. (2008), "A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press", *Discourse and Society*, Vol. 19 No. 3, pp. 273-306. DOI: 10.1177/0957926508088962

Beel, J., Gipp, B., Langer, S. and Breitinger, C. (2016), "Research-Paper Recommender Systems: A Literature Survey", *International Journal of Digital Libraries*, Vol. 17 No. 4, pp. 305-338. DOI: 10.1007/s00799-015-0156-0

Biber, D., Conrad, S. and Reppen, R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.

Bowker, L. (2003), "Lexical knowledge patterns, semantic relations, and language varieties: Exploring the possibilities for refining information retrieval in an international context", *Cataloging and Classification Quarterly*, Vol. 37 No. 1/2, pp. 153-171.

Bowker, L., Mackay, R., Kasama, D. and Buitrago Ciro, J. (2015), "Different views of textual 'aboutness': A recipient evaluation of the content descriptors proposed by professional indexers,

authors, readers and corpus analysis tools", in *Proceedings of the 43rd Annual Conference of the Canadian Association of Information Science*, Ottawa, June 3-6, 2015, available at: https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/921 (accessed 7 December 2017).

Bowker, L. and Pearson, J. (2002), *Working with Specialized Language: A Practical Guide to Using Corpora*, Routledge, London/New York.

Curado Fuentes, A. (2001), "Lexical Behaviour in Academic and Technical Corpora", *Language Learning and Technology* Vol. 5 No. 3, pp. 106-129.

Cruse, D. A. (1986), *Lexical Semantics*, Cambridge University Press, Cambridge.

Drouin, P. (2003), "Term extraction using non-technical corpora as a point of leverage", *Terminology* Vol. 9 No. 1, pp. 99-115.

Gauchi Risso, V. (2016), "Research methods used in library and information science during the 1970-2010", *New Library World*, Vol. 117 No. 1/2, pp. 74-93. DOI: 10.1108/NLW-08-2015-0055

Greene, M. and McMenemy, D. (2012), "The Emergence and Impact of Neoliberal Ideology on UK Public Library Policy, 1997-2010", in Spink, A. and Heinström, J. (Eds), *Library and Information Science Trends and Research: Europe*, Emerald Group Publishing Ltd., Bingley, UK, pp. 13-42.

Hider, P. and Pymm, B. (2008), "Empirical research methods reported in high-profile LIS journal literature", *Library and Information Science Research* Vol. 30 No. 2, pp. 108-114.

International Federation of Library Associations and Institutions (IFLA). (2013), *Riding the Waves or Caught in the Tide? Navigating the Evolving Information Environment. Insights from*

the IFLA Trend Report. IFLA, The Hague, available at: https://trends.ifla.org/ (accessed 22 February 2018).

Kehoe, A. and Gee, M. (2011), "Social tagging: A new perspective on textual 'aboutness'", *Studies in Variation, Contacts, and Change in English: Methodological and Historical Dimensions of Corpus Linguistics*, Vol. 6, available at: http://www.helsinki.fi/varieng/series/volumes/06/kehoe_gee/ (accessed 7 December 2017).

Kennedy, G. (1998), *An Introduction to Corpus Linguistics*, Longman, London.

Marshman, E., Morgan, T. and Meyer, I. (2002), "French Patterns for Expressing Concept Relations", *Terminology*, Vol. 8 No. 1, pp. 1-29.

Marshman, E., L'Homme, M.-C. and Surtees, V. (2008), "Portability of cause-effect relation markers across specialized domains and text genres: A comparative evaluation", *Corpora*, Vol. 3 No. 2, pp, 141–172.

McEnery, T. and Wilson, A. (1996), *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Nazarenko, A. and Aït El Mekki, T. (2007), "Building back-of-the-book indexes?", in Ibekwe-SanJuan, F., Condamines, A. and Cabré Castellví, M. T. (Eds), *Application-Driven Terminology Engineering*, John Benjamins, Amsterdam/Philadelphia, pp. 179-202.

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

Partington, A. (2003), *The Linguistics of Political Argumentation: The Spin-doctor and the Wolf-pack at the White House*, Routledge, London.

Schumann, A.-K. (2011), "A Case Study of Knowledge-Rich Context Extraction in Russian," in *Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*,

Paris, 8-10 November 2011, pp. 143-146, available at: http://tia2011.crim.fr/tia2011/Proceedings/index.html (accessed 7 December 2017).

Scott, M. (2015), *WordSmith Tools Manual*, version 6.0, available at: http://www.lexically.net/downloads/version6/HTML/index.html?getting_started.htm (accessed 7 December 2017).

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

Seadle, M. (2017), "Text box: Text Mining," in Silipigni Connaway, L. and Radford, M., *Research Methods in Library and Information Science, 6th Edition*, Libraries Unlimited, Santa Barbara, CA, p. 350.

Sinclair, J. (Ed.). (1987), *Looking Up: An Account of the COBUILD Project in Lexical Computing*, Collins, London.

Soler, V. and Alcina, A. (2010), "Patrones léxicos para la extracción de conceptos vinculados por la relación parte-todo en español", in Auger, A. and Barrière, C. (Eds), *Probing Semantic Relations: Exploration and identification in specialized texts*, John Benjamins, Amsterdam/Philadelphia, pp. 97-120.

Stubbs, M. (1995), "Corpus evidence for norms of lexical collocation", in Cook, G. and Seidlhofer, B. (Eds), *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson*, Oxford University Press, Oxford, pp. 245-256.

Turcios, M E., Agarwal, N. K. and Watkins, L. (2014), "How much of library and information science literature qualifies as research?", *Journal of Academic Librarianship*, Vol. 40 No. 5, pp. 473-479. DOI: 10.1016/j.acalib.2014.06.003

Webster's Online Dictionary, available at: https://www.merriam-webster.com/dictionary/habitual (accessed 7 December 2017).

White, M. D. and Marsh, E. E. (2006), "Content Analysis: A Flexible Methodology", *Library Trends* Vol. 55 No. 1, pp. 22-45.

Corpus analysis tools mentioned in this paper

AntConc, available at: http://www.laurenceanthony.net/software/antconc/ (accessed 7 December 2017).

TermoStat, available at: http://olst.ling.umontreal.ca/?page_id=91 (accessed 7 December 2017).

WordSmith Tools, version 6, available at: http://www.lexically.net/ (accessed 7 December 2017).