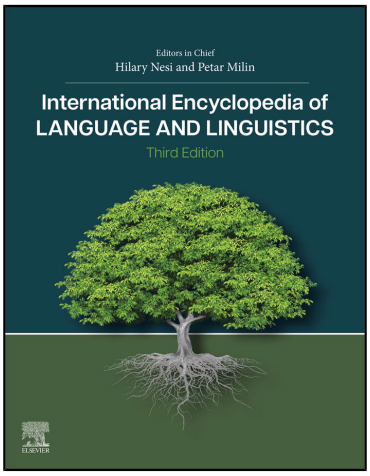

Cvrček, V. (2026). Keyword Analysis. In: Nesi, H., Milin, P. (Eds.), *International Encyclopedia of Language and Linguistics, 3e*. Vol 14., pp. 72–78. UK: Elsevier. https://dx.doi.org/10.1016/B978-0-323-95504-1.01263-1.
ISBN: 9780323955041

Keyword Analysis

Václav Cvrček, Department of Linguistics, Charles University, Faculty of Arts, Prague, Czech Republic

Key Points

- Defines and explains keyword analysis (KWA), a corpus-based method for extracting keywords (prominent linguistic unites) from text.
- Outlines methodological choices, including units of analysis, reference corpora, and statistical approaches.
- Discusses prominence of units in texts and how it can be measured via statistical methods and interpreted.
- Describes uses of keyword analysis in discourse analysis, literary interpretation, including possible extensions beyond lexical level.

Abstract

Keyword Analysis (KWA) is a quantitative method from corpus linguistics used to identify the statistically significant words that characterize a text. By comparing word frequencies in a target text against a large reference corpus, KWA pinpoints a text's distinctive themes and style, providing a data-driven foundation for analysis in fields such as literary studies and discourse analysis. This entry outlines the method's core principles, covering the step-by-step methodology for identifying keywords, common techniques for their interpretation, related advanced concepts, and a survey of key software tools.

Introduction

Keyword Analysis (KWA) is a corpus-based method used to identify the characteristic vocabulary of a text. It operates by statistically comparing word frequencies in a specific text (the target text) against their frequencies in usually a much larger collection of general language (the reference corpus).

KWA is based on the principle that every text has a preferred set of expressions. For example, a romance novel will use words like *love* and *heart* more frequently than is typical in general language. As Scott and Tribble (2006) note, "a word-form which recurs within the text in question will be more likely to be key in it". By automatically identifying words that are unusually frequent in the target text compared to the reference corpus, KWA produces a list of keywords. This list then serves as a starting point for both quantitative and qualitative study, usually with the help of other corpus linguistics techniques, such as examining collocation

International Encyclopedia of Language and Linguistics, Third Edition, 2026, 72–78

profiles, semantic prosody, or keyword links, i.e. the co-occurrences of keywords within a specific textual span (Scott & Tribble, 2006), among others.

The term "keyword" in this context is specific to this statistical method. It should be distinguished from other uses of the term, such as search engine queries, or the "cultural keywords" described by scholars like Williams (1985) or Wierzbicka (2010).

Ultimately, the keywords generated by KWA can signal a text's topic (its "aboutness") and its specific style or register. The latter application of KWs was illustrated by Xiao and McEnery (2005), who proposed using KWA as a complementary tool to multidimensional analysis of register, a method developed by Biber (1988, 1995). In general, keywords allow researchers to systematically explore the unique linguistic profile of a text, forming a solid basis for deeper interpretation. The methodological decisions involved in this process, such as the choice of reference corpus and statistical measures, will be outlined below.

Primary Uses of KWA and Its Applications

At its core, Keyword Analysis provides a systematic way to interpret a text by highlighting its most distinctive features. It allows researchers to move beyond intuition and identify objective, data-driven evidence for their analysis. The method has significant applications in two major fields: literary studies and discourse analysis.

In literary studies, KWA can uncover the unique stylistic fingerprints of authors, texts, or even individual characters. For example, researchers have used it to analyze Shakespearean plays, identifying the specific words that distinguish the dialog of characters in *Romeo and Juliet* (Culpeper, 2002; Walker, 2010). It can also be used for comparative analysis, revealing the unique thematic vocabulary of one play when set against others in a collection (Scott & Tribble, 2006).

Keyword Analysis is also a cornerstone of Corpus-Assisted Discourse Studies (CADS), where it is used to investigate how social issues are represented in public discourse. These studies often analyze large collections of news articles or journals to uncover underlying patterns and biases. For instance, KWA has been used to explore media representations of crime in British and German newspapers (Tabbert, 2015), to examine the language used to describe refugees and asylum seekers (Baker & McEnery, 2005), and to track the evolution of discourse surrounding LGBT issues (Baker, 2005).

The Methodology of Keyword Analysis

One of the key advantages of KWA is that the identification of KWs is grounded in a clear quantitative basis. As Culpeper and Demmen (2015, p. 90) note, KWA "is less subject to the vagaries of subjective judgments of cultural importance ... [and] it does not rely on researchers selecting items that might be important... but can reveal items that researchers did not know to be important in the first place." While the underlying statistics can be complex, the standard procedure follows several key steps. In the most common scenario, a unit (such as a word, lemma, or another item) is considered prominent, a keyword, if it meets two specific conditions (Fidler & Cvrček, 2018; Scott, 2010, p. 43):

(a) relative frequency of a unit (i.e. raw frequency divided by the size of the text) differ significantly in the target text/corpus and reference corpus, and
(b) relative frequency of a unit in the target text/corpus is reasonably higher than its relative frequency in the reference corpus.

Step 1: Selecting the Target and Reference Corpora

The analysis begins by defining two sets of texts:

- The target text (or corpus): This is the single text or collection of texts to be analyzed (e.g., a novel, a series of speeches).
- The reference corpus: This is often a much larger collection of texts that represents a baseline of "normal" language use (e.g., a multi-million-word corpus of modern English).

The choice of reference corpus is a crucial methodological decision, as it provides the context for what will be considered "unusual" or "key" (cf. Fidler & Cvrček, 2015).

Step 2: Calculating and Comparing Frequencies

Software is used to calculate the relative frequency of every word in both the target text and the reference corpus. Using relative frequency (e.g., occurrences per million words) ensures a fair comparison between corpora of different sizes. A word becomes a candidate for keyness if its relative frequency is substantially higher in the target text.

Step 3: Applying Statistical Tests for Prominence

To qualify as a keyword, a word generally must meet two conditions:

1. Statistical Significance: A statistical test, such as the Log-likelihood (Dunning, 1993) or Chi-square test, is applied. This determines whether the observed frequency difference is statistically significant, or if it could have simply occurred by random chance. For a comparison of these tests and further references, see Bertels and Speelman (2013), simple calculator for all these test can be found at https://korpus.cz/calc/.
2. Effect Size: Because significance tests can be misleading with very large corpora – they are typically "asymptotically true" with large datasets (Kilgarriff, 2005; Wilson, 2013) –, they are paired with an effect size measure. This score measures the *strength* or *magnitude* of the frequency difference. This value is crucial because it allows all the keywords to be ranked, showing which are the most prominent. Most often used effect size estimators are:
   - the "simple math" coefficient (Kilgarriff, 2009),
   - ProcDiff or %DIFF (Gabrielatos & Marchi, 2012),
   - Bayes Factor – BIC (Wilson, 2013),
   - Difference Index – DIN (Fidler & Cvrček, 2015), or
   - risk ratio (Zemánek & Milička, 2017).

Step 4: Generating the Final Keyword List

The output of the analysis is a list of keywords, ranked from most to least prominent according to their effect size score. This ranked list provides the objective, data-driven starting point for the researcher's interpretation.

Interpreting the Keyword List

Receiving a ranked list of keywords is not the end of the analysis, but rather the beginning of the interpretative work. The list provides the "what," and the researcher's job is to uncover the "why." Several methods are used to interpret the function and meaning of keywords within the discourse.

Close Reading and Concordance Analysis

The most common approach begins with a manual review of the keyword list, starting with the most prominent terms at the top. To understand a keyword's function, researchers examine it in its original context. This is done using a concordance, which is a tool that displays every instance of the keyword along with the words immediately surrounding it.

By analyzing these concordance lines, a researcher can identify patterns of use, semantic nuances, and the keyword's overall contribution to the text's message.

Analyzing Collocations and Co-Occurrence

Going beyond single words, interpretation often involves analyzing a keyword's typical neighbors. As Stubbs (2005) notes, context is crucial, as "collocations create connotations." This includes:

- Collocation: Examining which words frequently appear directly next to a keyword can reveal common phrases and associations.
- Keyword Links: Researchers may also look at the co-occurrence of keywords within a wider textual span (e.g., within the same sentence or paragraph). This helps to group keywords into related thematic clusters. For example, in a political text, the keywords *taxes*, *economy*, and *growth* might frequently appear together, revealing a key topic area (cf. Market Basket Analysis below).

Advanced Clustering and Recontextualizing of Keywords

Output of KWA, a keyword list, consists of units which are essentially decontextualized and therefore hard to interpret. To address this, additional methods beyond collocation and concordance analysis are employed to cluster keywords into interpretable groups or clusters. Unlike collocations, which concentrate on the immediate or close context surrounding a KW, methods such as keyword links, Market Basket Analysis (MBA), or Multiple Correspondence Analysis (MCA) help identify longer-range relationships between KWs.

Keyword links group KWs based on their co-occurrence within larger context windows (e.g., up to fifteen words on either side), which helps in identifying related topics. However, KW links lack a sophisticated mechanism for evaluating the importance or strength of these connections. As a result, researchers face challenges in selecting specific KW links for closer analysis from among the numerous connections identified.

Market Basket Analysis (Cvrček & Fidler, 2022), a data mining technique originally used in marketing to uncover consistent associations between items in a shopping cart, can reveal associations between KWs in a corpus of many texts. By identifying recurring associations between KWs and evaluating their strength using three metrics (support, confidence, and lift) MBA can address the lack of broader context (re-contextualize KWs).

The use of **Multiple Correspondence Analysis** (MCA) for interpreting keywords (KWs) was introduced by Clarke et al. (2021). This method involves a multidimensional analysis of the presence of individual KWs across different texts, enabling the identification of dimensions similar to those found in Biber's register analysis (MDA). These dimensions can then be interpreted at a more general level, providing deeper insights into the underlying patterns and themes within the corpus.

Advanced Topics and Related Issues

While the standard Keyword Analysis provides a powerful framework, several extensions, alternative approaches, and related concepts offer deeper analytical potential.

Extending the Unit of Analysis

KWA is not limited to single words. The principle of identifying "key" units can be applied to other linguistic features:

- **Key Multi-word Expressions**: Instead of single words, researchers can look for unusually frequent multi-word chunks or clusters (e.g., *hands in his pockets*), which can reveal more about discourse phraseology (Fischer-Starcke, 2009; Mahlberg, 2007).
- **Keymorph Analysis** (KMA): This technique, a part of Multi-level Discourse Prominence Analysis (Fidler & Cvrček, 2018), extends the analysis to grammatical categories (cf. Culpeper & Demmen, 2015). It identifies key parts-of-speech (e.g., more nouns than expected), key semantic domains based on semantic tagging (cf. Baker, 2009) or key morphological features (e.g., a specific verb tense). This has proven especially insightful for highly inflected languages like Slavic languages, where grammatical choices carry significant weight. For instance, studies based on English suggest that grammatical information conveyed by parts of speech contributes little to traditional KWA (Culpeper, 2009, pp. 54–55). In contrast, research on Slavic languages shows that grammatical prominence can be crucial for discourse interpretation (Fidler & Cvrček, 2018, 2019; Janda et al., 2022). This can be seen, for example, in the differentiation between noun-heavy and verb-heavy texts or in highlighting specific cases of keywords, which can indicate their particular function within the target text (e.g. analysis of V. Putin's speeches by Janda et al. (2022) revealed how use of case supports the portrayal of Russia as a dynamic, agentive actor as opposed to Ukraine as a static territory and NATO as an untrustworthy organization).

Ways of Units Quantification

Since KWA relies on comparing frequencies, the method of counting units directly influences the analysis results. The most common approach in KWA is to use relative frequencies, where word counts are divided by the size of the text or corpus. However, relative frequencies cannot account for the dispersion of units in a corpus, which is crucial for evaluating how common a word is (cf. Gries, 2022). As a result, several alternative quantification methods have been proposed. For example, Egbert and Biber (2019) suggest using document counts for KW identification instead of the number of occurrences, while Gries (2024) advocates for the use of information-theoretic measure called Kullback-Leibler divergence (DKL).

These approaches, which rely on the natural structuring of a corpus into smaller chunks (typically individual texts), are particularly useful for analyzing larger datasets or broader samples of discourse. However, when the subject of KWA is a single, homogeneous text (such as a novel), the text must first be arbitrarily divided into smaller parts. This process introduces a subjective element into the analysis, potentially affecting the objectivity of the results.

Choice of Reference Corpus

KWA can also be viewed as a method that facilitates our understanding of how a text was or might have been read. The choice of reference corpus, against which the target text is compared, is crucial in this context.

Interpretation of a text is always influenced by the reader's expectations and experiences. Similarly, a reference corpus can assist researchers in exploring how readers might have received a text. For example, in their analysis of New Year's addresses by Gustáv Husák, the last communist president of Czechoslovakia, Fidler and Cvrček (2015) examined the differences in KWs obtained using different reference corpora representing two distinct periods. This approach helped approximate the interpretations of different readers – contemporary readers versus readers from the past. For a discussion on the influence of the reference corpus on KWA outcomes, see Scott (2020).

In this approach, KWs do not simply indicate "aboutness" or style/register; rather, they provide insights into what might have been striking or surprising to different readers. Since readers adopt different linguistic and stylistic norms, they perceive various units as violations of their expectations, making the reference corpus a powerful tool in understanding the reception of a text.

Key Keywords and Lockwords

Researchers have also developed concepts to describe different types of word prominence across texts:

- **Key Keywords**: When analyzing a corpus of many texts (e.g., an author's complete works), "key keywords" are words that are consistently key across most or all of the individual texts. They represent the central, recurring themes of the entire collection (cf. Baker et al., 2006, p. 97; Scott & Tribble, 2006, p. 77)
- **Lockwords**: In contrast, "lockwords" are words that maintain a surprisingly stable frequency across different corpora or time periods. These are often grammatical words, and their stability can be a point of analysis, especially in studies of language change (Baker, 2011, p. 66; Taylor, 2013, p. 91).

Alternative Methods for Identifying Prominent Words

Other methods exist for extracting important words from a text, most of which operate without a reference corpus:

- **TF-IDF** (Term Frequency-Inverse Document Frequency): Widely used in information retrieval, TF-IDF identifies words that are frequent within a single document but rare across a larger set of documents (Manning & Schutze, 1999, p. 543). This highlights terms that are uniquely characteristic of that one document.
- **Thematic Concentration**: This method from quantitative linguistics identifies prominent "thematic" words based on their position in a text's rank-frequency distribution (Popescu et al., 2007). The first step is to locate the "h-point" in the frequency list of word types that occur in the text. The h-point corresponds to a word whose frequency is equal to its rank (e.g., the 57th word in the frequency list of types in a text has a raw frequency of 57 occurrences). This h-point divides the distribution into two distinct populations: words with a frequency higher than the h-point (typically grammatical words) and all other words (typically lexical or content words). Thematic concentration words are those lexical or content words that appear in the "grammatical part" of the distribution, i.e., above the h-point.

While useful, these methods differ fundamentally from KWA because they do not use an external reference corpus. As a result, they cannot model the reader's linguistic expectations or norms in the same way, making their outputs potentially more challenging to interpret from a discourse perspective.

Software and Tools for Keyword Analysis

A number of specialized software tools and platforms are available to perform Keyword Analysis. While they share core functions, they vary in their specific features, the statistical options they provide, and whether they include built-in reference corpora.

The most widely used tools include:

- **WordSmith Tools** (Scott, 2020, see https://www.lexically.net/wordsmith/): A powerful and long-standing desktop application for comprehensive corpus analysis.
- **Sketch Engine** (Kilgarriff et al., 2014, see https://www.sketchengine.eu): A major web-based platform that offers a suite of analysis tools and comes with many large, pre-loaded reference corpora.
- **AntConc** (Anthony, 2019, see http://www.laurenceanthony.net/software/antconc): A popular, free, multi-platform desktop tool that is highly versatile and requires users to provide their own corpora.
- **WMatrix** (Rayson, 2009, see http://ucrel.lancs.ac.uk/wmatrix/): A web-based tool from Lancaster University that provides corpus analysis and semantic tagging functionality.
- **LancsBox** (Brezina et al., 2015, see see http://corpora.lancs.ac.uk/lancsbox/): A free, modern desktop application developed at Lancaster University for corpus linguistics research.
- **KWords** (Horký et al., 2023, see https://kwords.korpus.cz): A web-based tool from the Czech National Corpus specifically designed for performing keyword analysis.
- **KonText** (Machálek, 2020, see https://www.korpus.cz/kontext): An open-source corpus query interface, also part of the Czech National Corpus project, which includes keyword functionality.

The choice of tool often depends on the researcher's project requirements, budget, and whether they need to use pre-existing reference corpora or build their own.

Conclusion

Keyword Analysis (KWA) is a robust and widely-used method in corpus linguistics for a data-driven examination of texts. By statistically identifying the prominent units that characterize a text or discourse, it provides researchers with an objective foundation for interpretation, moving beyond subjective impressions of a text's themes or style.

The method's strength lies in its comparative nature—evaluating a target text against a reference corpus—which allows it to uncover what is truly unique. This has proven invaluable in fields ranging from literary studies to social science, enabling nuanced analysis of everything from authorial style to media framing.

While the process requires careful methodological choices, the growing availability of sophisticated software and analytical techniques continues to enhance its power and accessibility. Ultimately, Keyword Analysis offers a versatile framework for systematically revealing the prominent features of a text, providing crucial insights into how language is used to create meaning.

See Also: Keyword Analysis.

References

Anthony, L. (2019). *AntConc (Version 3.5.8) [Computer software]*. Waseda University. http://www.laurenceanthony.net/software.

Baker, P. (2005). *Public discourses of gay men*. Routledge.

Baker, P. (2009). The question is, how cruel is it? Keywords in debates on fox hunting in the British House of Commons. In D. Archer (Ed.), *What's in a word-list? Investigating word frequency and keyword extraction* (pp. 125–136). Ashgate Publishing, Ltd.

Baker, P. (2011). Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics, 39*(1), 65–88. https://doi.org/10.1177/0075424210368368

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press.

Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of Language and Politics, 4*(2), 197–226.

Bertels, A., & Speelman, D. (2013). 'Keywords Method' versus 'Calcul des Spécificités': A comparison of tools and methods. *International Journal of Corpus Linguistics, 18*(4), 536–560. https://doi.org/10.1075/ijcl.18.4.04ber

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

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

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

Clarke, I., McEnery, T., & Brookes, G. (2021). Multiple correspondence analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. *Register Studies, 3*(1), 144–171. John Benjamins.

Culpeper, J. (2002). Computers, language and characterisation: An analysis of six characters in Romeo and Juliet. In U. Melander-Marttala, C. Ostman, & M. Kyto (Eds.), *Conversation in life and in literature: Papers from the ASLA symposium* (Vol. 15, pp. 11–30). Association Suedoise de Linguistique Appliquee.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29–59. https://doi.org/10.1075/ijcl.14.1.03cul

Culpeper, J., & Demmen, J. (2015). Keywords. In D. Biber, & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 90–105). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377

Cvrček, V., & Fidler, M. (2022). No keyword is an island: In search of covert associations. *Corpora, 17*(2), 259–290. https://doi.org/10.3366/cor.2022.0256

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

Egbert, J., & Biber, D. E. (2019). Incorporating text dispersion into keyword analyses. *Corpora, 14*(1), 77–104.

Fidler, M., & Cvrček, V. (2015). A data-driven analysis of reader viewpoints: Reconstructing the historical reader using keyword analysis. *Journal of Slavic Linguistics, 23*(2), 197–239. https://doi.org/10.1353/jsl.2015.0018

Fidler, M., & Cvrček, V. (2018). Going beyond "aboutness": A quantitative analysis of sputnik Czech Republic. In M. Fidler, & V. Cvrček (Eds.), *Taming the corpus: From inflection and lexis to interpretation* (1st ed., Vol. 1, pp. 195–225). Springer.

Fidler, M., & Cvrček, V. (2019). Keymorph analysis, or how morphosyntax informs discourse. *Corpus Linguistics and Linguistic Theory, 15*(1), 39–70. https://doi.org/10.1515/cllt-2016-0073

Fischer-Starcke, B. (2009). Keywords and frequent phrases of Jane Austen's Pride and Prejudice. *International Journal of Corpus Linguistics, 14*(4), 492–523. https://doi.org/10.1075/ijcl.14.4.03fis

Gabrielatos, C., & Marchi, A. (2012). Keyness: Appropriate metrics and practical issues. In *CADS International 2012. Corpus-assisted Discourse studies: More than the sum of Discourse Analysis and computing?*. Italy: University of Bologna. https://repository.edgehill.ac.uk/4196/.

Gries, S. T. (2022). Toward more careful corpus statistics: Uncertainty estimates for frequencies, dispersions, association measures, and more. *Research Methods in Applied Linguistics, 1*(1), Article 100002. https://doi.org/10.1016/j.rmal.2021.100002

Gries, S. T. (2024). *Frequency, dispersion, association, and keyness. Revising and tupleizing corpus-linguistic measures* (1st ed.). John Benjamins. https://doi.org/10.1075/scl.115

Horký, V., Vondřička, P., & Cvrček, V. (2023). *KWords (Version 2) [Computer software]*. FF UK: Institute of the Czech National Corpus. http://kwords.korpus.cz/.

Janda, L. A., Fidler, M., Cvrček, V., & Obukhova, A. (2022). The case for case in Putin's speeches. *Russian Linguistics, 47*, 15–40. https://doi.org/10.1007/s11185-022-09269-2

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory, 1*(2), 263–276. https://doi.org/10.1515/cllt.2005.1.2.263

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics 2009 Conference*. Liverpool: Corpus Linguistics 2009. http://ucrel.lancs.ac.uk/publications/cl2009/171_FullPaper.doc.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., et al. (2014). The sketch engine: Ten years on. *Lexicography*, 7–36.

Machálek, T. (2020). KonText: Advanced and flexible corpus query interface. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 7003–7008). European Language Resources Association. https://aclanthology.org/2020.lrec-1.865.

Mahlberg, M. (2007). Clusters, key clusters and local textual functions in Dickens. *Corpora, 2*(1), 1–31.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

Popescu, I.-I., Best, K.-H., & Altmann, G. (2007). On the dynamics of word classes in text. *Glottometrics, 14*, 58–71.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment [Computer software]*. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/.

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails. In M. Bondi, & M. Scott (Eds.), *Keyness in texts* (pp. 43–58). John Benjamins. https://benjamins.com/catalog/scl.41.04sco.

Scott, M. (2020). *WordSmith tools (Version 8) [Computer software]*. Lexical Analysis Software. http://www.lexically.net.

Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins. https://doi.org/10.1075/scl.22

Stubbs, M. (2005). Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature, 14*(1), 5–24. https://doi.org/10.1177/0963947005048873

Tabbert, U. (2015). Crime and corpus. In *Lal.20*. John Benjamins. https://benjamins.com/catalog/lal.20.

Taylor, C. (2013). Searching for similarity using corpus-assisted discourse studies. *Corpora, 8*(1), 81–113. https://doi.org/10.3366/cor.2013.0035

Walker, B. (2010). Wmatrix, key-concepts and the narrators in Julian Barnes' talking it over. In D. McIntyre, & B. Busse (Eds.), *Language and style* (pp. 364–387). Palgrave MacMillan. http://www.palgrave.com/products/title.aspx?pid=335511.

Wierzbicka, A. (2010). *Experience, evidence, and sense: The hidden cultural legacy of English* (1 ed.). Oxford University Press.

Williams, R. (1985). *Keywords: A vocabulary of culture and Society (Revised, Subsequent edition).* Oxford University Press.

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. 3–11). Peter Lang.

Xiao, R., & McEnery, A. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*(1), 62–82. https://doi.org/10.1177/0075424204273957

Zemánek, P., & Milička, J. (2017). *Words lost and found: The diachronic dynamics of the Arabic lexicon.* RAM-Verlag.