Marina Bondi (dir.), Mike Scott (dir.), *Keyness in Texts*

Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2010

Geneviève Bordet

This text was automatically generated on 2 November 2020.

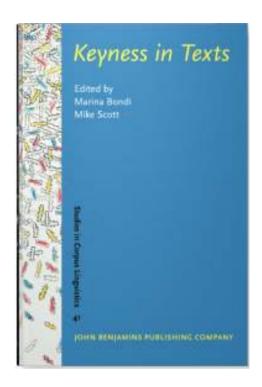# Marina Bondi (dir.), Mike Scott (dir.), *Keyness in Texts*

Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2010

Geneviève Bordet

REFERENCES

Bondi, Marina & Mike Scott (eds.). 2010. *Keyness in Texts.* Amsterdam/Philadelphia: John Benjamins Publishing Company. 251 pp. ISBN 978-90272-8766-3.

1   While this book, published in 2010, is no longer new, the issue it addresses remains a focus of major interest in the communities of researchers concerned with corpus linguistics, specialized discourse and text analysis. Keyness and keyness analysis have long been a field of interest because they offer a path towards text analysis based on corpus linguistics, as well as information retrieval based on both text analysis and corpus linguistics. Technological innovation with the creation of new software for keyword analysis is another factor which explains a surge of interest for this field of research. The book is edited by two major authors in the domain, with different perspectives. Marina Bondi is a Professor of English Studies and Translation at the University of Modena and Reggio Emilia (Italy) and her research approach combines attention to discourse analysis and corpus linguistics with a focus on language variation and phraseology. Mike Scott, from the University of Liverpool (UK) is the developer of the well-known software Wordsmith which its author presents as "an integrated suite of programs for looking at how words behave in texts." Therefore, the book offers an interesting combination of a wide perspective on corpus linguistics, text analysis and information retrieval, on the one hand, and a pragmatic and experimental approach of corpus and text, on the other hand.

2   The book is made up of thirteen individual contributions distributed into three sections. The first section explores the nature of keyness from various perspectives; the second studies the role of keyness in specialized discourse, while the third and final section offers critical and educational applications. I shall review each article before providing a general assessment of the book and of its contribution to the analysis of specialized languages.

Perspectives on keywords and keyness: an introduction, by Marina Bondi

3   With a quotation adapted from George Orwell's *Animal Farm*, "All words are equal but some are more equal than others," Marina Bondi offers diachronic and synchronic perspectives on the very nature of keyness, going from word to co-text, text and cultural context.

4   She first describes the evolution of the role of keyness in language studies. From a cultural perspective, J.R. Firth considers keywords as "pivotal" words whose distribution and use in context point to cultural values (1935: 40-41). Along the same line, R. Williams (1976) produces a dictionary of "cultural keywords" as a

representation of contemporary western culture. In an even more ambitious approach, A. Wierzbicka (2006) focuses on semantic universals and cultural underpinnings of a language. Her "Natural Semantic Metalanguage" aims to better identify the influence of "Anglo English" culture in language use. From a corpus linguistics perspective, keywords not only offer a key to culture, but also to the interpretation of texts. They are defined as words whose frequency (or infrequency) is statistically significant. Their study focuses on their typical co-occurrence with other lexico-semantic units. Keywords become key phrases and key clusters, thus opening perspectives for the study of collocations, phraseology and semantic preference. This study offers an insight on the exploration of the relation between text and form (Sinclair 2005) and the corpus study of semantic sequences (Hunston 2008).

5   Besides language studies, keywords also play a central role in knowledge management, to help identify the "aboutness" of texts in structured databases, notably with the development of textual databases. Following the creation of general and domain ontologies, the attention shifts from lexical units to underlying meaning units. Text mining therefore focuses on phraseological units and co-occurrences rather than on isolated words. Finally, a sample analysis of text and discourse gives evidence that keywords and phrases are indicative not only of a discourse community but also of the writer's identity and position.

6   A landmark text, *The General Theory of Employment, Interest and Money* by John Keynes, is compared with his previous book, *The Economic Consequences of the Peace*, a specialized reference corpus of current economic articles and a general corpus, the British National Corpus. A comparative analysis based on frequency shows that the study of keyword distribution provides significant information about the conceptual structure of a text while at the same time pointing to the evolution of economic discourse towards a denser form of writing, and a more formal type of reasoning.

Three concepts of keywords by Michael Stubbs

7   Michael Stubbs defines keywords as words which express evaluative social meaning and/or play a special role in a text or text-type. The central issue is to understand "how words relate to the world" (Searle 1969: 3), which implies identifying the connection between words, speech acts, speech events and social institutions. The aim of the paper is to assess the light corpus linguistics can shed on this connection and build bridges between Searle's speech act theory and corpus linguistics as two powerful models of communicative acts.

8   Keywords can be considered from various perspectives. From a cultural perspective, keywords not only label but also create cultural concepts (Williams 1976). From a statistical point of view, keywords are significantly more frequent words than expected in a sample of texts (Scott & Tribble 2006). They inform about the propositional content but not about text structure. Finally, in a corpus-driven perspective, keywords are significant units of meaning expressing shared values (Francis 1993).

9   The first approach identifies the connection between keywords and culture, the second between sets of keywords and text, the third one between phrased speech acts and cultural values. Corpus linguistics should be able to provide the missing link with a model of the relation between phraseology, speech acts, texts and social institutions by collecting material data on contextualized recurrent lexical patterns. In a

constructivist and Hallidayan vision of social reality as an edifice of meanings, and of culture as a network of meaning potential, discourse communities based on shared knowledge create and maintain the genres appropriate for the specific needs of social institutions. Searle explores the relation between speech acts and social institutions but cannot discuss text and text-types since he does not use empirical data. Corpus linguistics can help fill this gap by providing evidence of predictable co-occurrences of patterns which realize text and text-types; therefore contributing to tracing the interaction between phraseology, speech acts and social institutions. As a conclusion, one may say that speech act theory asks the right questions, but does not have the data; corpus linguistics has the data but cannot yet answer cognitive and social questions. Combining both approaches should help answer decisive questions such as the reason why extended phrasal units of meaning arise from recurrent collocations or, more generally, the interactions between social institutions and text-types seen as speech acts.

Problems in investigating keyness, or clearing the undergrowth and marking out trails by Mike Scott

10   Mike Scott explores the implications of the concept of keyness for corpus-based studies, positing that keyness is not language dependent, but text dependent. The topic has been widely explored since the 1990s, raising recurrent questions. One of these is the status of keywords when detected using machine methods. While the main objection was that machines could not detect humans "intentions," machines are also likely to detect regular patterns, including grammatical words that would not have been noticed by humans. Another recurrent issue is the size of the texts that should be investigated for keyness. While a software procedure such as Wordsmith's keywords tool can determine keywords at various levels, including chapters, whole texts, or a set of whole texts, what about taking into account multimodality, and more generally the mutation of texts on the web? The notion of text is "shifting sand." Besides, the software can identify keywords separately, using a reference corpus, but not the distribution of the whole set. Therefore, automatically identified keywords are just pointers for further prospection. The influence of the size of context and reference corpus should be taken into consideration when assessing the semantic value of keywords. Finally, the identification of related forms (synonymy and antonymy) is another issue that has yet to be addressed.

11   M. Scott then applies automatic identification of keywords to *Othello* and *Hamlet*, showing that it allows identification of keywords which would remain unnoticed, for example grammatical words. While this type of study provides interesting data as to the specific discourse of the main characters, these data only point to areas that should be explored for further literary interpretation.

Closed class keywords and corpus-driven analysis by Nicholas Groom

12   Prior to the development of keyword-identifying software such as Wordsmith or Antconc, the approach of keywords had to be deductive, based on a preselected list of terms to be studied. Keyword (henceforth KW) software, based on the comparison between the studied corpus and a larger reference corpus, allows to avoid this type of

bias and is able to uncover recurrent patterns or trends that would have remained undetected in a deductive approach.

13   One remaining major problem is the large number of results and the selection it implies for a detailed study; the usual "top-slicing" solution is not satisfactory since it is arbitrary and tends to eliminate interesting and meaningful types of KWs. The second usual answer is to eliminate "grammatical" words as not being directly related to the discourse analyst's concerns.

14   Groom advocates doing exactly the opposite and focusing exclusively on the "closed-class" KWs, based on a corpus-driven study of the academic discourse of history. A first step of analysis shows that in a wide majority of cases, *of* functions as a central linking preposition in complex noun phrases. The analysis goes beyond the already known dominance of complex nominal patterns in academic discourse when one does not focus on the repetition of isolated word forms, but on the repeated sequences of underlying meanings based on a conceptual analysis with repetitive patterns such as "PROPERTY + *of* + PHENOMENON." The recurrent semantic patterns yield interesting insight into the preferred meanings and values of the specialized discourse community, with a focus on process and properties combined with objects or phenomena. This is consistent with a discipline concerned with the impact of historical processes on social phenomena, and the re-conceptualization of given disciplinary problems in the light of new data, for example. The advantage of focusing only on closed-class KWs is that they allow the identification of repeated sequences of semantic elements beyond their lexical realization and beyond grammar patterns. Their study therefore uncovers discourse communities' "underlying commonalities of meaning" (p. 73) under apparently different lexical realizations.

Hyperlinks: keywords or key words? by Jukka Tyrkkö

15   With the explosive growth of digital media, electronic hypertexts and hyperlinks have become familiar to all. Texts are networks where coherence between two chunks of text connected by a hyperlink is based on the reader's ability to form expectations as to what the hyperlink will point to. Therefore the hyperlink is bound to occupy a salient semantic position. The paper focuses on a comparison between hyperlinks as keyness discursive devices, and statistical keywords. This leads to a differentiation between KWs as statistical facts and KWs as perceived by human readers. It is based on a case study of hypertextual fiction.

16   A hyperfiction story was selected; all the fragments were compiled as a unique corpus and a KW analysis was performed using Wordsmith 4 with the BNC as a reference corpus. All the hyperlinks were then extracted using Cmap Tools and compared to the KW lists of both sources – texts where they were found – and target fragments – texts they lead to. The cohesive role of hyperlinks was studied, based on the concept of "keyword linkage" i.e. a set of KWs shared by two texts or a set of texts, suggesting a shared discourse topic (Scott & Tribble 2006). Lexical cohesion was confirmed when a certain amount of repetition or collocation or co-reference was found between the hyperlink and the target fragment. Cohesion was found to be massively dominant, which tends to imply that hyperlinks function as key words, on a human perception basis. They are an example of Firth's (1935) focal words with an added formal markedness and redirective functionality.

Web semantics versus the semantic web: the problem of keyness by François Rastier

17   The semantic web has been presented since 1994 as an extension of the web which would allow access to the content of documents and would also allow reasoning in information research. This would be based on ontologies using denotational semantics. In this ontological approach, all contact with texts and documents is lost. The "web of data" replaces the "web of documents." An item of data is seen as a chain of characters, in a positivist logical context. This conception of data is based on economics and political aspects. Wide access to the web content and security concerns, specifically after 2001, require control of network making and of metadata. Databases must be interoperable to facilitate cooperative transparency. For situated cognitivists, "usage defines key properties for objects" (p. 98): an object can only be defined in relation to its various facets. One example of the limits of the ontological approach is Wordnet, of little use and limited interoperability. This approach does not allow any space for the concept of "keyness" which is not related to its referent in a semantic hierarchy. In fact, "the more superordinate a concept, the less subject to debate and the less key it is" (p. 99). Neither do ontologies take into account other semiotic clues such as typography which are decisive in digital documents. Therefore identifying "key information" should be considered as the objective of a praxeology, not an ontology.

18   Improvement of search engines requires applicable linguistics able to adapt to the task and the nature of documents implied. This type of linguistics should go beyond the combined use, in corpus linguistics, of computational linguistics and lexicometry. Full text search engines should take into account not only whole textual units ("passages") but also the contrast between discourse, genre and text sections. The text should not be considered as a chain of characters but a set of passages which can only be interpreted in context. Metadata must track the text and the context, with a philological and a hermeneutic aim. Keyness should be considered along three perspectives: subjective keyness based on the specific values of sections of the text such as peritext (title), intratext (body), infratext (notes, references). Subjective keyness is based on the reader's needs and interpretation while dynamic keyness, beyond subjectivity and objectivity, is constructed on document structure. Since there is great confusion as to what data and metadata are and as to their respective roles, a theory of textuality is needed to clarify this point and show that a text is not a reservoir of knowledge. Web Semantics based on corpus linguistics and semiotics should replace the Semantic Web.

Section 2 Keyness in specialized discourse

Identifying aboutgrams in engineering texts by Martin Warren

19   This research applies a computer-mediated research methodology, "concgramming," to the identification of the "aboutness" of texts. Concgrams consist of co-occurring words which can be either contiguous or combined in various positions. Two engineering research articles are studied to identify the most frequent two-word lexical combination. On this basis, a list of "aboutgrams" is selected and compared to a specialist corpus of general texts then to a general reference corpus, the BNC. The

comparison leads to the identification of the specific "aboutness" of each text. Moreover, for both texts, the lists of "aboutgrams" are compared to single word frequency lists in order to assess the impact of phraseology on the determination of the "aboutness" of the text. This methodology could be applied to determine genres or registers. It could also be used to raise students' awareness of the importance of phraseological variation.

Keywords and phrases in political speeches by Denise Milizia

20 The paper analyses n-grams and concgrams using WordSmith Tools, comparing speeches by Tony Blair and George Bush from early 2005 to late June 2007. The English corpus is comprised of one million words and the American one of five million words. A first list of top words is established and compared for each politician. Then the statistical prominence of keywords in Blair's speeches is determined, George Bush's speeches being used as a reference corpus. The first fifty keywords are obtained by referencing one million words uttered by Blair against five million words uttered by Bush. This list gives evidence of the importance of the term "climate" in Tony Blair's speeches. Its most frequent collocate is "change", as well as "issue" as in "the issue of climate change." The study of five-word concgrams focusing on this pattern shows that, while this pattern is dominant, there are also numerous variations such as "the climate change issue." The latter pattern, especially when "issue" is used in the plural form, is very often found in the neighborhood of the terms "Africa" and "trade." D. Milizia concludes to the relevance of this type of methodology to determine "aboutgrams" thus providing efficient indicators of discursive keyness.

Key words and key phrases in a corpus of travel writing by Andrea Gerbig

21 This paper offers a diachronic study of the evolution of keywords in travel writing from the 16th to the 21st century. Each of the six sub-corpora contains approximately 500,000 words. Two reference corpora were used for the extraction of keywords. A reference corpus of three million words was compiled for the three last centuries, based on spoken and written American, Australian and British sources. The three remaining sub-corpora, from the 16th century to the 18th, were compared with the Early Modern English section of the Helsinski Corpus. The 100 most frequent keywords for each century were then grouped into sets of semantically related words (countries, places, people, etc.). Results highlight the development in means of traveling with more and more distant destinations. They also show that the representation of travels and their objectives has changed, shifting from a focus on directions and positions in the three earlier centuries to a description of the destinations. The study's focus then shifts to "key key words", i.e. words that are frequent in a large number of texts in each sub-corpus, and "associates" i.e. words that co-occur significantly with a key-key word. It gives evidence of an evolution towards more nominalization, the earlier centuries using more verbs and prepositions, which confirms the interest for movement rather than places. Finally, the most frequent five-word-phrase frames over the centuries are described, confirming the earlier centuries' concern for movement and directions while more recent periods focus on places.

History versus marketing: keywords as a clue to disciplinary epistemology by Donatella Malavasi and Davide Mazi

22 Two comparable corpora (comprising research papers published in 1999 and 2000) of 2.5 million words were compiled for this study of keywords across two disciplines, history and marketing, with a focus on disciplinary epistemology. Based on the extraction of a keyword list for each discipline, using Wordsmith keyword function, five elements were selected regarding their relevance as to subjects, objects and procedures characterizing each discipline. Each selected keyword was concordanced in order to determine its collocational and intercollocational patterns and its co-occurrence with specific types of reporting verbs.

23 Results highlight the centrality of keywords in the epistemological characterization of each discipline. While history focuses on "historians" as sources, marketing uses "we" for the authors' self-representation. "Science" and "society" are the main objects of history while marketing concentrates on "effects." "Texts" are central for the procedures used by historians while marketing science is based on "data," "research" and "results." Study of the co-occurrences of keywords with specific types of reporting words in history highlights discourse acts ("claim", "tell", "declare") and cognition verbs ("reveal", "show"). In marketing, "data", "research" and "results" collocate with tentative verbs ("suggest", "may") and argumentative verbs ("support"). These contrasted data highlight the narrative trend of history in contrast with marketing which focuses on data and experimentation. Keywords appear as effective interpretive clues to specific disciplinary peculiarities.

Metaphorical keyness in specialized corpora, by Gill Philip

24 This research addresses the issue of the inter-relation between statistically-generated keywords and textual message by observing the interaction between keywords and evaluative language based on metaphors. A method for the semi-automatic identification of metaphors is described.

25 Following a discussion of the terminology of metaphors, in text and in corpora, the research sets out to identify potential metaphorical themes in a corpus comprised of speeches, communiqués and press releases by the Italian minister, Emma Bonino, from 2006 to 2007. Using WordSmith 4, a key-word list was generated with a wider political reference corpus. The words were lemmatized and grouped into semantic sets. To establish the criterion of incongruity, required for metaphor identification, it was hypothesized that low-frequency content words would provide a source of metaphorically-used lexis. Therefore the words were selected starting from the bottom of the list, stopping before the top 500 types. The selected words were then sorted in semantic sets, so as to be classified as part of potential metaphorical domains, and finally concordanced. Results show how metaphors become pervasive as they interact in significant ways with keywords. For instance, concordances of war-related lexis frequently co-occur with countries such as India or China. This creates a sort of covert keyness where keywords are neutral but key metaphors are not. The conclusion is that the evaluative role of metaphors can be understood through the examination of the collocational profile of each source domain lexical item.

Section 3 Educational and didactic perspectives

A contrastive analysis of keywords in newspaper articles on the "Kyoto Protocol" by Erica Bassi

26   This contrastive analysis of computed keywords in different languages compares the coverage of the Kyoto Protocol in two daily newspapers: the Italian *La Repubblica* and the American *New York Times* between 1997 and 2006. All articles in this period mentioning the Kyoto Protocol were selected and compared with the *New York Times* (NYT) portion of the American National Corpus and the *La Repubblica* corpus as reference corpora. Positive (statistically salient) and negative (statistically unusual) keywords were selected with WordSmith Tools 4. Keywords were then grouped manually into semantic fields. Negative keywords belong to the semantic fields of family and entertainment in both newspapers. As to positive keywords, three semantic fields are dominant in both cases: negotiation, change, and effect on the planet. The overall study gives contrastive evidence of the way each newspaper tackles the issues of negotiation and environment. In the United States, negotiation is felt as problematic while *La Repubblica* reveals a positive tendency towards agreement. *La Repubblica* highlights the negative consequences of climate change whereas the NYT stresses the lack of measures and decisions to tackle this issue.

Keywords in Korean national consciousness by Soon Hee Fraysse-Kim

27   This study focuses on the expression of national consciousness, termed as "weness" or "wuli" in Korean, through specific keywords comparing school-books used in four Korean communities: North Korea, South Korea, the Korean communities of, respectively, China and Japan. The aim is to better understand how Korean national identity is internalized and reproduced by school education. An electronic corpus of 3,000 words was established, composed of four sets of textbooks used in primary schools in 2006. Considering that the segmentation of words differs between the four communities, the statistical comparisons were performed on the basis of relative values with the amount of character occurrences as the standard. First, a list of the most frequent nouns and pronouns was established showing that "I" was dominant with the exception of North Korean where "we" prevails, followed by the noun "great-marshall." The collocational use of the concept of "weness" "or wuli" ("we", "our") was contrasted for the four communities. This comparison led to the conclusion that "weness" is mainly related to the concepts of "language" and "home." Then, the research goes beyond this sense of belonging by focusing on the notion of "otherness" or even "enemy" as it appears in textbooks. Investigations on proper nouns of people and places reveal that there is little agreement except for the name of a country: Japan (*Ilbon*). A classification of the occurrences of *Ilbon* was established based on the historical period and the global connotation of the narrative: antipathy, neutrality or sympathy. The conclusion is that antipathy is dominant in the colonial and the 16th century invasion periods while neutrality prevails in the present period. There are no occurrences for all four communities of sympathetic contexts of use of *Ilbon*, even for Korean residents in Japan. On the whole, the study shows that, despite variation, there are strong common points for all four communities, such as a sense of belonging based

on language, and a sense of "otherness" or even "common enemy" symbolized by Japan. These concepts appear as the major factor for the maintenance of "weness" and its conveyance by textbooks and school education.

General spoken language and school language: keywords and discourse patterns in history textbooks by Paola Leone

28   The concept of keyness is used here to compare the lexicon and lexical patterns young learners are confronted with in everyday life and in history school textbooks. The comparison is based on a self-compiled corpus of history textbooks used in secondary schools in Italy and a published corpus of Italian spoken language. The objective is to determine specific difficulties learners, especially young immigrants, might encounter when reading history textbooks. The study aims to determine the keywords that are specific to history textbooks, and the differences between language usage in textbooks and in spoken language. WordSmith Tools 5 was used to first establish a frequency-ordered list. The focus is set on the top 100 positive keywords in textbooks, as a result of a comparison of the two corpora. The study of these keywords shows that history textbooks tend to prefer tenses and verbs centered on time and transformation. They also show a preference for adverbs and linking adjuncts which construe causality and interpretation. While the vocabulary, centered on political authorities and entities, places and conflicts, is not very specific, it appears as acquiring new meanings in this specific context. Further investigation of this polysemy would require collocational analysis based on concordancing. This would allow an in-depth analysis of keywords and ultimately an improvement of teaching material, training and tests for minority students to help them cope with mainstream texts.

29   The perspectives offered by this book are wide and varied, sometimes complex. The concept of keyword explored in the book could be the missing link between the evidence produced by corpus linguistics analysis of corpora and the interactions between text-types and social institutions, between word, text and context. This contextualized approach of keyness in text and discourse paves the way for the development of new tools for information extraction and text analysis with full text search engines taking into account the contrast between discourse, genre and text. Corpus linguistics could thus be the cornerstone of the development of "Web Semantics."

30   Keyword analysis encompasses key phrases, key clusters and key linkage, thus opening perspectives on a better understanding of the textualizing role of phraseology. Based on specific software for keyword analysis, WordSmith Tools, the studies offer a wide range of approaches and methodologies for the exploration of the structuring role of keyness in specialized discourse. The concept appears as relevant for a contrastive approach of disciplinary epistemologies, as well as the ideology of textbooks, the diachronic evolution of travel writing or the analysis of political discourse.

31   However, the concept of keyness – while it is obviously operational – remains elusive and multi-faceted. Whether it signals the "aboutness" of the text or points at the textual structure, whether the key is conceptual or organizational, the definition of the very nature of keyness obviously requires further research. Understanding "how words relate to the world" (Searle 1969) is an ambitious objective. Beyond frequency analysis,

we need more information on its text structuring role and its interaction with textual patterning and phraseology. As Mike Scott puts it (p. 44):

> the aim is to mark out some trails and chop away some tangled undergrowth. [...] The only certainty is that our exploration will bring us some scratches and bruises. We may get to see some interesting views, however, when we get near the top of our path.

32 This is indeed a stimulating perspective and this book offers abundance of sometimes difficult but promising paths.

BIBLIOGRAPHY

FIRTH, John R. 1935. "Technique of semantics". *Transactions of the Philological Society*, 36–72.

FRANCIS, Gill. 1993. "A corpus-driven approach to grammar: Principles, methods and examples". In BAKER, M., G. FRANCIS & E. TOGNINI-BONELLI (eds.), *Text and Technology.* Amsterdam: John Benjamins, 137–156.

HUNSTON, Susan. 2008. "Starting with the small words: Patterns, lexis and semantic sequences". *International Journal of Corpus Linguistics* 13/3, 271–295.

SCOTT, Mike & Christopher TRIBBLE. 2006. *Textual Patterns.* Amsterdam: John Benjamins.

SEARLE, John R. 1969. *Speech Acts.* Oxford: Oxford University Press.

SINCLAIR, John. 2005. "What's in a phrase." Lecture given at University of Modena and Reggio Emilia, 15 November 2005.

WIERZBICKA, Anna. 2006. *English: Meaning and Culture*. Oxford: Oxford University Press.

WILLIAMS, Raymond. 1976. *Keywords: A vocabulary of culture and society*. London: Fontana Press.

AUTHORS

GENEVIÈVE BORDET

Université Paris Diderot Sorbonne Paris Cité, gbordet@eila.univ-paris-diderot.fr