**Translation Tools in the Sketch Engine**

**Adam Kilgarriff**

**Lexical Computing Ltd., Brighton, UK**

The Sketch Engine is a leading corpus query tool, in use for lexicography at OUP, CUP, Collins, Le Robert and Cornelsen, at national language institutes of eight countries, and for teaching and research in many universities. Its distinctive feature is the ‘word sketch’ a one page, automatic, summary of a word’s grammatical and collocational behaviour. Very large corpora and word sketches are available for fifty languages.

A number of tools and resources have recently been added with translators and terminologists in mind. These include automatic collocation dictionaries, parallel corpora, bilingual word sketches, and a term finder.

An automatic collocations dictionary is prepared by:

* Building a large, high-quality corpus (usually using web sources)
* Tokenising, lemmatising, part-of-speech-tagging, and dependency-parsing it
* Finding,
  + For each word
    - Its salient collocations, defined as those <grammatical relation, collocate> pairs that occur with it often and with high salience
    - For each collocation
      * An example sentence demonstrating a typical use

These components are then combined in a resource such as <http://forbetterenglish.com>

For translators, this will very quickly provide a large number of potential translation solutions. Trainee translators (following vocational courses, for example) and non-native speakers of a language translating into that language, can benefit from suggestions of phrases and structures that they can adapt or adopt.

Parallel corpora have proved of great value for translators, with Google translate, TAUS Data Association (<http://web2.tausdata.org:8801/>) and Linguee (<http://www.linguee.com>) – all built on parallel corpora -- proving three of the most significant additions to the translator’s toolbox in recent years. Our parallel corpus functionality is similar to Linguee (with less data per language pair, but for more pairs: currently around 100) with the addition of the ‘bilingual word sketch’, where we extend the widely used monolingual word sketches to include parallel-corpus-derived translation equivalents, for both single words and collocations (including terms). [[1]](#footnote-1) Over the last decade, word sketches have become a key resource for dictionary-making:

Editors have found that Word Sketches provide a compact and revealing snapshot of a word’s behaviour and uses. For many lexicographers with access to this kind of software, the lexical profile ha become the preferred starting point to their analyses of complex headwords. (Atkins and Rundell 2008, pp 110-111.)

We think it possible that bilingual word sketches will have a similar impact on translation over the next ten years.

The term-finder is based on the BootCaT procedure for instant web corpora in a particular domain. The user, typically a translator working in a domain where they are not an expert, inputs a few domain-specific ‘seed words’; these are sent to a search engine, and the hits identified by the search engine are gathered, cleaned, de-duplicated and processed to give a domain-specific corpus. This functionality has been found to be one of the best ways in which we can use the abundance of data on the web to support translators working with an unfamiliar genre or domain. Bernardini et al (2012) explore the benefits of using a range of resources for supporting trainee translators in translating Patient Information Leaflets (as found in drug packages) from English into Italian, and find that BootCaT provides a good tradeoff between how targeted a corpus is to the type of text, and how long it takes to build, allowing translators to work with a fairly well-targeted resource, of substantial size, in minutes.   
  
For some time, users have been benefiting from this functionality, and the opportunity it offers to find the keywords of the domain, in the Sketch Engine. The functionality has recently been extended so the user can find not just the key words, but the key terms, identified as those noun phrases in the text that display both ‘unithood’ (the two words belong together) and ‘termhood’ (they are much more common in the domain corpus than in a general reference corpus).

The presentation will include demonstrations, for various languages and language pairs, of all of this functionality.

**Acknowledgements**

The work described here was supported by the EU FP7 project PRESEMT.

**References**

B. T. S. Atkins and M. Rundell 2008. The Oxford Guide to Practical Lexicography. Oxford University Press.

S. Bernardini, A. Ferraresi and E. Zanchetta. Forthcoming 2012. Old needs, new solutions: comparable corpora for language professionals. In Sharoff, S., R. Rapp, P. Zweigenbaum, editors. Building and Using Comparable Corpora. Springer

1. Bilingual word sketches are currently in prototype and will be launched at the conference, provided the paper is accepted. [↑](#footnote-ref-1)