Discovering English with the Sketch Engine

James Thomas[1]

Abstract. *Discovering English with the Sketch Engine* is the title of a new book (Thomas, 2014) which introduces the use of corpora in language study, teaching, writing and translating. It focuses on using the Sketch Engine to identify patterns of normal usage in many aspects of English ranging from morphology to discourse and pragmatics. This paper discusses the findings from the most recent semester when corpora were used by three different groups of university students studying English for different ends. It was found that learners who are capable of drawing conclusions from data are quick to recognize the potential of corpora. Through training in the relevant aspects of linguistics and the software, all 60 students were able to create and exploit corpora. Although the book itself was still under development, its approach was piloted on these students using extended excerpts and was demonstrated to be of value in fulfilling quite different needs.

Keywords: Sketch Engine, guided discovery, corpus creation, corpus exploitation, learning language from language, patterns of normal usage.

1. Introduction

Discovering English with the Sketch Engine is the name of a new book (Thomas, 2014) which aims to inculcate descriptive and neo-Firthian views of English through teaching the Sketch Engine, a multi-faceted, web-based corpus tool that generates concordances, word sketches and various lists, has many pre-loaded corpora and facilitates corpus building (Kilgarriff, Rychlý, Smrz, & Tugwell, 2004). The book is aimed at teachers and trainees, advanced students and translators.

[1] thomas@phil.muni.cz.

How to cite this article: Thomas, J. (2014). Discovering English with the Sketch Engine. In S. Jager, L. Bradley, E. J. Meima, & S. Thouësny (Eds), *CALL Design: Principles and Practice*; *Proceedings of the 2014 EUROCALL Conference, Groningen, The Netherlands* (pp. 363-367). Dublin: Research-publishing.net. doi:10.14705/rpnet.2014.000246

The *discovering* of the title is manifested in the hundreds of questions whose answers the reader discover through learning how to form queries, how to use the Sketch Engine tools to process the data, and then to interpret it. I refer to this as *task-based linguistics*.

It is no exaggeration to say in 2014 that corpus linguistics has fundamentally changed our view of how language works. It is enough to open a mono-lingual learner dictionary (MLD), such as the ground-breaking COBUILD (1987) and a corpus-informed grammar such as Longman Grammar of Spoken and Written English (Biber, Johansson, Leech, Conrad, & Finegan, 1999) to see that the dictionary contains a vast amount of grammatical information and the grammar contains just as much lexical information. The interdependence of these two strands of language is empirically demonstrated on almost every page of these books and others of their ilk, through the use of corpora. Publishers such as Macmillan, Oxford University Press, Cambridge University Press, and HarperCollins all use the Sketch Engine in the preparation of dictionaries which makes it a resource worth putting in the hands of foreign language users.

Neo-Firthian linguists, such as Sinclair, Hoey, Hunston and Hanks were among many involved in the COBUILD project and at the core of their work, then and now, lies collocation and colligation. These underpin the work in multi-word units, grammar patterns, word templates and discourse studies whose linguistic significance is measured in terms of frequency: patterns of normal usage are invaluable to language pedagogy.

Discovering English with the Sketch Engine is particularly keen that language learners et al. understand the importance of patterns of normal usage so that they acquire them and a sense for them. This further empowers them to recognize language users' use of metaphors and other kinds of creative language, which Hanks (2013) refers to as *exploitations*. It is important that both norms and exploitations are recognized in all language input and output.

2. Method

2.1. Differing groups of students

Task-based linguistics has been evolving over a number of years and involves work with various groups of students. In two-year masters program training secondary school English teachers, we work with both in-service and pre-service teachers whose common European framework level is C2. In our bachelor program, whose students are hovering around C1, a new course called Collocation Plus was opened to experiment with corpus-based approaches to introducing neo-Firthian linguistics practically. At the Faculty of Informatics, doctoral students of academic writing, whose English ranges from A2 to C1, have long been trained in using the Sketch Engine to discover and use patterns of normal usage. None of this semester's 'informaticians' had ever heard of corpora, which made their wholesale acceptance of its value quite significant. Interestingly, they had not heard of the Sketch Engine, which was born at their faculty, and its international user base is managed from there as is its on-going development.

2.2. Student use of the Sketch Engine

The Sketch Engine is introduced via the BNC so that the concept of "core English" is well-established: it then serves as a standard against which their specialized corpus work is undertaken.

In the spring semester of 2014, all groups were taught to use the Sketch Engine tools that are relevant to their areas of study, and were required to produce work relevant to their future studies and careers. The students of *Collocation Plus* used the Sketch Engine's *Corpus Builder* to make a corpus of one novel that they downloaded from the Gutenberg Project. They explored the corpus looking for words and phrases they considered interesting. This assignment was partly inspired by an interview with David Crystal (2014) at IATEFL (online) in which he describes creating a glossary for a single book, something which he considers surprisingly innovative. The students published their glossary on a dedicated website and in addition, described the process in a five-minute video using JING.

The teacher trainees worked on graded readers. In pairs, they chose one graded reader, scanned it, uploaded it to our Graded Reader Corpus (GRC) and explored it to find examples of language features that we studied in the course, *Linguistic for Language Teachers* that ran in parallel to their methodology course, *Syllabus, Lesson and Material Design for ELT*. They produced draft teaching material that they presented using Moodle tools, some of which linked to the GRC using the Sketch Engine's permalinks, while other material they produced was related to the literary and cultural aspects of their chosen reader. They will use these resources next semester in their internal practice teaching.

The academic writing students are required to add ten personally relevant texts to the Informatics Reading Corpus (IRC) which I launched ten years ago and now contains almost eight million words. When uploading, the students add XML tags to generate metadata that includes the sections of papers and the field of Informatics in which they work. This permits targeted searches, e.g. the four word bundles that start sentences in the Future Work sections of papers in the field of natural language processing. The main assignment for their course is an academic paper they are currently working on.

3. Conclusions

From a self-assessment questionnaire of 90 linguistic, language acquisition and teaching constructs given to the masters students (25 pre-service and 11 in-service teacher trainees), it emerges that they acknowledge various levels of improvement in almost all areas, including the use of corpora. More objectively, the assignments of the bachelor and masters students evidence their grasp of many new linguistic concepts. It could also be seen in computer-equipped classrooms throughout the semester that the students gained a wide range of corpus skills as they interpreted data and added it to the course wiki under such headings as phrasal verb particles, bound and free prepositions and the word templates of specific nouns and verbs. As this was the first time that these assignments were set, future work will include a stronger focus on the relationships between discrete linguistic items and the discourse in which they appear. This manifests a top-down, bottom-up dichotomy that was introduced at the beginning of the courses, but needs reinforcing at later stages.

Accounting for the progress made by the informaticians is complicated by their various levels of English, as this dictates different uses of corpora. The students were able to derive information about collocation and colligation in general English from the BNC, whereas from the IRC, they were able to study terms and phrases related to their fields. They observed the use of parentheses, the use of first person, sexist language, writing numbers, the ubiquitous *the Noun of Noun* pattern. Some were able to observe hedging and tease out word templates.

As *Discovering English with the Sketch Engine* was under development during the semester described, the students in all courses piloted sections of it as it evolved, and were given chapters that pertained to the assignments as the semester drew to a close. Through learning new linguistic concepts, many of which are manifestations of corpus studies, and exploring them in corpora, students develop a new view of language itself. Such discoveries make concordancing a subversive activity (Thomas, 2009) as it displaces such notions as sentence grammar as the primary organizing unit for language study and teaching.

References

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

COBUILD. (1987). *Collins COBUILD English Dictionary*. Michigan: Collins.

Crystal, D. (2014). I*nterview with David Crystal - David talks about two projects he has recently been working on...* [video]. Harrogate Online. Retrieved from http://iatefl.britishcouncil.org/2014/sessions/2014-04-02/interview-david-crystal-david-talks-about-two-projects-he-has-recently-been-work

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Cambridge MA: MIT Press. doi:10.7551/mitpress/9780262018579.001.0001

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds), *Proceedings of the 11th Euralex Congress* (pp. 105-116). Lorient: Université de Bretagne Sud.

Thomas, J. (2009). Concordancing as a subversive activity. *Presentation at the PALC Conference* (unpublished).

Thomas, J. (2014). *Discovering English with the Sketch Engine*. Brno: Laptop Languages.