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.

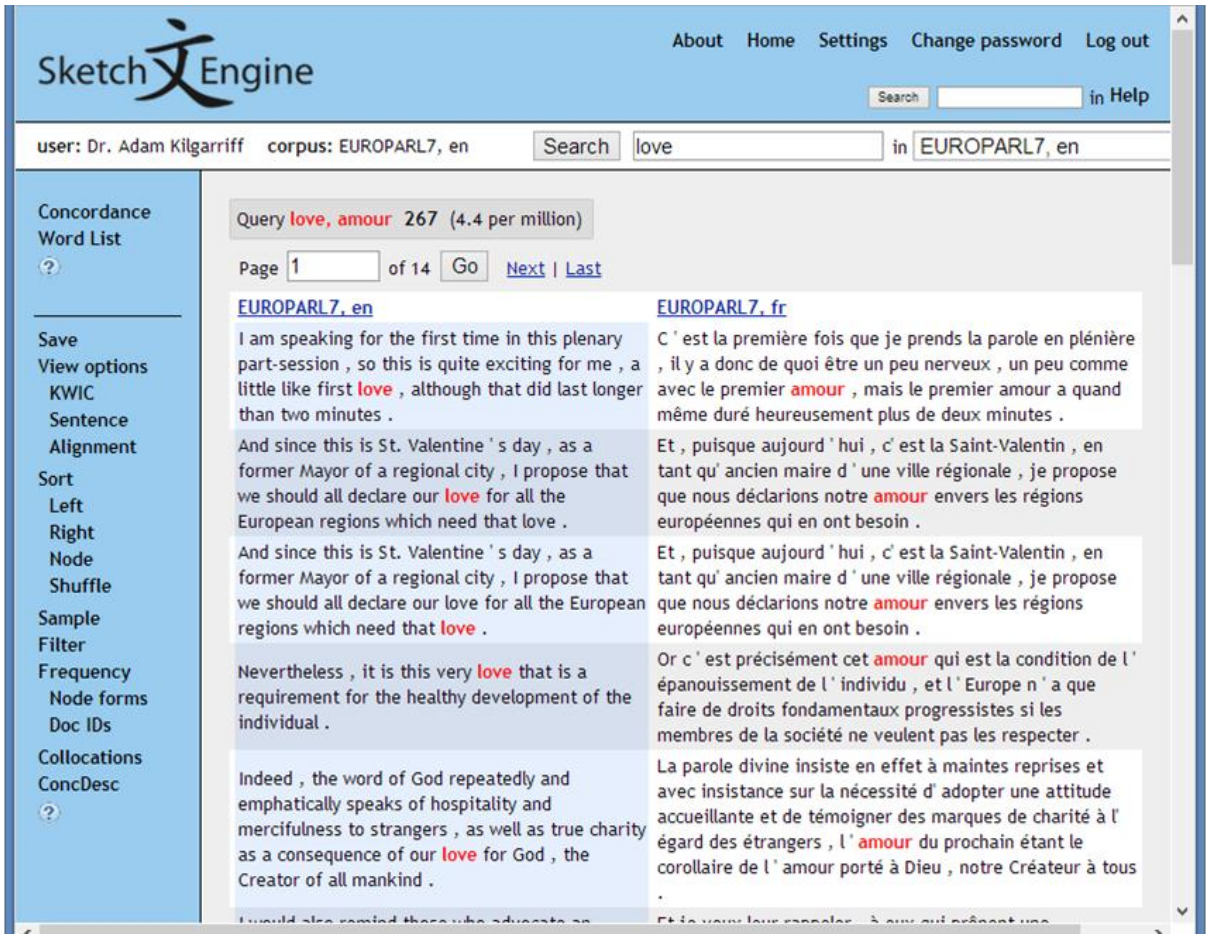


Figure 1. English-French parallel concordance for *love/amour*

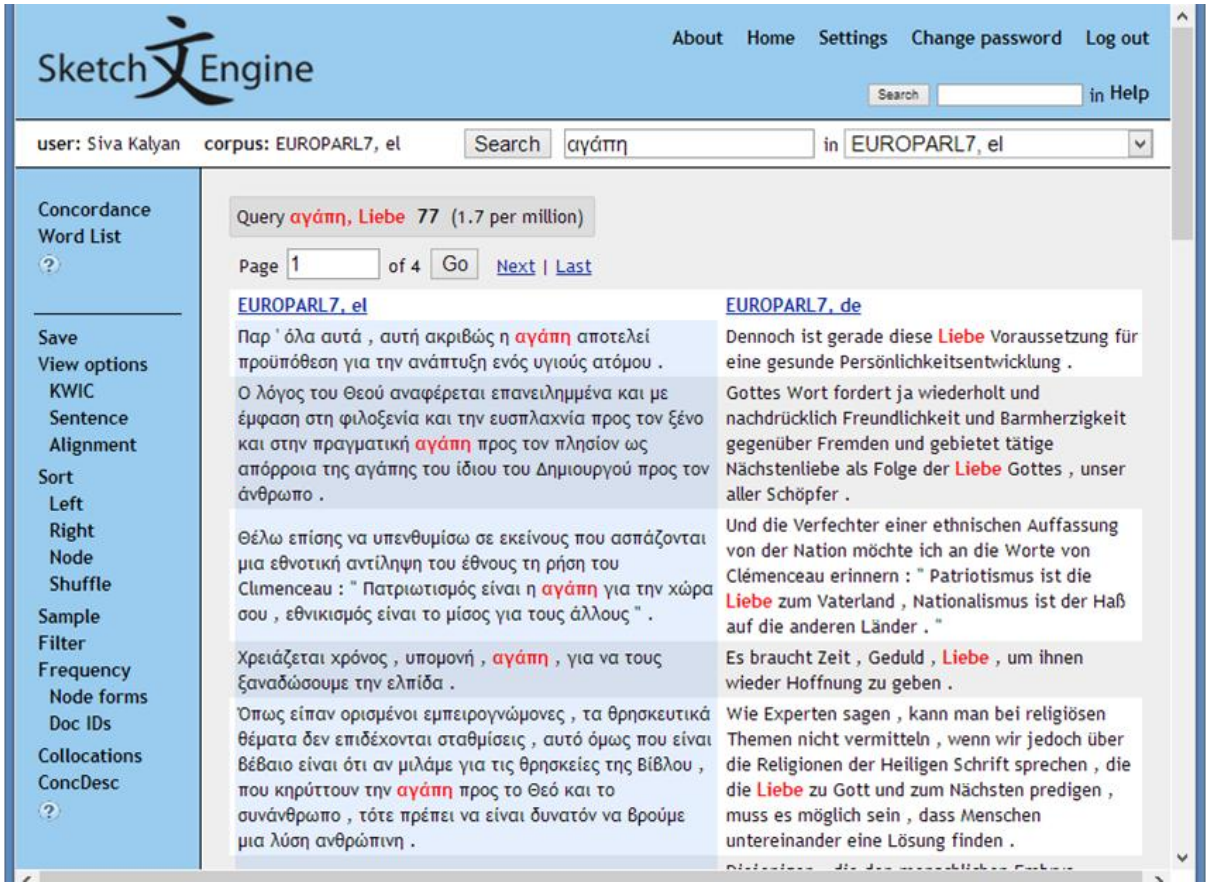


Figure 2. Greek-German parallel concordance for *αγάπη/Liebe*

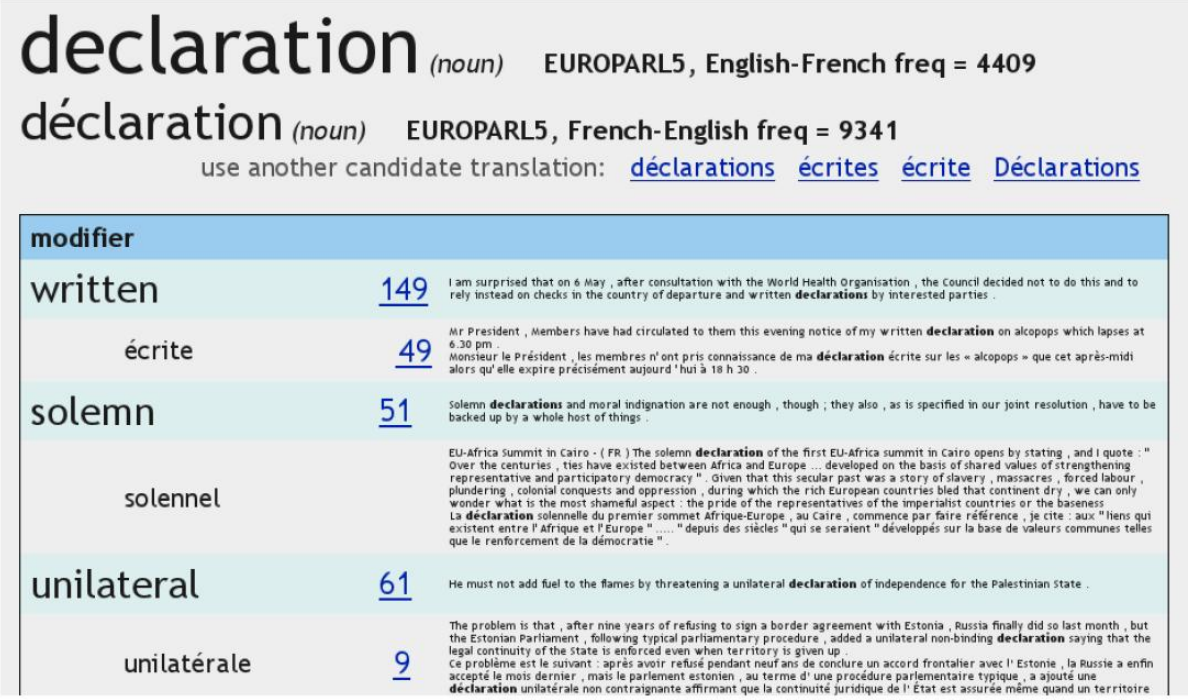
This is similar to Linguee, with less data per language pair, but for many more pairs: currently around 300. As the screenshot shows, the Sketch Engine offers many ways to further explore the concordances, including sorting, filtering, frequency reports and collocation reports. Recent additions include 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,*



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

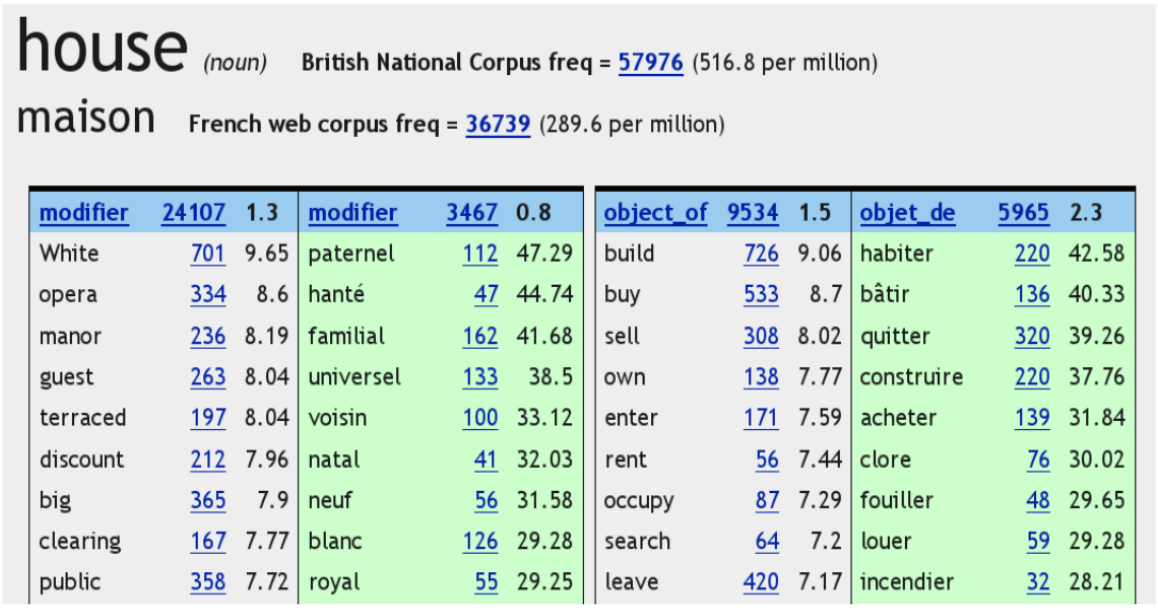
déclaration *(noun)* **EUROPARL5, French-English freq = 9341**

use another candidate translation: déclarations écrites écrite Déclarations

modifier		
written	149	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 **declarations** by interested parties .
écrite	49	Mr President , Members have had circulated to them this evening notice of my written **declaration** on alcopops which lapses at 6.30 pm . Monsieur le Président , les membres n' ont pris connaissance de ma **déclaration** écrite sur les « alcopops » que cet après-midi alors qu' elle expire précisément aujourd ' hui à 18 h 30 .
solemn	51	Solemn **declarations** 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 **declaration** 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 La **déclaration** 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	61	He must not add fuel to the flames by threatening a unilateral **declaration** of independence for the Palestinian State .
unilatérale	9	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 **declaration** saying that the legal continuity of the State is enforced even when territory is given up . 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 **déclaration** unilatérale non contraignante affirmant que la continuité juridique de l' État est assurée même quand un territoire

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.



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

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

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

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).

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

Figure 5. French terms in the mobile communications domain.

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

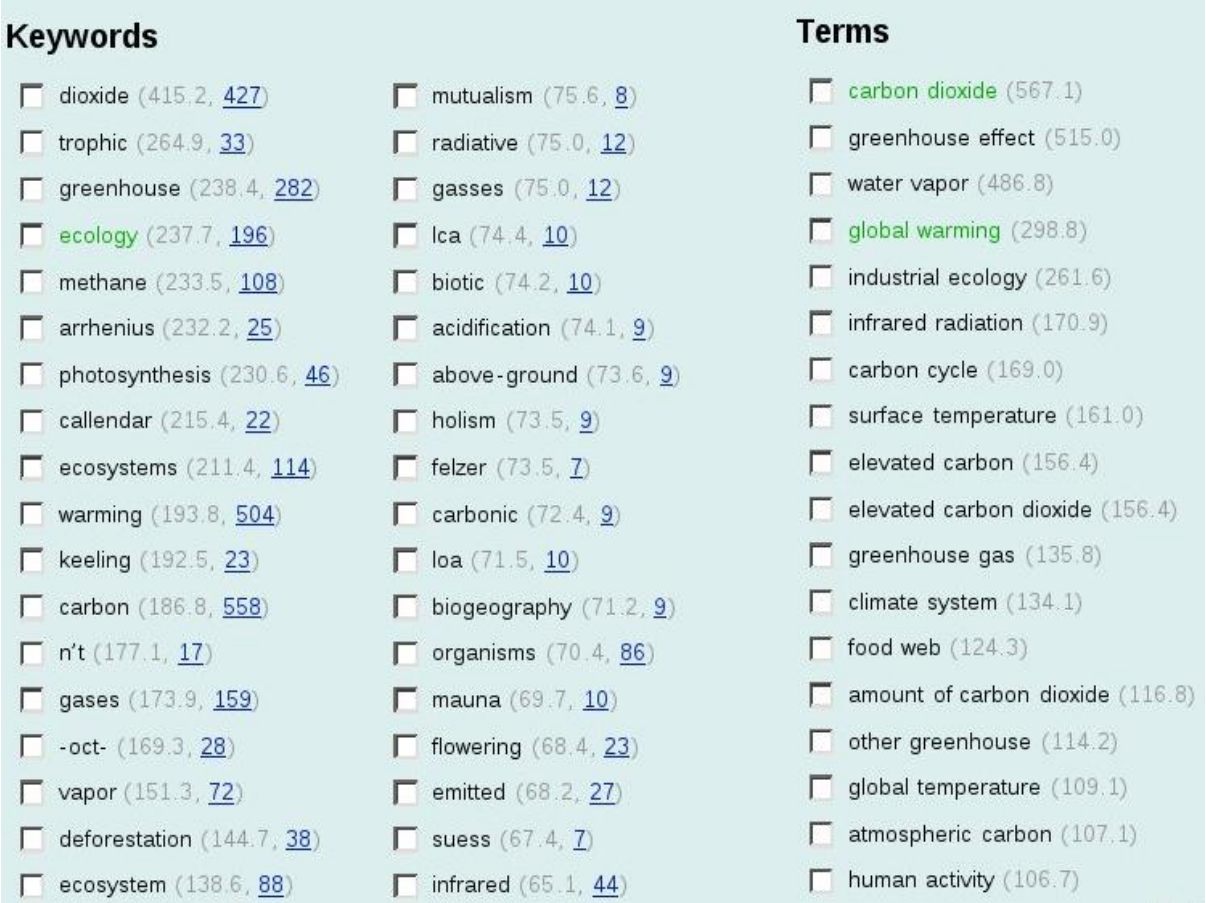


Figure 7. English key words and terms in the environment domain. The tickboxes are so the user can easily specify a new set of seed words and terms so they can refine the domain corpus by iterating the BootCaT procedure so they get more on-domain, and less off-domain text.

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