

## 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](#) 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.

The screenshot displays the Sketch Engine interface for a parallel concordance search. The top navigation bar includes links for 'About', 'Home', 'Settings', 'Change password', and 'Log out'. A search bar is present with the text 'Search' and 'in Help'. Below the navigation bar, the user information is shown as 'user: Dr. Adam Kilgarriff' and the corpus as 'corpus: EUROPARL7, en'. The search term 'love' is entered in the search box, and the results are shown for the 'EUROPARL7, en' corpus. The search results are displayed in a table with two columns: 'EUROPARL7, en' and 'EUROPARL7, fr'. The table shows several concordance pairs, with the word 'love' in the English column and 'amour' in the French column. The first row shows the English text: 'I am speaking for the first time in this plenary part-session, so this is quite exciting for me, a little like first love, although that did last longer than two minutes.' and the French text: 'C ' est la première fois que je prends la parole en plénière, il y a donc de quoi être un peu nerveux, un peu comme avec le premier amour, mais le premier amour a quand même duré heureusement plus de deux minutes.' The second row shows the English text: 'And since this is St. Valentine ' s day, as a former Mayor of a regional city, I propose that we should all declare our love for all the European regions which need that love.' and the French text: 'Et, puisque aujourd ' hui, c ' est la Saint-Valentin, en tant qu ' ancien maire d ' une ville régionale, je propose que nous déclarions notre amour envers les régions européennes qui en ont besoin.' The third row shows the English text: 'And since this is St. Valentine ' s day, as a former Mayor of a regional city, I propose that we should all declare our love for all the European regions which need that love.' and the French text: 'Et, puisque aujourd ' hui, c ' est la Saint-Valentin, en tant qu ' ancien maire d ' une ville régionale, je propose que nous déclarions notre amour envers les régions européennes qui en ont besoin.' The fourth row shows the English text: 'Nevertheless, it is this very love that is a requirement for the healthy development of the individual.' and the French text: 'Or c ' est précisément cet amour qui est la condition de l ' épanouissement de l ' individu, et l ' Europe n ' a que faire de droits fondamentaux progressistes si les membres de la société ne veulent pas les respecter.' The fifth row shows the English text: 'Indeed, the word of God repeatedly and emphatically speaks of hospitality and mercifulness to strangers, as well as true charity as a consequence of our love for God, the Creator of all mankind.' and the French text: 'La parole divine insiste en effet à maintes reprises et avec insistance sur la nécessité d ' adopter une attitude accueillante et de témoigner des marques de charité à l ' égard des étrangers, l ' amour du prochain étant le corollaire de l ' amour porté à Dieu, notre Créateur à tous.'

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

Sketch Engine

About Home Settings Change password Log out

Search in Help

user: Siva Kalyan corpus: EUROPARL7, el Search αγάπη in EUROPARL7, el

Concordance Word List

Save View options KWIC Sentence Alignment Sort Left Right Node Shuffle Sample Filter Frequency Node forms Doc IDs Collocations ConcDesc

Query αγάπη, Liebe 77 (1.7 per million)

Page 1 of 4 Go Next | Last

| EUROPARL7, el  | EUROPARL7, de   |
|--|---|
| Παρ ' όλα αυτά , αυτή ακριβώς η <b>αγάπη</b> αποτελεί προϋπόθεση για την ανάπτυξη ενός υγιούς ατόμου .   | Dennoch ist gerade diese <b>Liebe</b> Voraussetzung für eine gesunde Persönlichkeitsentwicklung .   |
| Ο λόγος του Θεού αναφέρεται επανειλημμένα και με έμφαση στη φιλοξενία και την ευσπλαχνία προς τον ξένο και στην πραγματική <b>αγάπη</b> προς τον πλησίον ως απόρροια της αγάπης του ίδιου του Δημιουργού προς τον άνθρωπο .  | Gottes Wort fordert ja wiederholt und nachdrücklich Freundlichkeit und Barmherzigkeit gegenüber Fremden und gebietet tätige Nächstenliebe als Folge der <b>Liebe</b> Gottes , unser aller Schöpfer .  |
| Θέλω επίσης να υπενθυμίσω σε εκείνους που ασπάζονται μια εθνοτική αντίληψη του έθνους τη ρήση του Clémenceau : " Πατριωτισμός είναι η <b>αγάπη</b> για την χώρα σου , εθνικισμός είναι το μίσος για τους άλλους " .  | Und die Verfechter einer ethnischen Auffassung von der Nation möchte ich an die Worte von Clémenceau erinnern : " Patriotismus ist die <b>Liebe</b> zum Vaterland , Nationalismus ist der Haß auf die anderen Länder . "  |
| Χρειάζεται χρόνος , υπομονή , <b>αγάπη</b> , για να τους ξαναδώσουμε την ελπίδα .  | Es braucht Zeit , Geduld , <b>Liebe</b> , um ihnen wieder Hoffnung zu geben .   |
| Όπως είπαν ορισμένοι εμπειρογνώμονες , τα θρησκευτικά θέματα δεν επιδέχονται σταθμίσεις , αυτό όμως που είναι βέβαιο είναι ότι αν μιλάμε για τις θρησκείες της Βίβλου , που κηρύττουν την <b>αγάπη</b> προς το Θεό και το συνάνθρωπο , τότε πρέπει να είναι δυνατόν να βρούμε μια λύση ανθρώπινη . | Wie Experten sagen , kann man bei religiösen Themen nicht vermitteln , wenn wir jedoch über die Religionen der Heiligen Schrift sprechen , die die <b>Liebe</b> zu Gott und zum Nächsten predigen , muss es möglich sein , dass Menschen untereinander eine Lösung finden . |

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,*

**declaration** (*noun*) EUROPARL5, English-French freq = 4409

**déclaration** (*noun*) EUROPARL5, French-English freq = 9341

use another candidate translation: [déclarations](#) [écrites](#) [écrite](#) [Déclarations](#)

| modifier    |                     |  |
|-------------|---------------------|--|
| written     | <a href="#">149</a> | I am surprised that on 6 May, after consultation with the World Health Organisation, the council decided not to do this and to rely instead on checks in the country of departure and written <b>declarations</b> by interested parties.   |
| écrite      | <a href="#">49</a>  | Mr President, Members have had circulated to them this evening notice of my written <b>declaration</b> on alcopops which lapses at 6.30 pm.<br>Monsieur le Président, les membres n'ont pris connaissance de ma <b>déclaration</b> écrite sur les « alcopops » que cet après-midi alors qu'elle expire précisément aujourd'hui à 18 h 30.  |
| solemn      | <a href="#">51</a>  | Solemn <b>declarations</b> and moral indignation are not enough, though; they also, as is specified in our joint resolution, have to be backed up by a whole host of things.   |
| solennel    |                     | EU-Africa Summit in Cairo - (FR) The solemn <b>declaration</b> of the first EU-Africa summit in Cairo opens by stating, and I quote: "Over the centuries, ties have existed between Africa and Europe... developed on the basis of shared values of strengthening representative and participatory democracy". Given that this secular past was a story of slavery, massacres, forced labour, plundering, colonial conquests and oppression, during which the rich European countries bled that continent dry, we can only wonder what is the most shameful aspect: the pride of the representatives of the imperialist countries or the baseness.<br>La <b>déclaration</b> solennelle du premier sommet Afrique-Europe, au Caire, commence par faire référence, je cite: aux "liens qui existent entre l'Afrique et l'Europe"... "depuis des siècles" qui se seraient "développés sur la base de valeurs communes telles que le renforcement de la démocratie". |
| unilateral  | <a href="#">61</a>  | He must not add fuel to the flames by threatening a unilateral <b>declaration</b> of independence for the Palestinian State.   |
| unilatérale | <a href="#">9</a>   | The problem is that, after nine years of refusing to sign a border agreement with Estonia, Russia finally did so last month, but the Estonian Parliament, following typical parliamentary procedure, added a unilateral non-binding <b>declaration</b> saying that the legal continuity of the state is enforced even when territory is given up.<br>Ce problème est le suivant: après avoir refusé pendant neuf ans de conclure un accord frontalier avec l'Estonie, la Russie a enfin accepté le mois dernier, mais le parlement estonien, au terme d'une procédure parlementaire typique, a ajouté une <b>déclaration</b> unilatérale non contraignante affirmant que la continuité juridique de l'Etat est assurée même quand un territoire  |

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

*leave/quit*. 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.

**house** (*noun*) British National Corpus freq = [57976](#) (516.8 per million)

**maison** French web corpus freq = [36739](#) (289.6 per million)

| modifier | <a href="#">24107</a> | 1.3  | modifier  | <a href="#">3467</a> | 0.8   | object_of | <a href="#">9534</a> | 1.5  | objet_de   | <a href="#">5965</a> | 2.3   |
|----------|-----------------------|------|-----------|----------------------|-------|-----------|----------------------|------|------------|----------------------|-------|
| White    | <a href="#">701</a>   | 9.65 | paternal  | <a href="#">112</a>  | 47.29 | build     | <a href="#">726</a>  | 9.06 | habiter    | <a href="#">220</a>  | 42.58 |
| opera    | <a href="#">334</a>   | 8.6  | hanté     | <a href="#">47</a>   | 44.74 | buy       | <a href="#">533</a>  | 8.7  | bâtir      | <a href="#">136</a>  | 40.33 |
| manor    | <a href="#">236</a>   | 8.19 | familial  | <a href="#">162</a>  | 41.68 | sell      | <a href="#">308</a>  | 8.02 | quitter    | <a href="#">320</a>  | 39.26 |
| guest    | <a href="#">263</a>   | 8.04 | universel | <a href="#">133</a>  | 38.5  | own       | <a href="#">138</a>  | 7.77 | construire | <a href="#">220</a>  | 37.76 |
| terraced | <a href="#">197</a>   | 8.04 | voisin    | <a href="#">100</a>  | 33.12 | enter     | <a href="#">171</a>  | 7.59 | acheter    | <a href="#">139</a>  | 31.84 |
| discount | <a href="#">212</a>   | 7.96 | natal     | <a href="#">41</a>   | 32.03 | rent      | <a href="#">56</a>   | 7.44 | clore      | <a href="#">76</a>   | 30.02 |
| big      | <a href="#">365</a>   | 7.9  | neuf      | <a href="#">56</a>   | 31.58 | occupy    | <a href="#">87</a>   | 7.29 | fouiller   | <a href="#">48</a>   | 29.65 |
| clearing | <a href="#">167</a>   | 7.77 | blanc     | <a href="#">126</a>  | 29.28 | search    | <a href="#">64</a>   | 7.2  | louer      | <a href="#">59</a>   | 29.28 |
| public   | <a href="#">358</a>   | 7.72 | royal     | <a href="#">55</a>   | 29.25 | leave     | <a href="#">420</a>  | 7.17 | incendier  | <a href="#">32</a>   | 28.21 |

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

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 has 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).

| Term                      | Frequency             | Freq/mill | Score  |
|---------------------------|-----------------------|-----------|--------|
| station de base           | <a href="#">28612</a> | 3292.2    | 3293.2 |
| station mobile            | <a href="#">12514</a> | 1439.9    | 1440.9 |
| communication sans fil    | <a href="#">8189</a>  | 942.3     | 943.3  |
| liaison montante          | <a href="#">6561</a>  | 754.9     | 737.5  |
| terminal mobile           | <a href="#">7406</a>  | 852.2     | 709.8  |
| liaison descendante       | <a href="#">5434</a>  | 625.3     | 626.3  |
| stations de base          | <a href="#">5010</a>  | 576.5     | 577.5  |
| réseau de communication   | <a href="#">4255</a>  | 489.6     | 490.6  |
| communication mobile      | <a href="#">4722</a>  | 543.3     | 462.5  |
| point d' accès            | <a href="#">3907</a>  | 449.6     | 450.6  |
| modes de réalisation      | <a href="#">3486</a>  | 401.1     | 402.1  |
| réseau d' accès           | <a href="#">3241</a>  | 372.9     | 373.9  |
| réseau sans fil           | <a href="#">2903</a>  | 334.0     | 335.0  |
| accès radio               | <a href="#">2412</a>  | 277.5     | 278.5  |
| transfert intercellulaire | <a href="#">2408</a>  | 277.1     | 278.1  |

Figure 5. French terms in the mobile communications domain.



| Term   | Frequency            | Freq/mill | Score  |
|--------|----------------------|-----------|--------|
| 移動局    | <a href="#">1374</a> | 2512.5    | 2442.6 |
| 基地局    | <a href="#">2324</a> | 4249.6    | 2048.5 |
| 無線基地局  | <a href="#">1025</a> | 1874.3    | 1787.7 |
| 移動端末   | <a href="#">702</a>  | 1283.7    | 1284.7 |
| 無線端末   | <a href="#">477</a>  | 872.2     | 865.4  |
| 無線リソース | <a href="#">430</a>  | 786.3     | 780.3  |
| 通信端末   | <a href="#">435</a>  | 795.4     | 716.2  |
| 制御部    | <a href="#">379</a>  | 693.0     | 656.0  |
| 送信部    | <a href="#">337</a>  | 616.2     | 602.8  |
| 送信電力   | <a href="#">326</a>  | 596.1     | 574.7  |
| 無線通信   | <a href="#">439</a>  | 802.7     | 569.2  |
| 無線通信端末 | <a href="#">304</a>  | 555.9     | 556.9  |
| 識別情報   | <a href="#">309</a>  | 565.0     | 539.6  |
| 制御情報   | <a href="#">298</a>  | 544.9     | 528.0  |
| ハンドオーバ | <a href="#">270</a>  | 493.7     | 492.7  |

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.

| Keywords   |   | Terms   |
|--|---|---|
| <input type="checkbox"/> dioxide (415.2, <a href="#">427</a> )       | <input type="checkbox"/> mutualism (75.6, <a href="#">8</a> )     | <input type="checkbox"/> carbon dioxide (567.1)           |
| <input type="checkbox"/> trophic (264.9, <a href="#">33</a> )        | <input type="checkbox"/> radiative (75.0, <a href="#">12</a> )    | <input type="checkbox"/> greenhouse effect (515.0)        |
| <input type="checkbox"/> greenhouse (238.4, <a href="#">282</a> )    | <input type="checkbox"/> gasses (75.0, <a href="#">12</a> )       | <input type="checkbox"/> water vapor (486.8)              |
| <input type="checkbox"/> ecology (237.7, <a href="#">196</a> )       | <input type="checkbox"/> lca (74.4, <a href="#">10</a> )          | <input type="checkbox"/> global warming (298.8)           |
| <input type="checkbox"/> methane (233.5, <a href="#">108</a> )       | <input type="checkbox"/> biotic (74.2, <a href="#">10</a> )       | <input type="checkbox"/> industrial ecology (261.6)       |
| <input type="checkbox"/> arrhenius (232.2, <a href="#">25</a> )      | <input type="checkbox"/> acidification (74.1, <a href="#">9</a> ) | <input type="checkbox"/> infrared radiation (170.9)       |
| <input type="checkbox"/> photosynthesis (230.6, <a href="#">46</a> ) | <input type="checkbox"/> above-ground (73.6, <a href="#">9</a> )  | <input type="checkbox"/> carbon cycle (169.0)             |
| <input type="checkbox"/> callendar (215.4, <a href="#">22</a> )      | <input type="checkbox"/> holism (73.5, <a href="#">9</a> )        | <input type="checkbox"/> surface temperature (161.0)      |
| <input type="checkbox"/> ecosystems (211.4, <a href="#">114</a> )    | <input type="checkbox"/> felzer (73.5, <a href="#">7</a> )        | <input type="checkbox"/> elevated carbon (156.4)          |
| <input type="checkbox"/> warming (193.8, <a href="#">504</a> )       | <input type="checkbox"/> carbonic (72.4, <a href="#">9</a> )      | <input type="checkbox"/> elevated carbon dioxide (156.4)  |
| <input type="checkbox"/> keeling (192.5, <a href="#">23</a> )        | <input type="checkbox"/> loa (71.5, <a href="#">10</a> )          | <input type="checkbox"/> greenhouse gas (135.8)           |
| <input type="checkbox"/> carbon (186.8, <a href="#">558</a> )        | <input type="checkbox"/> biogeography (71.2, <a href="#">9</a> )  | <input type="checkbox"/> climate system (134.1)           |
| <input type="checkbox"/> n't (177.1, <a href="#">17</a> )            | <input type="checkbox"/> organisms (70.4, <a href="#">86</a> )    | <input type="checkbox"/> food web (124.3)                 |
| <input type="checkbox"/> gases (173.9, <a href="#">159</a> )         | <input type="checkbox"/> mauna (69.7, <a href="#">10</a> )        | <input type="checkbox"/> amount of carbon dioxide (116.8) |
| <input type="checkbox"/> -oct- (169.3, <a href="#">28</a> )          | <input type="checkbox"/> flowering (68.4, <a href="#">23</a> )    | <input type="checkbox"/> other greenhouse (114.2)         |
| <input type="checkbox"/> vapor (151.3, <a href="#">72</a> )          | <input type="checkbox"/> emitted (68.2, <a href="#">27</a> )      | <input type="checkbox"/> global temperature (109.1)       |
| <input type="checkbox"/> deforestation (144.7, <a href="#">38</a> )  | <input type="checkbox"/> suess (67.4, <a href="#">7</a> )         | <input type="checkbox"/> atmospheric carbon (107.1)       |
| <input type="checkbox"/> ecosystem (138.6, <a href="#">88</a> )      | <input type="checkbox"/> infrared (65.1, <a href="#">44</a> )     | <input type="checkbox"/> human activity (106.7)           |

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](#). 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