doi: 10.15503/jecs20182.154.164

Using corpora to aid qualitative text analysis. An interdisciplinary approach

Jędrzej Olejniczak

Faculty of Letters
University of Wrocław, Plac Uniwersytecki 1, Wrocław
Email address: jedrzej.olejniczak@gmail.com

Abstract

Aim. The aim of this paper is to present and exemplify a number of basic uses of corpus-based text analysis tools that can supplement and provide additional insight for an otherwise qualitative analysis of a text. I attempt to show that nowadays certain corpus tools are easily accessible to any researcher and can be used to enrich the results of studies concerned with texts.

Methods. This paper comprises the basics of corpus building, the main types of data that can be drawn from a simple corpus and a detailed description of four methods that can aid text analysis: wordlists, concordances, dispersion plots, and keywords. Each of those four methods is thoroughly described, including a number of examples of its applications and indicates its possible limitations.

Results. The examples provided suggest that even performing a very simple corpus analysis of a text might unveil certain trends and phenomena not noticeable through the classic qualitative text analysis methods (*e.g.* close reading). The paper argues that corpus research can hence work as an extension of a quantitative analysis (or be its starting point) by examining themes and keywords present in a given text and enrich the results of a qualitative study with a fresh perspective. Finally, the paper claims that basic corpus analysis can, in fact, be successfully employed by researchers who do not have any prior experience with statistics or corpora.

Key words: Corpora, text analysis, concordance, wordlist, keyness, dispersion plot, corpus building

Introduction

A number of disciplines use texts as a resource for analysing certain phenomena, *e.g.* history, archeology, sociology, journalism, literary studies, linguistics, and translation studies. Many disciplines within the humanities limit themselves to qualitative research, however. Quantitative corpus analyses of texts that are performed in the interdisciplinary intersections between corpus linguistics and other disciplines (*e.g.* digital humanities, cultural analytics, text mining) are to a great extent done by those who specialise in the field of corpus linguistics. I would like to discuss the idea that due to the evolution of accessibility of the corpus tools, using them within the basic scope of text analysis no longer requires advanced computing skills or a great amount of knowledge of linguistic frameworks. Though these skills do obviously allow for more intricate analyses, they are no longer an absolute necessity.

The advantages of being able to conduct a quantitative analysis of language are quite remarkable. Recent research in clinical psychology has discovered ways of identifying depression based on studying how individuals used absolutist words (Al Mosaiwi, 2018, pp. 1-14). The corpus-based analyses of Jane Austen's works unveil and highlight, among other features of the texts, a number of previously undiscovered strategies used by the writer to establish a relationship with her readers (Fischer-Starcke, 2010, pp. 74-87). Rybicki uses corpora to investigate a large number of translated texts in order to examine whether it is possible to identify each individual translator's style (Rybicki, 2012, pp. 231-248). O'Sullivan, Bazarnik, Eder, and Rybicki inspect the stylistic influences of James Joyce on Flann O'Brien via qualitative methods and find a number of strong relationships between the two writers texts (2018, pp. 1-25).

As enumerated above, examining patterns in language can reveal many fascinating facts about texts. Most of these findings are normally unavailable to the naked eye: what characterises any given writer's style, what patterns they follow (they always do!), how they structure their texts and what makes them different from the others. Even though some of these features are concealed and require advanced knowledge in statistics and linguistics to be properly explored, a vast chunk of data obtained via corpus analysis of texts is rather transparent and accessible to any researcher willing to take their chances. The aim of this paper is to discuss these more mundane applications of corpus research and to present a number of readily available approaches that can be used in text analysis. The main purpose of the methods discussed herein is to facilitate and enhance close-reading through accessing and exploring keywords, themes, and patterns in texts and through discovering new research perspectives.

A general overview of readily available software

Having a background in corpus linguistics, I used various corpus analysis programs throughout my research. Though my assessment of the available software is no doubt subjective, I believe that both Laurence Anthony's AntConc (Version 3.5.7, Anthony, 2018) and Mike Scott's WordSmith Tools (Version 7.0, Scott, 2016) are fairly intuitive and provide a wide array of corpus analysis tools.

AntConc is a freeware program that includes all the essential tools for conducting text analysis and obtaining statistics from a corpus. AntConc's huge asset is the transparency, simplicity, and accessibility of its interface, which make the program easy to learn and user-friendly. WordSmith Tools have to be purchased but come with a number of additional, advanced functions allowing its users to customise and refine their search queries. At their core, both programs offer the same types of tools: concordancing, word frequency lists, dispersion plots, keyness analysis, and collocation/word cluster sampling. The use of all these tools has been described and exemplified in this paper, whereas both programs include very detailed manuals regarding the access to and the proper use of those tools.

Building a corpus for text analysis

Corpus Size and Representativeness

The issue of corpus representativeness has been raised by many theorists. Within the framework of linguistic research, corpora need to be large enough to be representative for the particular phenomena being studied (Sinclair, 2005, p.4). However, a corpus used to aid a particular (*e.g.* literary) study can contain just a single text. In his paper on corpus sampling, Biber (1993) notes that "[d] ifferent overall corpus designs represent different populations and meet different research purposes" (p. 244). Biber (1993) also notes that the size of the corpus depends entirely on the hypotheses and research problems (pp. 244-245). When a corpus is indeed testing hypotheses about language, certain constraints are in place. If the purpose of a corpus is to simply examine a singular text to capture its lexical peculiarities, or to compare it with another text of similar type, a text of any size can be considered a complete corpus in its own right (Baker, 1993, pp. 244-245).

As all these theorists point out, however, the corpus size does indeed restrict the number of methods that can be used in research. While a single average-sized novel or a play should provide multiple research perspectives, a short story or a poem will usually not suffice. Corpus methods are quantitative in their core and thus require larger chunks of data.

The above issues can be alleviated by expanding the corpus with additional texts. For instance, much can be unveiled about a given poet's style and inclinations when a corpus based on their work becomes large enough. Phenomena such as lexical preferences (*e.g.* keywords that indicate particular recurring themes) or a structure that is repeated by a given poet throughout their work can very often lead to discoveries inaccessible through qualitative approaches.

Text Sampling and Corpus Cleaning

Basic corpus-based analysis methods discussed in this paper do not involve any elaborate corpus management skills. All corpus analyses do, however, require a digital copy of the texts that are to be included in the corpus. A vast number of texts is nowadays available online for free, for instance through the websites such as Project Gutenberg ("Project Gutenberg", 2018). The obtained corpus material needs to be saved as a text file (.txt format) and cleaned, which entails removing all paratextual information (*e.g.* cover text, editorial pages, footnotes, foreword). This ensures that no external data disrupts the results.

Corpus sampling should not be a large concern either. When a single, average (10 000 words and above) literary text is being examined, its full content will usually satisfy the researcher's corpus-building goals. Keyness tests (to be discussed herein) will, however, require another text of similar size. When the research goal is to examine shorter texts of a given author (*e.g.* poems, short stories), the research will only be viable when a larger collection of these texts has been accumulated (perhaps around 5 000 words or more).

Data is beautiful: corpus tools and their uses in aiding text analysis

Word Frequency Lists/Wordlists

A wordlist created from a corpus represents the frequency of appearance of every word therein (Scott, 2010a). Tools such as AntConc or WordSmith tools generate these lists almost instantly and hence a wordlist is an easily accessible but a very raw resource. An example of a wordlist is shown in Fig. 1.

The ability to filter through the overwhelming amount of data present in a wordlist can reveal a number of interesting facts about the text and its author. Certain themes present in a given text can be discovered on the wordlist level and then explored qualitatively.

Grammatical and Lexical Words. Understanding the distinction between these two classes is crucial for analysing corpora within the framework proposed in this paper. According to Quirk and Greenbaum (1973), lexical words (or open-class items) comprise verbs, adjectives, nouns, and most adverbs, the distinctive feature of the class being its ability to expand. Each lexical word also carries its own independent meaning. By contrast, grammatical words include the parts of speech such as prepositions, pronouns, conjuncts, auxiliary verbs, articles, and particles. Grammatical words (or closed-class items) do not carry their own meaning and are used to form structures within sentences or to add to the meaning of the lexical words (Quirk, & Greenbaum, 1973, pp. 18-20).

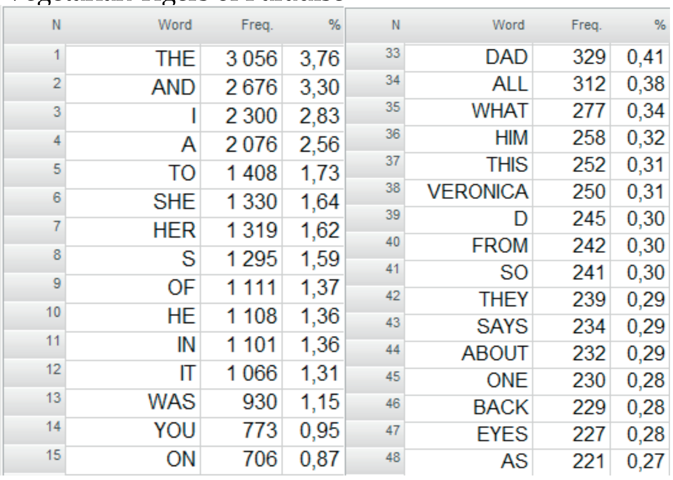
Though wordlists contain both grammatical and lexical words, only the latter can be expected to provide meaningful information when the corpus is rather small. Grammatical words are by no means irrelevant as they allow for the study of language structures used across the corpus. For these structures to reoccur on a statistically reliable level, however, a corpus needs to contain at least 1.4 million words (Biber, 1993, pp. 243-257).

Grammatical words, due to the fact that their inventory is limited, will always occupy the top of almost every wordlist. Nonlinguistic analyses can safely ignore them and explore the lexical words instead.

Lexical words. Lexical words might reveal, among other things, themes, motifs, peculiarities in the use of adjectives or adverbs, and deliberate or accidental overuse of certain words. These qualities, though present in the text, are usually visible only at the statistical level. Corpora provide an insight into their patterns of use and distribution, potentially giving the researcher a number of new ideas to explore.

I would like to illustrate the use of this tool with an analysis of Crystal Jeans's (2016) novel *Vegetarian Tigers of Paradise*. The author sets out to describe her childhood, her adolescent years and adulthood in Cardiff's Paradise Place in the 1990s, focusing on her relationship with her mother, her sister, and her friend Jessica. The book gives the reader a bitter taste of Jeans's everyday life and of the challenges she had to face when growing up. I decided to choose *Vegetarian Tigers* as the main source of examples for this paper as the patterns present in the book are very pronounced, providing solid examples of the applications of corpus tools. The novel has also never been analyzed via corpus tools before. The wordlist created using the book's full content is shown in Fig. 1.

Figure 1. Sections of the wordlist created from the corpus built from the full text of "Vegetarian Tigers of Paradise"



N	Word	Freq.	%	N	Word	Freq.	%
1	THE	3 056	3,76	33	DAD	329	0,41
2	AND	2 676	3,30	34	ALL	312	0,38
3	I	2 300	2,83	35	WHAT	277	0,34
4	A	2 076	2,56	36	HIM	258	0,32
5	TO	1 408	1,73	37	THIS	252	0,31
6	SHE	1 330	1,64	38	VERONICA	250	0,31
7	HER	1 319	1,62	39	D	245	0,30
8	S	1 295	1,59	40	FROM	242	0,30
9	OF	1 111	1,37	41	SO	241	0,30
10	HE	1 108	1,36	42	THEY	239	0,29
11	IN	1 101	1,36	43	SAYS	234	0,29
12	IT	1 066	1,31	44	ABOUT	232	0,29
13	WAS	930	1,15	45	ONE	230	0,28
14	YOU	773	0,95	46	BACK	229	0,28
15	ON	706	0,87	47	EYES	227	0,28
				48	AS	221	0,27

Source: own research. Generated with WordSmith Tools (Version 7.0, Scott, 2016)

The "Freq.: column represents the number of appearances of a given words in corpus whereas the "%" column indicates the percentage of the book's volume that this word occupies. The table to the left represents the most common 15 words appearing in the text; as mentioned before, grammatical words prevail in the highest ranks and for the purposes of this analysis can be ignored. The table to the right represents the ranks 33-48, in which first lexical words such as nouns make their appearance. One of the most frequent lexical words that appears in the text is hence the plural noun "eyes" (the other words being nouns that are used to refer to the characters in the book, *e.g.* "Veronica," "mum," and "dad").

Though the fact that the word "eyes" is used so commonly in the text does not seem to be in any way remarkable, it can be used as a starting point of a qualitative analysis using the concordance tool (see Concordance lists section below). It can also be easily checked against other corpora to see how often other writers use this term.

According to the literary segment of British National Corpus (Davies, 2004) for the years 2000 and later, the weighted volume of "eyes" is 0.13%, which means that on average it is more than twice less common than in *Vegetarian Tigers*. Testing particular literary works from different periods and different genres gives similar results; be it Joseph Heller, Jane Austen, Terry Pratchett, Nicolas Sparks, David Lodge, or Saul Bellow, their interest in their characters' eyes gravitates around 0.10%. It may be hence concluded that the high volume of the word "eyes" is somewhat extraordinary and perhaps requires some attention.

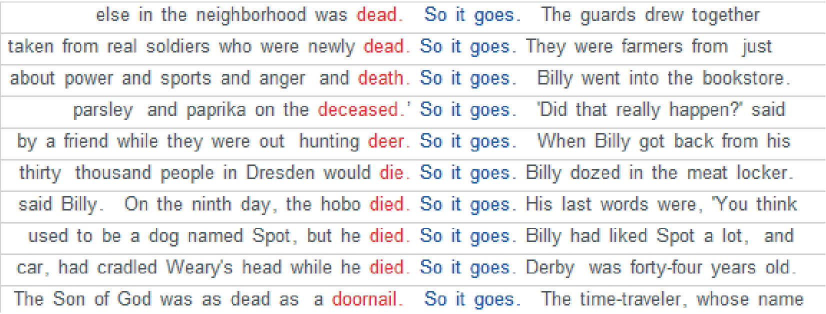
Many other words also stand out in the wordlist based on *Vegetarian Tigers*, in particular various words indicating strong language and nouns referring to parent figures. The strong language distribution is discussed herein in the dispersion plot section.

Concordance lists

A concordance list, being a list of all appearances of a given word or phrase and their immediate or extended contexts within a corpus, is the most straightforward corpus tool available. Despite that, it may find multiple uses in text analysis. The ability to cross-examine all the appearances of a keyword or phrase that interests the researcher can drastically improve the efficiency of qualitative research while possibly capturing certain recurring patterns (Rauscher, Swiezinski, Riedl, & Biemann, 2013, pp. 66-69).

The words or phrases can be selected for concordance analysis through qualitative analysis (*e.g.* keywords or core concepts discovered through close reading). In this case, a concordance list will greatly facilitate the speed at which the text is analysed and provide a lot of material for further examination. A stellar example of this approach could be the analysis of Kurt Vonnegut's *Slaughterhouse-Five* (Vonnegut, 1969) and the phrase "so it goes" that continuously reoccurs throughout the text. The concordance list for the phrase generated in WordSmith Tools looks as follows:

Figure 2. A fragment of concordance list for the phrase "so it goes" in Kurt Vonnegut's *Slaughterhouse Five*



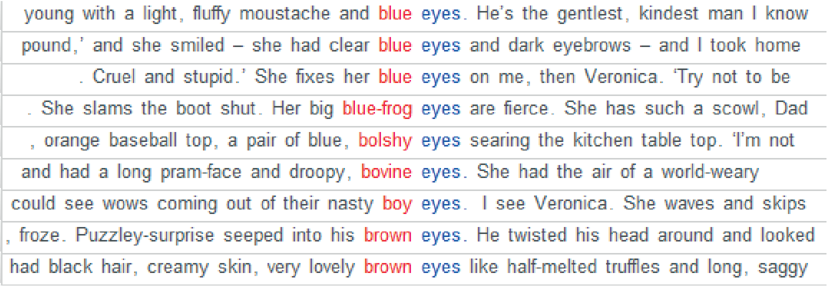
else in the neighborhood was dead.	So it goes.	The guards drew together
taken from real soldiers who were newly dead.	So it goes.	They were farmers from just
about power and sports and anger and death.	So it goes.	Billy went into the bookstore.
parsley and paprika on the deceased.'	So it goes.	'Did that really happen?' said
by a friend while they were out hunting deer.	So it goes.	When Billy got back from his
thirty thousand people in Dresden would die.	So it goes.	Billy dozed in the meat locker.
said Billy. On the ninth day, the hobo died.	So it goes.	His last words were, 'You think
used to be a dog named Spot, but he died.	So it goes.	Billy had liked Spot a lot, and
car, had cradled Weary's head while he died.	So it goes.	Derby was forty-four years old.
The Son of God was as dead as a doornail.	So it goes.	The time-traveler, whose name

Source: own research. Generated with WordSmith Tools (Version 7.0, Scott, 2016)

The phrase appears 99 times throughout Vonnegut's text. It appears in a number of contexts but what concordance analysis makes apparent is the fact that "so it goes" is associated with death and disease. All the uses express how death is sudden and inevitable and how much of a routine death has become to those who participated in the War. This aspect of the text is now open for an in-depth analysis explaining the particular details of Vonnegut's deliberate strategy and the structure of the text.

With concordance list such as the one in *Fig. 2.*, each particular appearance of the expression can be expanded and thoroughly inspected, allowing the researcher to focus on the details in their analysis. Another interesting aspect of concordances is that they can be graphically represented with the use of so called concordance plots or dispersion plots (also discussed herein), which allow to graphical tracking of the clusters of occurrences of a given word or phrase within a given corpus. This gives the researcher a possibility to assess the way in which a text is structured with regard to a given theme.

Figure 3. A section of concordance list for the word "eyes" in Crystal Jeans's *Vegeterian Tigers of Paradise*



```
   young with a light, fluffy moustache and blue eyes. He's the gentlest, kindest man I know
  pound,' and she smiled – she had clear blue eyes and dark eyebrows – and I took home
           . Cruel and stupid.' She fixes her blue eyes on me, then Veronica. 'Try not to be
  . She slams the boot shut. Her big blue-frog eyes are fierce. She has such a scowl, Dad
  , orange baseball top, a pair of blue, bolshy eyes searing the kitchen table top. 'I'm not
  and had a long pram-face and droopy, bovine eyes. She had the air of a world-weary
could see wows coming out of their nasty boy eyes.  I see Veronica. She waves and skips
, froze. Puzzley-surprise seeped into his brown eyes. He twisted his head around and looked
had black hair, creamy skin, very lovely brown eyes like half-melted truffles and long, saggy
```

Source: own research. Generated with WordSmith Tools (Version 7.0, Scott, 2016)

As mentioned before, resources for concordancing can be drawn from lexical words appearing in wordlists. In *Vegeterian Tigers*, I observed a high density of the noun "eyes." An analysis of concordance list based on the word "eyes" (see *Fig. 2.*) reveals that eyes are in fact an inherent component of Jeans's dialogues and appear almost exclusively therein. Close examination of the concordances then allows us to draw a number of conclusions regarding the internal layout of the text. When introducing or describing her characters, Jeans is rather reluctant. She does not go into much detail, focusing on two or three distinctive or defining features of appearance or traits of character, usually within the limits of a single sentence. When presenting her characters' appearance, Jeans mentions their eyes extremely often, normally in the first line of the description itself. Perhaps coincidentally, Jeans assigns various shades of blue to the eyes most of her characters: Mum, Veronica, Dad, Neil, Tim, Janice, April, Sidney, Chantelle, and even Kev the Milkman all happen to have blue eyes even though only three of them are actually related. More interestingly, when a character's eyes are not blue, they are rarely mentioned at all. Characters' gaze is also a focal point of the dialogues. What the characters do with their eyes is deliberately mentioned in the majority of human interactions described in the book: they cover them, open them, roll them and close them (those are the most common verb collocations of the noun "eyes"). Together with the eyes but to a lesser extent, face and facial expressions are also mentioned in the descriptions. In contrast with that, the author of *Vegetarian Tigers* pays very little attention to body movement or the clothes of her characters.

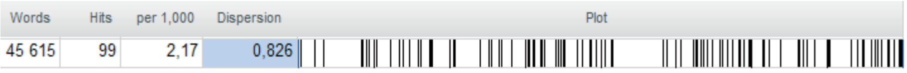
An excellent example of what a concordance and wordlist analysis can produce can be found in the monograph by Bettina Fischer-Starcke (2010) *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. Fischer-Starcke discusses the word "heroine," which stood out as a very frequent word in the wordlist. The word is used to refer to the book's protagonist, Catherine, who, as Fischer-Starcke aptly notes, has in no way been referred to as a heroine by literary theorists; the perception of her has been quite the opposite (Fischer-Starcke, 2010, pp. 74-77). Through a concordance-based examination conducted thanks to this finding, Fischer-Starcke established the word is in fact used as a tool used by Austen to connect with the reader using the colligation "our heroine" (Fischer-Starcke, 2010, pp. 86-87). This, in turn, lead Starcke to close-examining other means through which Austen referred to the readers throughout her books.

Concordances can be used to conduct in-depth analysis of a given concept or notion across a single text or a number thereof. Rauscher, Swiezinski, Riedl, and Biemann use their own concordance tool to explore the descriptions and associations related to four large cities (Frankfurt, Dortmund, Birmingham, & Glasgow) across over 240 crime novels. They seek to confirm the hypothesis that no common notion of large city exists and that every city appearing in literature will carry its own peculiar, recurring set of associations. Being presented with a large amount of data, they are indeed able to discover associations shared across numerous literary works that are particular to every given city examined (Rauscher, Swiezinski, Riedl, & Biemann, 2013, p. 64-67).

Studying Concordance/Dispersion Plots

Dispersion plot is a graphical representation of all the recurrences of a given word or phrase throughout the corpus. When examined, dispersion might reveal a marked distribution pattern within a given corpus. Application of such patterns are fairly vast: they can serve to illustrate the development of a certain known theme throughout the text, unveil structural elements of a text or track the development of certain ideas. Dispersion plots can also be used to track multiple words/phrases (*e.g.* words centered around a common theme) at the same time. Dispersion itself is a mathematical value than ranges from 0 to 1 and represents how uniform the distribution of a given word across the text is. The dispersion value of 0.9 to 1 indicates high uniformity whereas the value between 0 and 0.1 indicates that the distribution is particularly uneven (Scott, 2010b). Values below 0.8 indicate that the distribution is somewhat uneven, suggesting certain deliberate structure within the text.

Figure 4. The dispersion plot of the expression "so it goes" in Vonnegut's *Slaughterhouse Five*



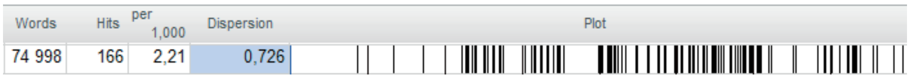
Words	Hits	per 1,000	Dispersion	Plot
45 615	99	2,17	0,826	

Source: own research. Generated with WordSmith Tools (Version 7.0, Scott, 2016)

I would like to draw the first example from Vonnegut's *Slaughterhouse-Five*. As I established above, the expression "so it goes" is used therein mostly in relation to death. In dispersion plots, black segments represent all the appearances of a given word or expression throughout the text. Tracking the pattern inside the text (which is fairly uniform with dispersion value being slightly over 0.8) reveals that the consistently white areas (where the expression does not appear) represent the protagonist's "escapes" to Tralfamadore.

Dispersion can be also used to track the development of certain deliberate themes or structures present in a given text. For instance, the examination of the use of strong language in *Vegeterian Tigers* reveals a logical pattern. At first glance, it appears that Jeans's characters swear very often and, in fact, quite consistently. A dispersion plot in Fig. 5. indicates that the author made a very deliberate decision as to the distribution of strong language, however:

Figure 5. Dispersion plot for strong language in *Vegetarian Tigers of Paradise*



Words	Hits	per 1,000	Dispersion	Plot
74 998	166	2,21	0,726	

Source: own research. Generated with WordSmith Tools (Version 7.0, Scott, 2016)

The first two chapters of the book refer to the protagonist's childhood, where she does not at first notice the use of strong language at all. The harsh reality does, however, gradually find its way to the protagonist's world (and later speech as well). Tracking strong language in concordance list reveals another feature. Initially, strong language exclusively appears in the dialogues, used by characters that surround the protagonist. That language then gradually devolves and increases in density to eventually find its way into the protagonist's speech, thoughts and narration. It is particularly excessive and uncalled for in the chapters describing her teenage years and then it is used commonly, though more sensibly throughout the remainder of the book. The gaps where no strong language occurs represent the chapters which go back to the protagonist's childhood.

Keyness tests in comparative text analysis

The final component of corpus analysis I should like to discuss is the keyness factor. Keyness tests are conducted in order to discover the differences in vocabulary between two wordlists. This entails that keyness is therefore used exclusively in comparative analyses as to highlight the contrasts between two different texts or two parts of the same text: the so called main corpus and the reference corpus. The results of keyness analysis will highlight keywords: the words and phrases that stand out in the main corpus (Stubbs, 2010, p. 21).

Keyness itself is a figure obtained through a G-Test, which is closely related to Chi-square test but is more accurate when the sizes of the samples compared differ (Scott, 2010c). Though the test accounts for uneven sizes of corpora, it is important not to cross logical boundaries when conducting the test (*i.e.* remaining within the same genre and making educated choices as to which texts are being compared, similarly to qualitative comparative analyses). The keyness coefficient is calculated for every word present in the main corpus; generally, the higher the coefficient of a given word, the more it stands out.

I personally believe that in terms of text analysis, keyness tests find their most successful uses in the analyses of multiple translations of the same source text. They facilitate the discovery of the different approaches two translators may have taken with respect to a single source text. A keyness analysis will very easily highlight the differences in translation choices of proper names (as they will be completely unique and thus receive very high keyness scores) as well as particular idiolectual features of a translator's language.

Denise Milizia (2010) in her paper "Keywords and Phrases in Political Speeches" analyses the keywords in Tony Blair's and George Bush's speeches. Since she uses texts that are naturally rather concise, her corpora consisted of all the speeches given by both politicians released by the White House and by 10 Downing Street (p. 129). Milizia's analysis focuses on the keyness of the collocations of certain phrases, such as "people." The most common frame of reference in this context was "the American people" for Bush and "the Iraqi people" and "the European people" for Blair, perhaps pointing to the internal versus external foci of the two politicians. Another large contrast unveiled by the analysis was how often Blair would refer to the concept of "climate change" compared to Bush, which did lead to the further exploration of the topic by the author (Milizia, 2010, pp. 135-142).

Conclusions

Focusing on the use and distribution of lexical words through wordlists, concordances, dispersion plots and keyness tests can reveal a number of facts about a given text or a group of texts. The results of corpus research can be analysed and discussed as an additional component of an otherwise qualitative analysis or to feed the qualitative analysis with new ideas and perspectives. These corpus methods can be used in almost every discipline concerned with texts and reveal details, themes, motifs, and ideas that might have gone unnoticed in a qualitative analysis. Even though the more advanced methods (such as stylometry or grammatical word analysis) are definitely going to provide even more results, they are not necessary within the scope of every corpus analysis.

References

1. Al-Mosaiwi, M., & Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 1-14. https://doi.org/10.1177/2167702617747074
2. Anthony, L. (2018). *AntConc* (Version 3.5.7) [Computer Software]. Tokyo: Waseda University. Available from http://www.laurenceanthony.net/software
3. Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In Baker, M., Francis, G. and Tognini-Bonelli, E. (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233-250). Amsterdam: John Benjamins Publishing Company.
4. Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing 8(4)*, 243-257. https://doi.org/10.1093/llc/8.4.243
5. Davies, M. (2004-). *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available online at https://corpus.byu.edu/bnc/
6. Fischer-Starcke, B. (2010). *Corpus Linguistics in Literary Analysis: Jane Austen and her Contemporaries*. London: Continuum Publishing.
7. Jeans, C. (2016) *Vegetarian Tigers of Paradise*. Aberystwyth: Honno Welsh Women's Press.
8. Milizia, D. (2010). Keywords and Phrases in Political Speeches. In Bondi, M., Scott, M. (Eds.), *Keyness in Texts* (pp. 127-145). Amsterdam/Philadelphia: John Benjamins Publishing Company.
9. O'Sullivan, J., Bazarnik, K., Eder, M., & Rybicki, J. (2018). Measuring Joycean Influences on Flann O'Brien. *Digital Studies*, *8(1)*, 1–25. https://doi.org/10.16995/dscn.288
10. Project Gutenberg. (n.d.). Retrieved on 24 April 2018 from www.gutenberg.org.
11. Quirk, R., & Greenbaum, S. (1973). *A University Grammar of English*. London: Longman.
12. Rauscher, J., Swiezinski, L, Riedl, M., & Biemann, C. (2013). Exploring Cities in Crime: Significant Concordance and Co-occurrence in Quantitative Literary Analysis. In Kazantseva, A. & Szpakowicz, S. (Eds.), *Proceedings of the Computational Linguistics for Literature Workshop at NAACL-HLT 2013* (61-71). Atlanta, GA, USA: Association for Computational Linguistics.
13. Rybicki, J. (2012). The great mystery of the (almost) invisible translator: stylometry in translation. In Oakley, M. and Ji, M. (Eds.), *Quantitative Methods in Corpus-Based Translation Studies* (pp. 231-248). Amsterdam: John Benjamins.
14. Scott, M. (2010a). *WordSmith Tools Manual*. Retrieved on 16 April 2018 http://www.lexically.net/downloads/version5/HTML/index.html?wordlist_overview.htm
15. Scott, M. (2010b). *WordSmith Tools Manual*. Retrieved on 12 April 2018 http://www.lexically.net/downloads/version5/HTML/?dispersion_basics.htm
16. Scott, M. (2010c). *WordSmith Tools Manual*. Retrieved on 24 April 2018 from http://www.lexically.net/downloads/version5/HTML/index.html?keyness_definition.htm
17. Scott, M. (2016). *WordSmith Tools Version 7, Stroud: Lexical Analysis Software*. Available from http://lexically.net/wordsmith/
18. Sinclair, J. (2005). Corpus and text - basic principles. In Wynne, M. (Ed.) *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books.
19. Stubbs, M. (2010). Three Concepts of Keywords. In Scott, M. and Bondi, M. (Eds.), *Keyness in Texts* (pp. 21-42). Amsterdam: John Benjamins Publishing.
20. Vonnegut, K. (1969). *Slaughterhouse-Five or the Children's Crusade*. New York: Bantam Doubleday Dell Publishing Group.