

SKETCH ENGINE: A TOOLBOX FOR LINGUISTIC DISCOVERY¹

THOMAS, James: *DISCOVERING ENGLISH WITH SKETCH ENGINE: A CORPUS-BASED APPROACH TO LANGUAGE EXPLORATION*. 2nd ed. Brno: Versatile, 2016. 228 p. ISBN: 9788026083603

THOMAS, James: *DISCOVERING ENGLISH WITH SKETCH ENGINE: WORKBOOK AND GLOSSARY*. Brno: Versatile, 2016. 140 p. ISBN: 9788026095798

A new paradigm in linguistics, new language analysis tools along with easy access to a substantial amount of electronic texts have made researchers, lexicographers, terminologists, translators and language instructors increasingly interested in corpora. This has led to a higher demand for practical corpus knowledge and skills: which query tools and corpora are available, what they can do and how to use them. However, learning to use query tools requires a greater awareness of the linguistic theory within the neo-Firthian paradigm, which focuses on the interconnection of lexis and grammar. Operational knowledge is inseparable from the theoretical implications that guide research questions and help in the interpretation of corpus data. Indeed, using a corpus tool means understanding and accepting a certain amount of linguistic assumptions incorporated into it by the designers (Anthony, 2013).

In language teaching, the gap between scientific advances and the methodologies in use tends to be wide. There is a clear lack of instructional materials and theory-backed know-how in the field. In this context, the new comprehensive textbook on *Sketch Engine (SkE)* by James Thomas is especially timely. This book constitutes a practical hands-on introduction to corpus linguistics built around SkE and focused on English language teaching (ELT) and learning.

SkE is a set of software tools for corpus analysis developed by Lexical Computing Ltd. The system was created by a British lexicographer and corpus linguist, Adam Kilgarriff, and a Czech programmer, Pavel Rychlý. In 2004 this corpus query system was made available as a commercial product. Initially supplied with corpora in just three languages, Czech, Irish and English (Kilgarriff et al., 2004), the system was immediately appreciated by major dictionary projects. Today lexicographers from Cambridge University Press, Macmillan, Harper Collins, Oxford University Press use *SkE* as one of their corpus analysis tools of choice (Kilgarriff et al., 2014a, p. 15).

SkE is a leading online corpus analysis service with a range of highly flexible functions to build and analyze KWIC concordances for items ranging from lemmas to CQL query strings. It also offers common statistical methods to produce frequency statistics, calculate co-occurrence patterns, visualize contrasts and to explore user and multilingual corpora (Kilgarriff et al., 2014a). At the time of writing (May 2017), *SkE* contains 400 ready-to-use corpora in over 90 languages.²

The book is aimed at teachers and students of English, linguists and translators. It introduces corpus resources available through the *SkE* interface, and teaches *SkE* functionality.

¹ This work has been supported by the Russian Foundation for Basic Research within Project No. 17-06-00107.

² <https://www.sketchengine.co.uk>

The core system functions consist of the following tools (Herman, Kovář, 2013; Kilgarriff et al., 2014a; Kilgarriff et al., 2014b):

1. *Concordance* searches a corpus for a word form, a lemma, a phrase, a part of speech tag, etc. The system converts all queries into Corpus Query Language (CQL) which can be used directly.
2. *Word List* generates frequency lists of words, lemmas, n-grams or key words.
3. *Keywords and Terms* enables extraction of core lexis in a corpus using “keyness score”.
4. *Collocations* calculates words that are statistically associated with the query term. The system uses several measures to find collocation candidates: T-score, MI, log likelihood, logDice, etc.
5. *Word Sketch* generates summaries of a word’s grammatical and collocational behaviour using “sketch grammar”.
6. *Word Sketch Difference* offers a comparison of two words based on collocations.
7. *Thesaurus* creates a distributional thesaurus based on common collocation. The resulting list of words includes items in various semantic relationships.
8. *Trends* helps to conduct a diachronic analysis of word usage.
9. *WebBootCaT* is a set of programs to compile a user web corpus.

It would be a gross understatement to say that *DESKE* is limited to demonstrating the potential of *SkE*. It is actually an excellent guide to linguistic exploration. The author does not so much walk the reader through the tools as induces the reader to resort to them in search for solutions to the linguistic riddles he poses. He engages the reader in linguistic games that integrate accessible theory with guided practical activities leading towards a discovery. The author offers 354 contextualized questions that are in fact organized creative activities of progressing complexity and levels of independence. For example, in the beginning of the book, Thomas sets up an investigation into the origins and nature of the phrase ‘one swallow does not a summer make’ to demonstrate the difference between patterned and occasional language (Thomas, 2016a, p. 26). He instructs the reader on which corpus to use, how to build the concordance and how to read it. Without spelling out the answers, the task is full of leading questions that guide the reader in the process. This is also a good example of how intriguingly interconnected the book is: semantically the idiom relates to the basic principle of pattern hunting. Another entertainingly artistic example of the way the author chooses to set tasks is textbook question 215: “All question tags are contracted, are they not?” (p. 141). As the book progresses, the tasks get more sophisticated. Although they rely on prior knowledge and require more independence, the author does not fail to offer friendly tips, reminders and back-references where necessary. It is important that he motivates the reader to consider alternative solutions and notice seemingly insignificant by-products of the searches. This develops linguistic awareness and being able to find “something valuable unintentionally” (p. 63). The *DESKE* how-to instructions align conceptually with the ideas of data-driven learning (Johns, 1991; Boulton, 2010), a methodological approach in ELT which promotes learners’ autonomy, individual pace and inductive, self-directed language learning. Above all, it equips the learner with language learning skills that can be used independently.

To support his arguments the author refers to a broad variety of recognized authority from domains ranging from classical pedagogy (J. Piaget) to modern learner corpus research

(S. Hunston), from general linguistic theory (R. Jakobson) to up-to-date corpus linguistics (T. McEnery, P. Hanks). The extensive and versatile reference list makes this textbook theoretically well-grounded without turning it into a dry, overly scientific account. There is a selection of simple examples to illustrate concepts that might be challenging for a newcomer to the field. Without making any assumptions about students' prior linguistic knowledge, the author introduces concepts relating to many branches of linguistics, including corpus linguistics (e.g., *grammatization, word family, binominal, lexicogrammar, Idiom Principle, troponym, FASI, colligation, template, connotation, lemma, node, part-of-speech tagging, metacognitive strategies, association measures* like *(log)Dice* and *MI*, etc.). The terminology used in the book is explained in a glossary included with the workbook.

The linguistic aspect of *DESKE* is strengthened by links to external linguistic resources, e.g., P. Hank's *Pattern Dictionary of English Verbs*,³ intended to enhance readers' language competence, to stimulate interest in research and to develop knowledge about language.

James Thomas points out that corpus approach is underestimated in language education. According to the British Council Annual Report (2010), a large proportion of the 15 million English teachers in the world do not know anything about corpora and do not realize how they can be used in learning and teaching language. However, modern ELT classrooms are usually well-equipped to incorporate corpus methods (pp. 17 – 19).

The author convincingly demonstrates the importance of corpus research skills for foreign language learners. Analyzing authentic data helps to reveal numerous cases of specific language use which are unlikely to be discussed in grammar books, dictionaries or traditional textbooks, but could make a learner's speech vivid and idiomatic. The meaningful use of corpora develops linguistic intuition and promotes understanding of language variation; for example, the author demonstrates that the sentence "*I was sat watching TV*" presents a grammar structure which is quite widespread in the UK (p. 109).

This book demystifies the corpus approach to language teaching and shows that corpus software is no magic wand, but a useful tool to obtain language information that was hitherto almost inaccessible. The availability of corpus data opens up opportunities for DIY linguistic discoveries for everyone. Obviously, learning to use corpus software requires some effort, but they are comparable to finding one's way with electronic devices, such as smart phones.

The workbook supplement to this edition contains the 354 questions from the book and allows some space for answers. Each chapter is followed by a set of quiz questions on language or linguistic issues touched upon in the chapter and a set of research and discussion questions, that also come in the provided in a photocopy-friendly format, which can be used to facilitate discussion in class. The questions help to establish important connections between linguistics, corpus linguistics and language education objectives, along with issues related to translation, terminology use and lexicography. The glossary includes reader-friendly term definitions devoid of academic formalism. The textbook carries extensive links to screen shots and other support illustrations stored on *SkE* servers and a companion-site, although many of them are unavailable for users without a commercial license to *SkE*. The latter is also a bar to completing some of the tasks based on commercially available corpora.

³ <http://pdev.org.uk>

Although the questions included in *DESKE* are meant to be answered by querying English corpora (some unavailable), the ideas and methods can be adapted to other corpus managers and resources. One of the obvious and productive ways to practice these skills outside *SkE* and English is the family of comparable web corpora called *Aranea* (Benko, 2014). They are made available through an open-source project, *NoSketch Engine* (Rychlý, 2007), on the website of Comenius University in Bratislava, Slovakia.⁴ As of now, the family includes corpora in 17 languages. The text material for the corpora is processed and tagged in a uniform manner, which ensures comparability of language data. Importantly, this data can be accessed through a single interface, without having to adjust to the differences between the unique interfaces that often come with one-off corpora. The *Aranea* corpora come in three sizes: the freely-available *Minus* format (120 mln tokens), the *Maius* corpora (over 1 bln tokens accessible after a free registration) and the *Maximum Aranea* for just a few languages (Czech, Slovak and Russian). The *NoSketch Engine* platform inherits most of the *SkE* functionality, which makes chapters 1–10 and chapter 13 of the book by J. Thomas fully applicable to this resource.

Generally, *DESKE* is a witty, highly-readable narrative written by an enthusiast with vast experience in ELT, whose line of argumentation is very compelling and engaging. It is rich in metaphors, imagery, real-world and literary references which deliver ideas in a more immediate, easy-going and memorable way than traditional academic prose. It elegantly combines theory and practice, which one always hopes for in a textbook. *DESKE* will make a great coursebook in general practice-oriented corpus linguistics as part of the education of language teachers or translators. We are sure that many language learners will benefit if they are introduced to corpus skills and it undoubtedly makes sense to incorporate a *DESKE*-based course into any advanced language training program.

References

ANTHONY, Laurence: A critical look at software tools in corpus linguistics, *Linguistic Research*, 30(2), pp.141–161.

BENKO, Vladimír: *Aranea: Yet Another Family of (Comparable) WebCorpora*. In: Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014. pp. 257–264.

BOULTON, Alex: Data-driven learning: Taking the computer out of the equation. In: *Language Learning*, 2010, Vol. 60 Issue 3, pp. 534–572.

HERMAN, Ondřej – KOVÁŘ, Vojtěch: Methods for Detection of Word Usage over Time. In: VII Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2013. Brno: Tribun EU 2013. pp. 79–85.

JOHNS, Tim: Should you be persuaded: Two samples of data-driven learning materials. In: *English Language Research Journal*, 1991, Vol. 4, pp. 1–16.

KILGARRIFF, Adam – RYCHLÝ, Pavel – SMRŽ, Pavel – TUGWELL, David: *The Sketch Engine*. In: Proceedings of the XI EURALEX International Congress. Lorient: Université de Bretagne-Sud, 2004, pp. 105–116. Accessed on: https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf

⁴ http://unesco.uniba.sk/aranea_about

KILGARRIFF, Adam – BAISA, Vít – BUŠTA, Jan – JAKUBÍČEK, Miloš – KOVÁŘ, Vojtěch – MICHELFEIT, Jan – RYCHLÝ, Pavel – SUCHOMEL, Vít: The Sketch Engine: Ten Years On. In: Lexicography ASIALEX, 2014a, Vol. 1, pp. 7–36. Accessed on: <http://link.springer.com/article/10.1007/s40607-014-0009-9>

KILGARRIFF, Adam – JAKUBÍČEK, Miloš – KOVÁŘ, Vojtěch – RYCHLÝ, Pavel – SUCHOMEL, Vít: Finding Terms in Corpora for Many Languages with the Sketch Engine. In: Proceedings of the Demonstrations at the 14th Conference the European Chapter of the Association for Computational Linguistics. Sweden, April 2014b, pp. 53–56. Accessed on: https://www.sketchengine.co.uk/wp-content/uploads/Finding_Terms_2014.pdf

RYCHLÝ, Pavel: Manatee/Bonito – A Modular Corpus Manager. In: 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. pp. 65-70.

THOMAS, James: Discovering English with Sketch Engine: A Corpus-Based Approach to Language Exploration. 2nd ed. Brno: Versatile, 2016a. 228 pp.

THOMAS, James: Discovering English with Sketch Engine: Workbook and Glossary. Brno: Versatile, 2016b. 140 pp.

Maria Kunilovskaya – Marina Koviazina
Department of Philology and Journalism
University of Tyumen, Russia