Language Sketching

Dragoș Ciobanu, from the translation department of the University of Leeds, wrote to me a few weeks ago:

> The reason I'm writing is to ask whether you've had a chance to play with Sketch Engine (see sketchengine.co.uk). In Leeds, we've been using it in our corpus linguistics work a lot. It's got brilliant features, from the terabytes of super useful multilingual data which it already comes with, to features for term extraction, specialized corpus building, thesaurus, collocations, and tons more! It's really, really cool and I'm only writing to you because the translators I know who have been playing with it also like it a lot.

Not sure whether you could tell, but Dragoș really likes Sketch Engine. And in a way, I could stop this column right here, because he already said it all—sort of.

After spending some time looking at Sketch Engine, I felt embarrassed that I hadn't known more about it. As Dragoș said, it's really, really cool. It's also a monster of a tool (size-wise) and it's not particularly easy to navigate when you first encounter it. (According to Ondřej Matuška of the Sketch Engine team, one of the areas they're trying to focus on in the immediate future is to make the product more user-friendly.)

But first, what exactly is Sketch Engine and what does it do?

It's a corpus tool developed by the Czech company Lexical Computing Limited. Lexical Computing was originally founded in 2003 by the late Brit Adam Kilgarriff and Pavel Rychlý, a professor at Masaryk University in Brno. The idea of corpus tools, and this corpus tool in particular, is to find how language behaves based on large collections of data. For this purpose, Sketch Engine built corpora in more than 80 languages (as well as "time-stamped" corpora in a slightly different set of 18 languages for the purpose of comparing word usage over time). The sizes of the corpora differ widely (from just a few million words in Maori to more than 800 billion in English), and they are available for a number of analysis purposes for any paying trial user. (The annual subscription price is 100 euros for non-academic users, with the trial period ending after 30 days.)

The analyses you can do on these corpora with Sketch Engine include the following:

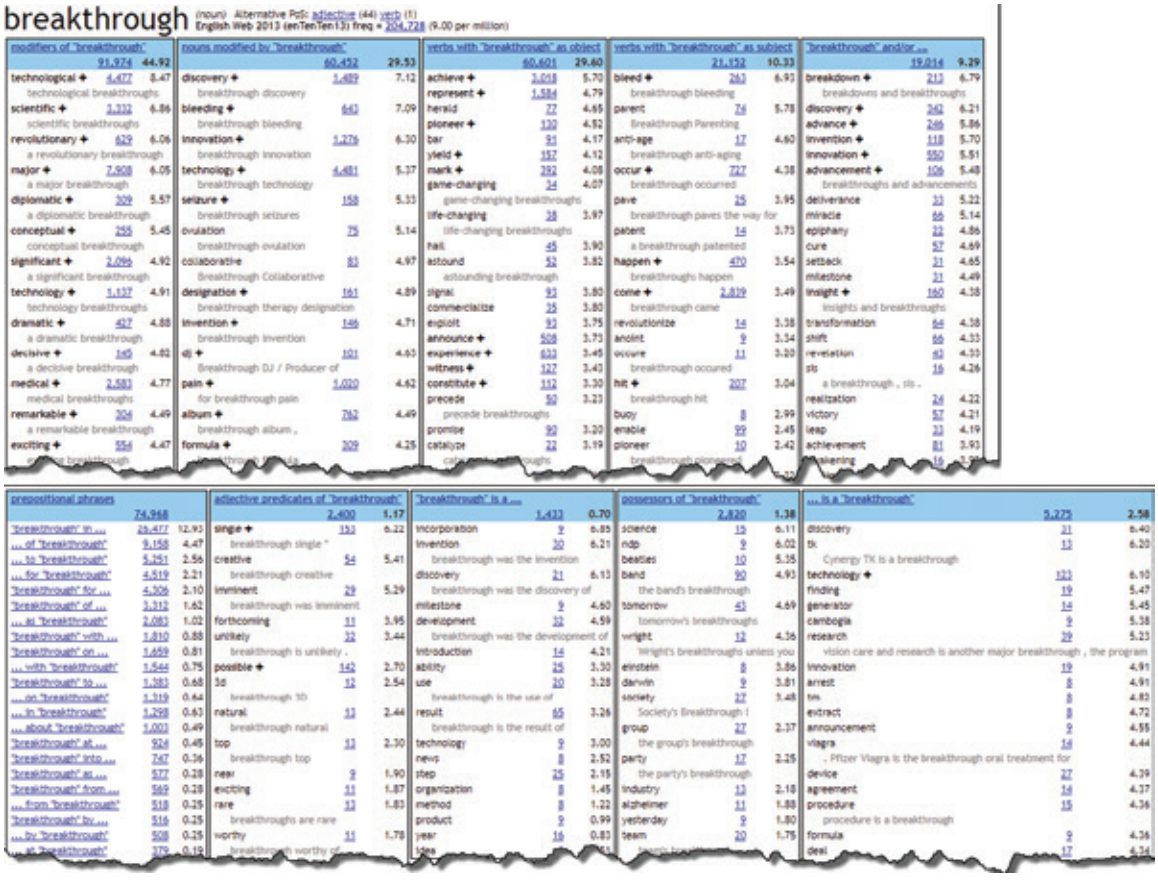

Figure 1: An example of a Word Sketch

Remember, if you have any ideas and/or suggestions regarding helpful resources or tools you would like to see featured, please e-mail me at jzetzsche@internationalwriters.com.

Word Sketches: This is where the program got its name, and it's what Kilgarriff brought to the table. A word sketch is a summary of a word's grammatical and collocational behavior (collocational refers to the analysis of how often a word co-occurs with other words or phrases. (See Figure 1 on page 36.) Since the data in the corpora is lemmatized (i.e., words are analyzed so they can be brought back to their base or dictionary form), the results are a lot more meaningful than what most of our translation environment tools provide when they're unable to relate different forms of one word to each other. Another word sketch option that Sketch Engine offers is the comparison of word sketches of similar words.

Thesaurus: The ability to retrieve a detailed list or a graphical word cloud with similar words, including links to create reports on word sketch differences for those terms to understand the exact differences in actual usage.

Concordance: Searches for single words, terms, or even longer phrases. Since the data in the supported languages is tagged, it's also possible to search for specific classes of words or specific classes of words that surround the word in question.

Parallel Corpus: Retrieval of bilingual or multilingual sets of words or phrases within the contexts. (See Figure 2.) Presently this is available only for on-screen data viewing, but it will soon be offered as downloadable data. This is especially helpful when uploading your own translation memories (see below).

Word Lists: The possibility of creating lists of words and the number of occurrences, either as lemmas (the base form of each word) or in each word form.

Creating Your Own Corpus: This is likely the most exciting feature for translators. You can either upload your own translation memories or use the tool's own search engine mechanism (which relies on Microsoft Bing) to create a list of bilingual websites that contain the terms that are relevant to your field, have them automatically align, and form a corpus. I don't need to explain to you the possibilities this offers to translators who don't have

Figure 2: The parallel corpus feature

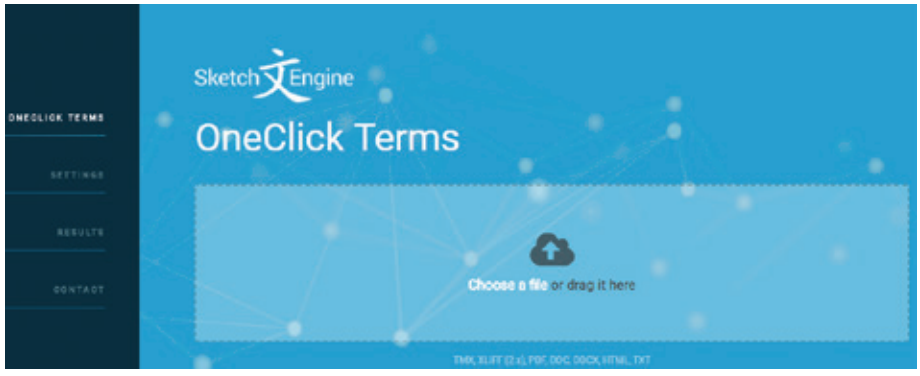

Figure 3: Preview of OneClick Terms

the privilege of having high-quality translation memories or termbases for a particular subject matter that they need to translate. As a logical extension of this feature, not only can you perform any of the functions mentioned earlier, but it's also possible to run a keyword search on the user-created corpus, identify the terms that are relevant, and download that into an Excel or TBX file. This feature is currently available for Chinese, Czech, Dutch, English, French, German, Italian, Japanese, Korean, Polish, Portuguese, Russian, and Spanish. The bilingual version of this is just around the corner.

By the way, you can find an example of the up-and-coming increased user-friendliness of Sketch Engine in OneClick Terms (terms.sketchengine.co.uk) that allows you to extract terms from TMX, XLIFF, PDF, DOC, DOCX, HTML, or TXT files in essentially one or two clicks. (See Figure 3.)

Translators have been one of the primary target groups for the makers of Sketch Engine. One immediate result of that focus is the availability of a plug-in for SDL Trados Studio (see http://bit.ly/SDL-appstore-sketch and http://bit.ly/user-guide-sketch). The plug-in itself is free, but it requires a trial or paid registration to be usable. It allows you to perform collocation, thesaurus, and concordance searches and will soon offer term extraction. According to Ondřej Matuska at Sketch Engine, talks with makers of other translation environment tools are under way to offer plug-ins or add-ons for those tools as well.

Can you believe you've never heard about this tool before? Well, maybe you were quicker than I to find this, but the good thing is that now we all know.

Jost Zetzsche is chair of ATA's Translation and Interpreting Resources Committee. He writes the "Geekspeak" column for *The ATA Chronicle*. He is also the co-author of *Found in Translation: How Language Shapes Our Lives and Transforms the World*, a robust source for replenishing your arsenal of information about how human translation and machine translation each play an important part in the broader world of translation. Contact: jzetzsche@internationalwriters.com.