See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/305314795

Sketch Engine for Bilingual Lexicography

Article in International Journal of Lexicography · July 2016

DOI: 10.1093/ijl/ecw029

| citations 0 | 5 | reads 98 | |
|----------------|------------------------------------------------------------------------------------------|-------------|---------------------------------------------------------------------------------|
| 3 author | rs: | | |
| | Vojtěch Kovář Masaryk University 33 PUBLICATIONS 183 CITATIONS SEE PROFILE | | Vít Baisa Masaryk University 24 PUBLICATIONS 139 CITATIONS SEE PROFILE |
| | Miloš Jakubíček Lexical Computing Ltd 44 PUBLICATIONS 214 CITATIONS SEE PROFILE | | |

Some of the authors of this publication are also working on these related projects:



SDL Trados plugin View project

AQA: Automatic Question Answering for Czech View project

All content following this page was uploaded by Vojtěch Kovář on 04 January 2018.

Sketch Engine for bilingual lexicography

Vojtěch Kovář, Vít Baisa and Miloš Jakubíček

Abstract

Sketch Engine is a leading corpus query and corpus management tool that has been used for many large dictionary projects. The paper summarizes its features supporting bilingual lexicography and the creation of bilingual learner's dictionaries. Some of these features have been added recently; some of them have been part of the software for a rather long time, but they have been recently improved.

1. Background

Sketch Engine¹ is a leading corpus querying and corpus management tool created by Adam Kilgarriff (Kilgarriff et al. 2014). Since then, it has been used in many significant learners' dictionary projects led by recognized international institutions, including e.g. Oxford University Press or Macmillan.

The system is based on a specialized database engine Manatee (Rychlý, 2007) which enables fast evaluation of queries over very large corpora. Sketch Engine imposes practically no limitations on corpus size; the biggest corpus that has been tested in practice so far contained about 80 billion words. It comes with a graphical user interface called Bonito, which is web-based since version 2 (2004) – so the whole system is presented to the users as a web application.

In 2010, Sketch Engine introduced a corpus building and corpus management module called Corpus Architect that enables straightforward creation of user corpora from the users' own data, or from the web. All parts of the system are language independent.

A prominent feature of Sketch Engine is a so called *word sketch* – a one-page summary of collocational behaviour of a particular word, statistically derived from the corpus data and structured according to grammatical patterns in which they occur (also called *grammatical relations*, or simply *relations*). An example of a word sketch is shown in Figure 1. Apart from this key feature the system offers:

- concordancer with rich query and view options
- compilation of various word and n-gram lists
- defining sub-corpora and working with them
- creating frequency distributions according to various criteria
- statistical thesaurus derived from the word sketch data
- visual difference between words, based on word sketch data
- algorithm for suggesting good dictionary examples (Rychlý et al. 2008)
- *tickbox lexicography* application for streamlining lexicography work (Kilgarriff et al. 2010)
- over 400 preloaded corpora for more than 80 languages

Corpus building facilities incorporated into the Corpus Architect include:

- building corpora from users' data in various formats
- building corpora from the web, using key words or lists of URLs
- various support functions for management and sharing corpora
- built-in third party tools for tokenisation and morphological tagging of user corpora

| modifiers of "fire" | | | nouns and vert | os modif | verbs with "fire" as | | | |
|---------------------|-----------------|----------|-----------------|--------------|----------------------|------------|--------------|-------|
| | <u>3,392</u> | 0.90 | <u>"fire"</u> | | | object | | |
| log | <u>96</u> | 9.64 | | <u>3,168</u> | 0.90 | | <u>3,334</u> | 2.40 |
| a log fire | | | brigade + | <u>282</u> | 11.11 | light + | <u>208</u> | 10.34 |
| gas + | <u>150</u> | 9.48 | the fire briga | de | | catch + | <u>197</u> | 9.34 |
| the gas fire | | | engine + | <u>174</u> | 9.67 | caught fi | ire . | |
| electric + | 110 | 9.26 | fire engine | | | open + | <u>255</u> | 9.20 |
| electric fire | | | extinguisher | <u>69</u> | 9.43 | opened | fire on | |
| coal | 71 86 | | fire extinguis | shers | | set + | <u>280</u> | 8.87 |
| a coal fire | | | escape | <u>72</u> | 9.29 | set fire t | 0 | |
| open + | 144 | 8 56 | the fire esca | pe | | machine- | 37 | 8 /17 |
| an open fire | <u></u> | 0.00 | crew | <u>78</u> | 9.01 | gun | 57 | 0.47 |
| artillery | rtillon 2E 9.00 | | fire crews | | | of machi | ine-gun f | fire |
| artillary fire | <u>55</u> | 0.20 | fighter | <u>61</u> | 8.88 | start + | <u>124</u> | 8.33 |
| aninery me . | 20 | 0 1 1 | fire fighters | | | blaze | <u>37</u> | 8.31 |
| | <u></u> | 0.11 | alarm | 47 | 8.57 | a blazing | g fire | |
| assurance | | uiu iiie | fire alarm | _ | | roar | <u>29</u> | 8.02 |
| forest | 46 | 8.02 | station + | 167 | 8.37 | a roaring | g fire | |
| a forest fire | | 0.02 | the fire static | n <u></u> | 0.01 | extinguish | 27 | 7.97 |
| harn | 26 | 7 78 | hazard | 37 | 8 22 | to exting | uish the | fire |
| died in a harn fire | 20 | | a fire hazard | <u>01</u> | 0.22 | burn | <u>37</u> | 7.83 |
| eningr | 22 | 7 79 | ovit | 25 | 7.92 | fire burn | ed | |
| aniper fire | <u>20</u> | 1.10 | the fire cuit | <u>20</u> | 7.62 | fight | 44 | 7.76 |
| sniper nie . | | | the me exit | | | 5 | | |

fire (noun) Alternative PoS: <u>verb</u> (3,176) British National Corpus (BNC) freq = <u>14,172</u> (126.20 per million

Figure 1. An example of word sketch for 'fire'

Originally, the system was designed only for monolingual corpora. However, in due course the need arose and Sketch Engine was subsequently enhanced with features for bilingual and multilingual corpus data. Also, large volumes of the multilingual data have become much more easily available in recent years, e.g. with Europarl, OPUS and EUR-LEX (Baisa, 2016) public data sets, which make the bilingual features increasingly relevant.

Recently, the bilingual and multilingual features have been substantially improved – some functions were newly introduced, others were substantially extended. This paper is a comprehensive description of the bilingual functions of Sketch Engine, many of which are based on the state-of-the-art results in computational linguistics. We believe that these

functions can be extremely useful in the process of creating bilingual learners' dictionaries, as well as in other types of bilingual or multilingual studies.

2. Glossary

In this section we explain several terms that may not be commonly known but are used in this paper.

Parallel corpus (or a set of parallel corpora) is a bilingual or multilingual corpus where the monolingual parts are aligned to each other using some small text units, typically sentences or paragraphs. In other words, the data contain information about which sentence(s) in language 1 corresponds to which sentence(s) in language 2, etc. The alignment does not have to be 1:1 and there may be untranslated segments.

Comparable corpora are not aligned to each other – they do not contain translations of the same texts (at least it is not guaranteed and the eventual translations are not marked up in the data). However, they are comparable in size and types of text contained – e.g. a 100 million corpus of British newspapers is comparable to a 150 million corpus of French newspapers, even if the topics covered by the news are different in most cases.

Source and target language – for us the distinction between the source and target language is rather weak compared to what it means in the field of bilingual learners' dictionaries – source language is simply the language that we start from when formulating the query to the system, and target language are the results from the other language, obtained in a semi-automatic way. However, the direction itself mostly corresponds to the notions of source and target languages as used in bilingual learners' dictionaries.

3. Bilingual concordance

A bilingual concordance allows searching in aligned segments of parallel corpora according to various conditions and displaying the results side by side.

3.1 Bilingual queries in corpus query language

Technically, a parallel corpus in Sketch Engine is a set of two or more monolingual corpora that are only linked together by the meta-data (e.g. sentence identifiers). The corpus query language (CQL, Jakubíček 2010) used in Sketch Engine² contains an operator for binding the two (or more) parts: the *within* operator. The syntax is:

<query1> within <corpus2>:<query2>

where <corpus2> is an identifier of the aligned corpus for the target language, and <query1>, <query2> are any arbitrary CQL queries. For example, the following query on the German part of the Europarl corpus:

[lemma="Katze"] within europarl7_en: [lemma="cat"]

returns all occurences of 'Katze' in the German part where the corresponding aligned segment contains 'cat' – in most cases, this will mean 'Katze' translated as 'cat'.

It is also possible to look for the complement, by the operator *!within* ('not within'):

[lemma="Katze"] !within europarl7_en: [lemma="cat"]

This will find such occurences of 'Katze' where the corresponding aligned segment *does not* contain 'cat' – in other words, cases where 'Katze' is not translated as 'cat'.

By these operators, it is possible to join also 3 or even more corpora.

3.2 User-friendly interface

Writing such complex CQL queries is not user-friendly. Therefore, we have enhanced the basic Sketch Engine concordance form by full query forms for each of the aligned parts of the corpus, which enables users to query each part separately and then link them together (by 'within' or 'not within'). There is an option not to input any query on the target language part(s) – then we just show the aligned segments from the target language without constraints. For this case, there is a check-box for filtering occurrences where the translation is not empty. The form also contains a selection from the languages available, and the query form appears only after selecting a language. The form is illustrated in Figure 2.

The resulting concordance is then displayed side by side, as illustrated in Figure 3. Users can switch between the languages and perform various concordance operations (such as filtering or sorting) on both parts of the concordance.

| Simple query: cat Ma | ke Concordance |
|-------------------------------------------|----------------|
| Query types Context Text types | |
| Parallel query | |
| EUROPARL7, Bulgarian (europarl7_bg) | |
| EUROPARL7, Czech (europarl7_cs) | |
| EUROPARL7, Danish (europarl7_da) | |
| EUROPARL7, Dutch (europarl7_nl) | |
| EUROPARL7, Estonian (europarl7_et) | |
| EUROPARL7, Finnish (europarl7_fi_1) | |
| EUROPARL7, French (europarl7_fr) | |
| EUROPARL7, German (europarl7_de) Contains | |
| Simple query: Katze Does NOT contain | Query types |
| filter out empty lines | |
| EUROPARL7, Greek (europarl7_el) | |

Figure 2. Parallel corpus concordance form – searching for English 'cat' translated as German 'Katze'

| Query cat, Katze 119 (1.96 per million) (1) | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Page 1 of 6 Go <u>Next</u> <u>Last</u> | |
| EUROPARL7, English | EUROPARL7, German |
| Incidentally, it is ironic that the problem is much less serious for the feedingstuffs of our dogs, <u>cats</u> and other domestic animals. | Am Rande sei erwähnt, daß dieses Problem unglaublicherweise sich in weitaus geringerem Maße bei den Futtermitteln für unsere Hunde, <u>Katzen</u> und anderen Haustiere stellt. |
| Please say something about the risk of infection from <u>cats</u> . | Sagen Sie bitte dazu etwas zur Gefahr, von $\underline{\text{Katzen}}$ infiziert zu werden. |
| Mr President, Mrs Roth-Behrendt, Mr Böge, Commissioner, we have before us a proposal for a regulation on the prevention of transmissible spongiform encephalopathies, which is to say degeneration of the brain in bovine animals, sheep, primates, antelope, <u>cats</u> and even humans, since there have been 53 cases of atypical Creutzfeldt-Jakob in Great Britain, all fatal. | Wir befassen uns mit dem Vorschlag für eine Verordnung zur Verhütung der transmissiblen spongiformen Enzephalopathien, d. h. einer Form der Degeneration des Gehirns bei Rindern, Schafen, Affen, Antilopen, <u>Katzen</u> und auch beim Menschen, denn es hat ja in Großbritannien 53 atypische Creutzfeldt- Jakob-Fälle gegeben, die alle tödlich verlaufen sind. |
| l identify <u>cats</u> here in particular. | Katzen möchte ich hier besonders hervorheben. |
| That is the advice I have been given in relation to \underline{cats} . | Das sind die Informationen, die ich im Zusammenhang mit <u>Katzen</u> erhalten habe. |
| However, I cannot accept the exemption for feeding <u>cats</u> with ruminant protein from the ban. | Der Ausnahmeregelung für die Verfütterung von aus Wiederkäuergewebe gewonnenem Protein an <u>Katzen</u> kann ich jedoch nicht zustimmen. |

Figure 3. English-German parallel concordance in Sketch Engine – 'cat' translated as 'Katze'

4. Statistical translation dictionaries

Sentence alignment in parallel corpora enables us to compute various statistics over the number of aligned pairs, and to quantify the probability (or other metric) that word X translates to word Y, for each pair of words in the corpus. The procedure is similar to training a translation model in statistical machine translation (Och and Ney 2003). Our implementation uses the logDice association score (Rychlý 2008), the same measure that is used in scoring collocational strength in word sketches, depending on:

- the frequency of co-occurrence of the two words (e.g. 'Katze' and 'cat') the higher this frequency, the higher the resulting score
- standalone frequencies of the two words the higher these frequencies, the lower the resulting score

By computing these scores for all word pairs across the corpus, we are able to list the strongest 'translation candidates' for each word, according to the score. For our purposes, we store the top 10 candidates for each word, as a statistical translation dictionary.

The procedure is computationally demanding (quadratic to the number of types) and we exploit an algorithm for computing bi-grams to make it feasible even for very large corpora which is part of the Manatee indexing system used in Sketch Engine.

4.1 *Highlighting translation candidates*

The statistical translation dictionaries enable us to highlight translation candidates in the parallel concordance, even if no query is explicitly given, as illustrated in Figure 4. If we find any translation of the source word (lemma) in the target segment, we highlight the particular

target word. Of course, it is not 100% successful (the actual average accuracy is around 85% and there are big differences between different language pairs), but it streamlines the process of going through the concordance results significantly. The function can be also roughly described as 'show me what X might translate as'.

| Query cat 252 > Filter by aligned corpus 236 (3.89 per million) (1) | | | | | | | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|--|--|--|--|--|
| Page 1 of 12 Go <u>Next Last</u> | | | | | | | |
| EUROPARL7, English | EUROPARL7, German | | | | | | |
| We all oppose social dumping when perpetrated by fat <u>cat</u> industry, we must also oppose it when perpetrated by governments. | Wir alle verurteilen Sozialdumping, wenn es von der Wirtschaft ausgeht, wir müssen es aber auch verurteilen, wenn es vom Staat ausgeht. | | | | | | |
| Incidentally, it is ironic that the problem is much less serious for the feedingstuffs of our dogs, <u>cats</u> and other domestic animals. | Am Rande sei erwähnt, daß dieses Problem unglaublicherweise sich in weitaus geringerem Maße bei den Futtermitteln für unsere Hunde, Katzen und anderen Haustiere stellt. | | | | | | |
| This applies in particular to proposed Amendment No 50, which includes <u>cat</u> food in the proposal tabled by Mr Böge, who I know takes a very careful approach to everything. | Das ist insbesondere der Änderungsantrag 50, der Katzenfutter mit aufnimmt in den Vorschlag von Herrn Böge, von dem ich genau weiß, daß er sehr sorgfältig mit allem umgeht. | | | | | | |
| But you have said that you have a problem with <u>cat</u> food. | Aber ich habe von Ihnen auch gehört, daß Katzenfutter für Sie ein Problem ist. | | | | | | |
| Please say something about the risk of infection from <u>cats</u> . | Sagen Sie bitte dazu etwas zur Gefahr, von Katzen infiziert zu werden. | | | | | | |
| Mr President, Mrs Roth-Behrendt, Mr Böge, Commissioner, we have before us a proposal for a regulation on the prevention of transmissible spongiform encephalopathies, which is to say degeneration of the brain in bovine animals, sheep, primates, antelope, <u>cats</u> and even humans, since there have been 53 cases of atypical Creutzfeldt-Jakob in Great Britain, all fatal. | Herr Präsident, Frau Berichterstatterin, Herr Verfasser der Stellungnahme des mitberatenden Ausschusses und Herr Kommissionsvertreter! Wir befassen uns mit dem Vorschlag für eine Verordnung zur Verhütung der transmissiblen spongiformen Enzephalopathien, d. h. einer Form der Degeneration des Gehirns bei Rindern, Schafen, Affen, Antilopen, Katzen und auch beim Menschen, denn es hat ja in Großbritannien 53 atypische Creutzfeldt-Jakob-Fälle gegeben, die alle tödlich verlaufen sind. | | | | | | |

Figure 4. English-German parallel concordance with highlighted candidate translations in the target language – 'cat' without given translation (the highlights are automatic)

5. Bilingual word sketch

We have been experimenting with bilingual word sketches for quite a long time – mainly because it has not been clear how they should work, and especially, what the users need to see. Several variants of the bilingual word sketch have emerged from these experiments (Kilgarriff 2013), all of which are currently operational, and we describe them in the rest of this section. We start with the simple variants and proceed with the more complex ones (although this was not the chronology in which the features were developed).

Unlike the parallel concordances, most features of bilingual word sketches are available for parallel as well as comparable corpora.

5.1 Bilingual word sketch with aligned relations

When there was no bilingual word sketch feature, some people used two browser windows with word sketches, for two different languages, with two words that were translations of each other, to analyse the mutual relations between the words.

The simplest version of a bilingual word sketch allows users to do this within one window: in the word sketch form, there are new input boxes for target language, target corpus, and the word in the target language. Any target corpus in any language can be selected. The result is a merged word sketch for the two words in different languages – distinguished by colour – with *compatible* relations placed next to each other, as in Figure 5. We explain in Section 5.3 what exactly we mean by *compatible* relations.

This type of bilingual word sketch is referred to as *bilingual manual* (BIM) word sketch in older Sketch Engine materials.

5.2 Translate button

With the automatic statistical dictionary derived from a parallel corpus, we do not need users to input the translation in the target language, but have 10 good candidate translations stored in the dictionary. We can also pre-set the most suitable corpus for each target language. For parallel corpora, it would be the parallel part in the particular language; for non-parallel corpora, we can select the best fitting comparable corpus.

This leads us to a more sophisticated variant of a bilingual word sketch – a 'translate button'. When a user creates an ordinary monolingual word sketch, we offer them to 'translate' the word sketch into one of the available languages – see the highlighted menu in the left panel in Figure 5.

The function finds the best translation in our statistical dictionary and shows the bilingual word sketch with this word, using the pre-configured parallel or comparable corpus. In case this choice is not correct, or the user is interested in a different translation equivalent, we offer them links to the bilingual word sketch for the other candidate translations stored in our dictionary.

5.3 *Compatible relations*

We mentioned (in Section 5.1) the alignment of *compatible* relations in the user interface; how do we decide whether two particular relations are compatible?

This is not an easy question. We cannot expect that all collocations from one relation will translate only to collocations in one particular relation in the target language, as different things are expressed by different means in different languages. On the other hand, *most* of the collocations from a certain relation are often translated using a single equivalent relation in the target language. Our solution is as follows:

- If the relations have identical names, they are aligned to each other (e.g. 'modifier' will be very similar across many languages).
- We have implemented an option for mapping relations to their English (near) equivalents a directive in the word sketch grammar called *UNIMAP. Using this directive, some relations can be manually marked compatible with some English relations. Compatible relations for two languages other than English are then those that are compatible with the same English relation.

Still, it remains a problem. In some cases, the number of aligned relations is small due to principal inconsistencies between the sketch grammars. Generally, the sketch grammars for different languages are rather heterogeneous – they were historically developed by people across the world from very different linguistic environments, and assigning equivalents to a large part of the relations is simply not possible. In the future, we want to invest in the development of compatible sketch grammars (similar to the grammars created by Benko (2014) but with more linguistically descriptive relations) that will contain more compatible relations. The alignment – and therefore the whole bilingual word sketch application – will thus be improved.

| Concordance | decla | rati | on | (noun) EUROPARI | .7, Englis | sh freq = | - <u>5,532</u> dé | cl | ara | ation |
|----------------|---------------------------|-----------------------|----------|----------------------------|-----------------|------------|-----------------------------|------------|--------|-------------------------|
| Word list | Use another can | didate tra | Inslatio | n: <u>écrire</u> <u>si</u> | <u>ner</u> pro | cès-verb | al intention De | éclara | ations | unilatéral in |
| Word sketch | | oo to ooo | 000 10 | oiprocal biling | | | translated college | ations | | |
| Thesaurus | Click off collocat | es lo acc | essie | ciprocai bilingi | iai searci | r or initi | | auons | 2 | |
| Sketch diff | object of | | | <u>objet de</u> | | | subject of | | | <u>sujet de</u> |
| Sketch eval | | <u>2,014</u> | 2.70 | | <u>6,702</u> | 3.50 | | <u>826</u> | 1.90 | |
| СРА | write | <u>118</u> | 9.15 | écrire | <u>1,494</u> | 11.73 | annex | <u>10</u> | 7.87 | écrire |
| Corpus info | Oral quest written dec | ions and larations | (| . Déclara article 14 | tions écri 2 | ites (| declaration to the Treat | anne v | exed | . Déclara article 14 |
| My jobs | submission | n) | (| appeler | 820 | 9.80 | emphasize | 6 | 6.03 | annexer |
| | sign | <u>139</u> | 8.65 | ordre du | iour appe | elle la | accompany | 8 | 5.72 | la déclar. |
| | issue | <u>99</u> | 8.46 | déclaratio | on de la | | state | 17 | 5.33 | prononcer |
| User guide 🗹 | issued a d | eclaratio | n | Commission | | | a declaration stating | | | la déclar. |
| | annex | <u>10</u> | 6.98 | signer | <u>150</u> | 8.71 | that | // 314 | ung | nublier |
| Save | append | <u>8</u> | 6.90 | publier | <u>140</u> | 8.60 | reaffirm | <u>3</u> | 5.28 | une décl. |
| Change options | attach | <u>23</u> | 6.76 | publié un | e déclara | ation | include | <u>48</u> | 5.00 | figurer |
| Cluster | declaration | n attache | d to | saluer | <u>122</u> | 7.83 | see Minute | s Wri | tten | dóclarati |
| Sort by freq | the | | | salue la c | léclaratio | n | declaration | s incl | luded | no |
| Hide gramrels | submit | <u>39</u> | 6.45 | prononcer | <u>65</u> | 7.60 | In the regis | ter | | lo déclor |
| More data | translate | <u>10</u> | 6.37 | déclaratio | on pronoi | ncée | condemn | <u>8</u> | 4.90 | la declari |
| Less data | ring | <u>5</u> | 6.15 | faire | <u>1,015</u> | 7.40 | clarify | <u>5</u> | 4.78 | emaner |
| Translate | publish | <u>17</u> | 6.14 | faire une | déclarati | ion | entitle | <u>3</u> | 4.54 | déclaratio |
| - Bulgarian | contradict | <u>6</u> | 5.97 | annexer | <u>34</u> | 7.27 | reiterate | <u>3</u> | 4.49 | contenir |
| - Czech | • • | | | 1 1/ 1 | | | read | 4 | 4 30 | faire |

Figure 5. English-French bilingual word sketch with aligned relations – 'declaration' translated as 'déclaration'

5.4 Reciprocal bilingual word sketch

When viewing a bilingual word sketch with aligned relations, it is quite typical that some of the collocations are translated by target language words present on the page (mostly within the aligned relation) and some are not – not because of a mistake but because the collocation is expressed by a different word in the target language. In that case, users need a straightforward way to go to the word sketch of the particular problematic collocate.

For example, when viewing an English-Portuguese word sketch for *brown* and *marrom*, there will probably be *rice* on the English side, in the 'modifies' relation, without an equivalent on the Portuguese side. A small gadget in the interface enables the user to click on *rice*, input a Portuguese translation *arroz* and get to the bilingual word sketch of *rice* and *arroz*, which will reveal that *brown rice* is an equivalent of *arroz integral* (i.e. no direct equivalent of *brown* is used at all). An illustration of this function is in Figure 6.

| marro | m, | tTenTe | en11 freq = | rice (noun) enTen12 freq = <u>268089</u> (20.7 | | | | | | |
|--------------|----------------------------|--------|------------------------------------------------|---------------------------------------------------|------|---|-----------|----------------|------|---------------|
| brown en | e q = <u>374425</u> | arroz | DZ ptTenTen11 freq = <u>86646</u> (26.7 | | | | | | | |
| <u>N mod</u> | 7,998 | 0.4 | modifies | <u>223,516</u> | 0.3 | | modifier | <u>100,314</u> | 0.1 | rice N mod AD |
| marrom ADJ | 1,000 | • | rice _ | <u>13,109</u> | 9.28 | | brown | 13,071 | 9.53 | parboilizado |
| coloração | <u>340</u> | 7.49 | Lemma: | arroz | | | fried | 3,480 | 9.31 | cozido |
| anā | <u>71</u> | 7.4 | Lemma. | anozj | | - | basmati | 1,592 | 8.98 | carolino |
| corrimento | <u>58</u> | 6.92 | Query | | | | sticky | 1,651 | 8.06 | integral |
| linhaça | <u>48</u> | 6.9 | trout | 3,425 | 8.26 | | alutinous | 739 | 7.88 | arbóreo |
| borra | <u>42</u> | 6.55 | sugar | 11.161 | 8.21 | | pasta | 1.590 | 7.44 | doce |
| aranha | <u>100</u> | 6.35 | bear | 2,754 | 7.11 | | jasmine | 647 | 7.42 | polido |

Figure 6. Illustration of reciprocal bilingual word sketch – from 'brown rice', we can easily get to 'arroz integral', although 'brown' is not a probable translation of 'integral'

5.5 Bilingual word sketch with aligned collocations

The automatic statistical dictionary can be used not only for translating the headword but also for translating the collocations. By doing so, we can shift the automatic analysis yet a bit further – the user does not have to search for target language translations; the system will search them itself (in the easy cases).

We align the collocates across grammatical relations, because the correct translation does not have to be within the same relation; we also provide one example for each of the collocations in both source and target languages.

This feature is available for parallel as well as for comparable corpora. In case of parallel corpora, we show only examples that are aligned to each other, and we hide collocation translations that have no such example, even if they are present in the data. The feature is available under the 'translated collocations' link visible in Figure 5. The resulting bilingual word sketch with translated collocations is then illustrated in Figure 7.

declaration (noun) EUROPARL7, English déclaration

Use another translation: <u>écrire</u> <u>signer</u> <u>procès-verbal</u> <u>intention</u> <u>Déclarations</u> <u>unilatéral</u> <u>indépend</u>

| object_of | | |
|------------|------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| write | <u>118</u> | |
| écrire | <u>113</u> | Subject : Remembrance of the Holocaust Following up the Swedish Prime Minister G arranging the well-received Stockholm International Forum on the Holocaust in Janua written declaration on 7 July 2000 on the remembrance of the Holocaust . Objet : Souvenir de l' Holocauste Suite à l' initiative prise par le premier ministre suéd en janvier 2000 l' important Forum international de Stockholm sur l' Holocauste , le P juillet 2000 , adopté une déclaration écrite sur le souvenir de l' Holocauste . |
| sign | <u>139</u> | |
| signer | <u>70</u> | Therefore , Madam President , in addition to calling upon you to make authoritative re Government , I also invite all the Members of Parliament to sign the declaration call dedicated to the Shoah , to the holocaust , to be observed throughout the European C' est la raison pour laquelle , Madame la Présidente , je vous demande d' intervenir slovène et invite tous les députés à signer la déclaration qui demande l' institution d Shoah , à l' Holocauste , au nom de toute l' Union européenne . |
| convention | <u>1</u> | More and more countries are signing fine declarations and international conventions complied with . De belles déclarations , des conventions internationales - de plus en plus de pays le qu' ils ne les appliquent pas . |
| annex | <u>10</u> | |
| annexer | <u>1</u> | Parliament 's future involvement in strategy documents is now covered by two decl a Interinstitutional Agreement . La participation future du Parlement dans les documents stratégiques est maintenan déclarations annexées à l accord interinstitutionnel . |
| Eiguro 7 I | Englic | b French bilingual word skatch with aligned collocations (declaration) |

Figure 7. English-French bilingual word sketch with aligned collocations – 'declaration' translated as 'déclaration'

6. Bilingual terminology extraction

This section describes how Sketch Engine automatically extracts translation candidates for terms from a specialized parallel corpus. The feature can be used straightforwardly for compiling a list of entries for a specialized bilingual dictionary and for automatic drafting of translation candidates.

6.1 Terminology from a monolingual corpus

Sketch Engine has a method for automatic extraction of term candidates from a specialized corpus (Kilgarriff et al. 2014). In short, it uses queries in corpus query language to describe what a term may look like (using particular words and morphological tagging – mostly, term candidates are particular types of noun phrases). Candidates gained by this process are further evaluated using a score expressing how special the word is for the particular type, in comparison with a general corpus (Kilgarriff 2009). The basic formula is

score = (freq_per_million + 1) / (ref_freq_per_million + 1)

where *freq_per_million* is per-million relative frequency of the candidate in the specialized corpus, and *ref_freq_per_million* is the same frequency in a general corpus of the given

language. The candidates are then sorted according to this score: the higher the score, the more specialized the term is in the particular domain.

6.2 Alignment of terms

The process of drafting the statistical dictionary described in Section 4 can be extended to work with term candidates instead of words. For each sentence, we look at the term candidates extracted from it, and for each pair of term candidates T1 and T2, we compute their co-occurrence score in the same way as described in Section 4.

The results can be sorted either by their association score (so that most salient translation candidates appear at the top), or by their term score in the source language (so that the most specific terms for the particular domain are on the top). A sample output is shown in Figure 8.

| L1 term | L2 term | Co-freq | L1 freq | L2 freq |
|----------------------|-----------------------|------------|------------|------------|
| mobility | mobilité | <u>23</u> | <u>27</u> | <u>25</u> |
| retention | rétention | <u>49</u> | <u>60</u> | <u>51</u> |
| consent | consentement | <u>49</u> | <u>58</u> | <u>53</u> |
| analysis | analyse | <u>410</u> | <u>467</u> | <u>464</u> |
| secondary school age | pourcentage du nombre | <u>11</u> | <u>12</u> | <u>13</u> |
| climate | changement climatique | <u>102</u> | <u>122</u> | <u>110</u> |
| hydration | hydratation | <u>29</u> | <u>36</u> | <u>30</u> |
| chain | chaîne | <u>29</u> | <u>34</u> | <u>32</u> |
| methodology | méthodologie | <u>49</u> | <u>58</u> | <u>54</u> |
| non-discrimination | non-discrimination | <u>14</u> | <u>16</u> | <u>16</u> |
| young woman | jeune femme | 7 | <u>8</u> | <u>8</u> |
| nutrition | nutrition | <u>263</u> | <u>307</u> | <u>295</u> |
| litre | litre | <u>31</u> | <u>35</u> | <u>36</u> |
| solid waste | déchets solides | <u>24</u> | <u>25</u> | <u>30</u> |

Figure 8. An example of the bilingual terminology extraction feature – a parallel corpus of UNICEF texts

7. User parallel corpora from TMX

Although there had been some support for creating user parallel corpora before, it was technically quite a complicated process: users needed to prepare two corpora with special meta-data marking the alignment, so the alignment needed to be exactly 1:1 (any mistake would break the whole alignment) and they needed to link the two corpora together using complicated settings in the corpus configuration file.³ Therefore, the support for parallel user corpora was not really useful for most users.

Within the corpus building component of Sketch Engine, the so-called Corpus Architect, we have significantly streamlined the process, allowing to create all the components of a parallel corpus at once, from one file, with a few clicks. The file format is TMX (Translation Memory eXchange format)⁴ and it is used on a daily basis by translators and systems for computer-aided translations. There are a lot of datasets (in fact, parallel corpora) stored in the TMX format, or easily convertible to it, but a large part are under intellectual property protection (for instance, when translation companies do not want to share their private data with the public). This new feature makes possible very easy conversion of the TMX files to a parallel corpus, which enables all the previously described functions instantly, e.g. specialized word sketch using the particular data, or terminology extraction, including drafting translation candidates for future translations from the same domain. When the data are proprietary, at least its owner can benefit from the functionality (user corpora in Sketch Engine are private, unless shared by their owner).

Technically it works in the following way: the system parses the TMX file according to specifications, finds the aligned segments, splits them into two (or more) aligned corpora and creates the 1:1 alignment at the same time.

8. M:N alignment

Last but not least, unlike before, Sketch Engine (in particular the Manatee library) supports alignments of parallel corpora that are not strictly 1:1, so it is possible to express that 'these three sentences are translated as this one sentence' without having additional 1:1 alignment meta-data. This is especially useful for multilingual parallel data where the granularity of alignment between one pair of languages is different from the granularity of another pair, so creating a 1:1 helper alignment meta-data is not even possible. Also, it allows including non-translated segments (i.e. any of M, N can be 0).

However, setting up such a parallel corpus is still technically rather complicated: users have to create a mapping file using sentence identifiers and link this file in the configuration.³ Therefore, this machinery is used for preloaded corpora in Sketch Engine rather than for user corpora (although that is also possible). On the other hand, a typical use case for a user parallel corpus is a bilingual (rather than multilingual) corpus where 1:1 alignment is mostly sufficient, so the current state may not be a problem.

9. Conclusion

The corpus query system Sketch Engine used to be primarily a monolingual tool, despite its capability to process many languages. This is no longer true.

We have recently introduced significant enhancements in relation to bilingual lexicography and working with bilingual corpora. Above, we have described principles that the features rely on and illustrated them with examples. We believe these improvements will help lexicographers to streamline their work on bilingual learner's dictionaries, and thus contribute to their quality.

Future directions include improving all the above-mentioned functions according to feedback from users. We have mentioned some possible development directions earlier in the text while describing the particular functions. There are also options to enhance the automatic alignment methods as well as the visual part of the application in order to provide better data to the users in a more comprehensible way.

Acknowledgements

This work has been partly supported by the Grant Agency of Czech Republic within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

Notes

- ¹ www.sketchengine.co.uk
- ² www.sketchengine.co.uk/corpus-querying
- ³ www.sketchengine.co.uk/setting-up-parallel-corpora
- ⁴ en.wikipedia.org/wiki/Translation_Memory_eXchange

References

- Baisa, V., Cukr, M. and Ulipová, B. 2015. Bilingual terminology extraction in Sketch Engine. In 9th Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2015), 65-70.
- Baisa V., Michelfeit J., Medved' M. and Jakubíček M. 2016. European Union Language Resources in Sketch Engine. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016).*
- Benko, V. 2014. Compatible Sketch Grammars for Comparable Corpora. In *Proceedings of the XVI EURALEX International Congress*. Bolzano: EURAC research, 417-430.
- Jakubíček, M., Kilgarriff, A., McCarthy, D. and Rychlý, P. 2010. Fast Syntactic Searching in Very Large Corpora for Many Languages. In *PACLIC*, 24, p741-747.
- **Kilgarriff, A. 2009.** Simple Maths for Keywords. In *Proceedings of Corpus Linguistics Conference CL2009.* University of Liverpool.
- **Kilgarriff, A. Kovář, V. and Rychlý, P. 2010.** Tickbox Lexicography. In *eLexicography in the 21st century: New challenges, new applications*. Brussels: Presses universitaires de Louvain, 411-418.
- Kilgarriff, A., Kovář, V. and Frankenberg-Garcia, A. 2013. Bilingual word sketches: three flavours. In *Electronic Lexicography in the 21st Century: Thinking outside The Paper (eLex 2013)*, 17-19.
- Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P. and Suchomel., V. 2014. Finding Terms in Corpora for Many Languages with the Sketch Engine. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: The Association for Computational Linguistics, 53-56.
- Kilgarriff, A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P. and Suchomel V. 2014. The Sketch Engine: ten years on. *Lexicography*, 1.1: 7-36.
- **Och, F.J. and Ney, H. 2003.** A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29.1: 19-51.
- Rychlý, P. 2007. Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2007), 65-70.
- **Rychlý, P. 2008.** A Lexicographer-Friendly Association Score. In *2nd Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2008).* Brno: Masarykova Univerzita, 6-9.
- Rychlý, P., Husák, M., Kilgarriff, A., Rundell, M. and McAdam, K. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the*

XIII EURALEX International Congress. Barcelona: Institut Universitari de Lingüística Aplicada, 425-432.