**Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine**

**Adam Kilgarriff**

adam@lexmasterclass.com

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

The [Sketch Engine](http://www.sketchengine.co.uk/) is a leading corpus query tool, in use for lexicography at OUP, CUP, Collins, Le Robert and Cornelsen, and 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, corpus, derived summary of a word’s grammatical and collocational behaviour. Very large corpora and word sketches are available for sixty languages.

A number of tools and resources have recently been added with translators and terminologists in mind. The resources are parallel corpora: EUROPARL-7 and the various datasets available in the OPUS collection. The tools are bilingual word sketches and the term finder.

**Parallel concordancing**

Parallel corpora have proved of great value for translators, with Google translate, TAUS Data Association (<http://web2.tausdata.org:8801/>) and <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 concordancing is shown in Figures 1 and 2.



Figure 1. English-French parallel concordance for *love/amour*



Figure 2. Greek-German parallel concordance for *αγάπη/Liebe*

This is similar to Linguee, with less data per language pair, but for many more pairs: currently around 300. As the screenshot shows, the Sketch Engine offers many ways to further explore the concordances, including sorting, filtering, frequency reports and collocation reports. Recent additions include querying in both languages simultaneously, so, eg, the aligned segments in Figure 2 are only those with both *αγάπη* in the Greek and *Liebe* in the German.

**Bilingual word sketches**

We have also developed the ‘bilingual word sketch’, where we extend the widely used monolingual word sketches to include data for two languages. In one version, "bip" or "bilingual-parallel" sketches, we derive matched headwords and collocations from parallel corpora, as in Figure 3. Here we can see that the tool has automatically identified the three English collocations (*written declaration, solemn declaration, unilateral declaration)*  and the corresponding French collocations (*déclaration écrite,* *déclaration solennel* *déclaration unilatérale),* also provided corpus citations for each.

In "bim" or "bilingual-manual' word sketches the user specifies which translation-pair of words they want to compare word sketches for, and they are then shown a word sketch with corresponding grammatical relations matched, as in Figure 4. Here the user has specified that they want to see English *house* and French *maison* side-by-side.

We see the pairs, under the 'object\_of/objet\_de' columns, *build/bâtir, buy/acheter, rent/louer,*

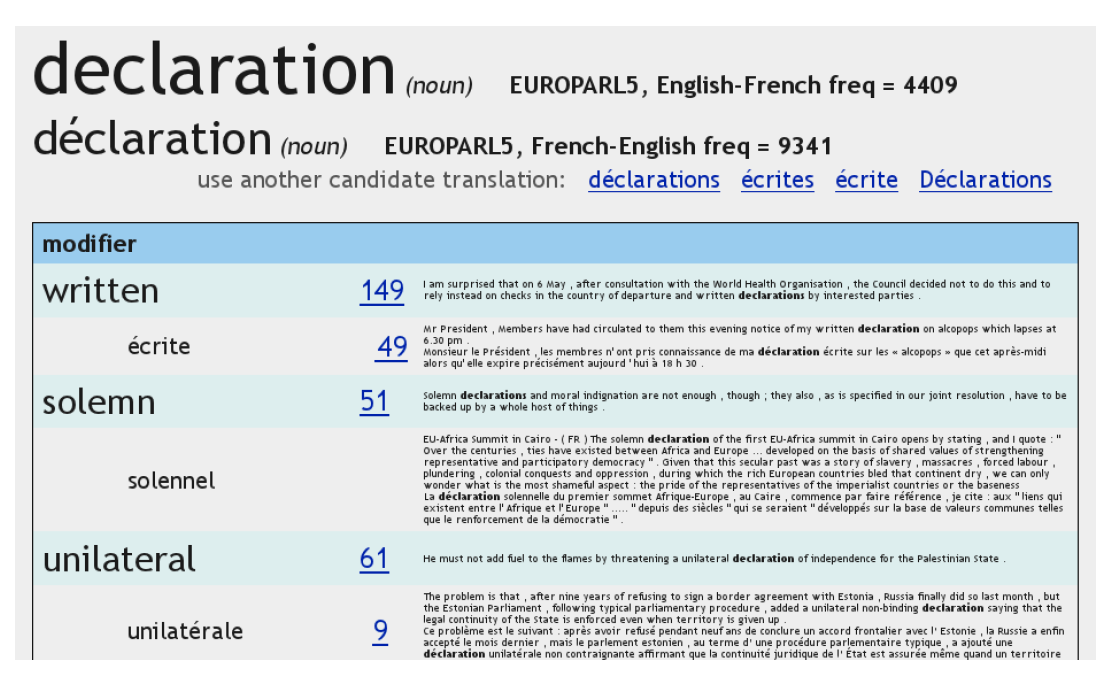


Figure 3: *Bip* word sketch for English *declaration,* with French *déclaration*

*leave/quitter.* This may well prove useful for language learners and translators. For lexicographers, it is perhaps what is missing that is most useful: which collocations for *house* do **not** have a French equivalent with *maison?* These are the items needing explicit mention in a bilingual dictionary. We are currently adding to the functionality to support that question.

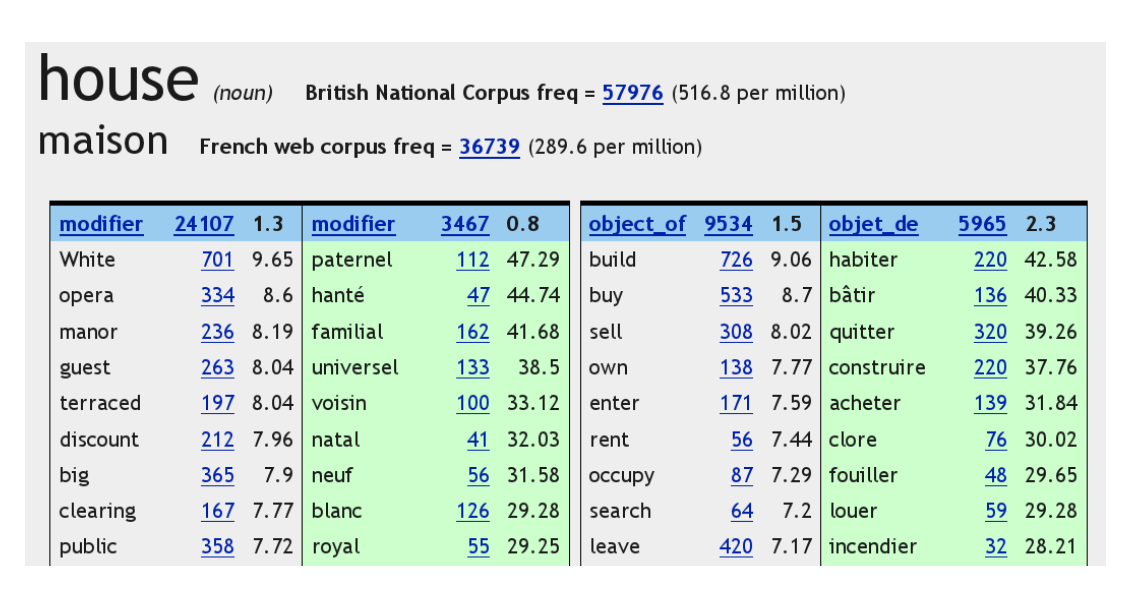


Figure 4: *Bim* word sketch for English *house,* with French mais*on*

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.)

Perhaps bilingual word sketches will have a similar impact on translation over the next ten years.

**Term finding**

The term-finder starts from a domain corpus, and a reference corpus. First it finds all the noun phrases, and their frequencies, on both corpora. It then takes the ratio, and the items with highest ratios will be terms, as in Figures 5 and 6 (where the data was supplied by the first users of this technology, the World Intellectual Property Organisation).



Figure 5. French terms in the mobile communications domain.

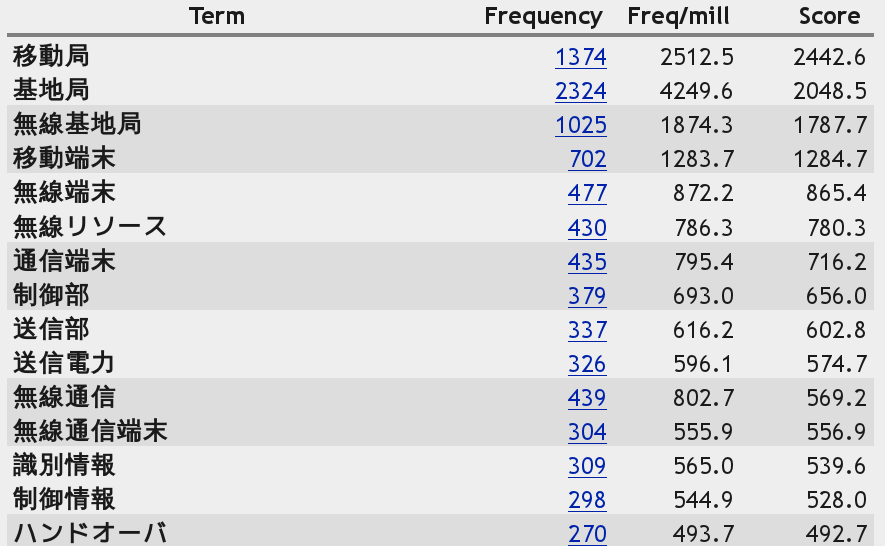


Figure 6. Japanese terms in the mobile communications domain.

In some cases, as with WIPO, the user will have domain corpora, but in others they will not. In that case they may use the BootCaT procedure (Baroni and Bernardini 2004). 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 support translators well (Bernardini et al 2013). For some time, the Sketch Engine has incorporated a BootCaT tool, allowing users to create an instant corpus for a domain, which means they can then compare this corpus with a reference corpus to find the keywords of the domain. The functionality has recently been extended so the user can find the terms alongside key words. Thus, where the user has Bootcatted an English environment corpus, the Sketch Engine provides the "key words and terms" report shown in Figure 7.

The requirements for the term-finding functionality are:

* a processing chain, comprising tokeniser, lemmatiser and part-of-speech tagger, installed and ready to apply to the user's domain corpus
* a reference corpus processed with the processing chain
* a term grammar.

At time of writing, these are all in place for Chinese, English, French, German, Japanese, Korean, Russian, Spanish and Portuguese. More languages will be added over the coming year.

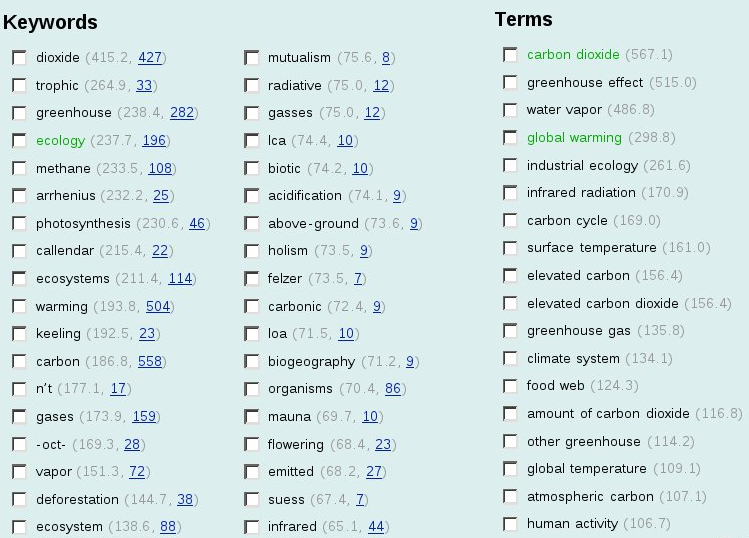


Figure 7. English key words and terms in the environment domain. The tickboxes are so the user can easily specify a new set of seed words and terms so they can refine the domain corpus by iterating the BootCaT procedure so they get more on-domain, and less off-domain text.

**In sum**

The Sketch Engine has for some years been a leading tool for lexicography and corpus linguistics. Over that period, it has built up corpus resources and functionality which are relevant for translators and terminologists, but not specialised for them. In the last year, translators and terminologists have been the target of our development efforts, and we now have a number of tools designed specifically for them: many parallel corpora covering many language pairs; improved parallel concordancing; bilingual word sketches; and term finding. We hope you will find them interesting.

**References**

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

M. Baroni and S. Bernardini. 2004. [BootCaT: Bootstrapping corpora and terms from the web](http://clic.cimec.unitn.it/marco/publications/lrec2004/bootcat_lrec_2004.pdf). Proceedings of LREC 2004, Lisbon: ELDA. 1313-1316.

S. Bernardini, A. Ferraresi and E. Zanchetta. 2013. Old needs, new solutions: comparable corpora for language professionals. In Sharoff, S., R. Rapp, P. Zweigenbaum, P. Fung, editors. Building and Using Comparable Corpora. Springer