Bilingual Word Sketches: the *translate* Button

Vít Baisa, Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý
Lexical Computing Ltd, UK; Faculty of Informatics, Masaryk University, Czech Republic
{xbaisa,jak,xkovar3,pary}@fi.muni.cz, adam@lexmasterclass.com

Abstract

We present bilingual word sketches: automatic, corpus based summaries of the grammatical and collocational behaviour of a word in one language and its translation equivalent in another. We explore, with examples, various ways that this can be done, using parallel corpora, comparable corpora and bilingual dictionaries. We present the formalism for specifying equivalences between gramrels in the two languages. We show how bilingual word sketches can be useful for dictionary-making and we present additional functionality to make them more useful. We state the language pairs for which bilingual word sketches are currently available, and our plans for adding more pairs.
Keywords: word sketch; corpus lexicography; lexical computing

1 Introduction

Word sketches are one-page, automatic corpus-based accounts of a word's grammatical and collocational behaviour (Kilgarriff et al 2004). Since their introduction in 1998 they have come to be widely used in lexicography, often serving as the first port of call for a lexicographer analysing a word. Until recently, they have been monolingual. Bilingual lexicographers would like to see the word and its grammar and collocations, matched up with its translation and its grammar and collocations. In this paper we discuss how this might be done, and present the solution we have adopted and now make available within the Sketch Engine.

Two open questions are:

- how should the source word's translation be identified?
- how should the grammar and collocations be matched up?

We first describe three responses, based on different answers to the first question: *bics, bips* and *bims.* We then describe how we amalgamate all three to give a single, easy-to-use ***translate*** function, with a report as illustrated in Figure 1.

Figure 1: Bilingual word sketch for fire/Feuer. The user can click on alternative translations to see alternative sketches.

In all the approaches discussed, we start from the word sketch for one language (hereafter L1) and augment it with information from the other language (L2). Note that, here, L1 and L2 are neither 'source' and 'target' as understood by translators, nor 'mother tongue' and 'language being learnt' as in the language learning literature. They simply reflect the fact that the user (and algorithm) starts from one language and adds information from another. At some point we may develop symmetrical, direction-independent bilingual word sketches but we have not done so yet.

2 BIPs

Bips are bilingual word sketches based on **p**arallel corpora. A dictionary is not needed because the connections between the languages can be inferred, by looking to see which <L1, L2> pairs of words are frequently found in aligned chunks <L1, L2> chunks (where the chunks are usually sentences). We first count occurrences in aligned chunks for all <L1, L2> word pairs, and then use the Dice coefficient to identify candidate translations.

Where a lemmatiser is available for the language, the corpus is first lemmatised and the dictionary-induction process is applied to lemmas rather than word forms.

This provides a bilingual dictionary, with each lemma in each language having a list of candidate translations, with confidence scores. The default setting is that the ten top candidates for each lemma are retained.

Similar methods are used in GIZA++ (Och and Ney, 2000) and other tools for statistical machine translation, though usually applying to word forms rather than lemmas, and with different statistics and objectives.

Once the blingual dictionary is in place, the algorithm for creating the bip sketch is as follows:

(1) the user inputs an L1 headword.

(2) take the set of L1 collocates[1] from the L1 word sketch and translate them using the bilingual dictionary.

(3) take the top translation candidate for the L1 headword: call it the L2 headword

(4) take the set of collocates from the word sketch of the L2 headword

(5) perform an intersection between the translations-of-collocates from step 2 and the collocates-of-translation from step 4

(6) for each item in the intersection,
- is there at least one pair of aligned chunks in the parallel corpus where
  - the L1 collocation occurs in the L1 chunk, and
  - the L2 collocation occurs in the L2 chunk?
- if yes
  - present as a translation candidate for the L1 collocation,
  - illustrate with the aligned chunk in the two languages

(7) For L1 collocates with no translation candidates in the intersection, or where there were items in the intersection but there were no instances of corresponding collocations in aligned chunks, present the L1 collocate monolingually, without any candidate translations.

A bilingual sketch produced using this method, also showing the aligned chunks with both L1 and L2 collocations, is shown in Figure 2.

Figure 2: Bip word sketch for *declaration*/*déclaration*.

1 Our terminology here is that a *collocation* comprises a *headword* and a *collocate* (in a specific grammatical relation).

The 'intersection' method follows Grefenstette (1999): to find translations for compositional collocations like English *work group* (into, e. g., French) he looked up *work* and *group* in an English-French dictionary, where he found three translations for *work,* five for *group.* That gives 3 x 5 = 15 possible combinations. He then checked to see which was commonest in a French corpus, and presented that as the candidate translation. This core method is explored in much comparable corpus work (see Sharoff et al 2013 for the state of the art).

Linguee[2] provides users with some similar functionality, with aligned bilingual concordance data together with the dictionary translations of the search and some matched pairs of collocations.

2.1 In Praise of EUROPARL

A central constraint on methods using parallel corpora is the quality, genre, size and availability of parallel data for each language pair. For the 22 official EU languages[3] and corresponding 253 language pairs, we are fortunate: the EUROPARL corpora are large, contain professional quality translations, and are of a text type - European parliamentary speeches - which, while far from perfect for general-language lexicography, is far more general than the language of, for example, software manuals or patient information leaflets, two other domains where parallel data is available in bulk. Moreover the EUROPARL corpora have been prepared and made ready for language technology use (Koehn 2005). We are currently only exploring parallel-corpus methods for EU language pairs, for this reason.

3 BICs

Bics are bilingual word sketches based on **c**omparable corpora. They require a bilingual dictionary, as well as two comparable corpora, as input. Our first attempts at bilingual word sketches took a comparable-corpus approach, with dictionaries from publishers (Kilgarriff et al 2011). Our conclusion was that this left us too dependent on dictionaries from publishers, which were highly variable in availability and licence terms, not to mention format, size, quality and lexicographic approach. The approach was unviable for extending to multiple language pairs.

Now that we can build bilingual dictionaries for all EU-language pairs, these dictionaries can be used as free-standing resources for building bic sketches**.** The algorithm again uses the intersection method, and is as presented above for bips, except that step 6 is not available. Wherever an L2 headword's collocate is a translation of an L1 collocate, it is presented as a candidate translation, as shown in Figure 3.

2 http://www.linguee.com

3 Excluding Irish: irish ahs a distinct status to the other 22 languages and there is far less data available.

Figure 3: Bic word sketch for *declaration*/*déclaration*.

4 BIMs

Bims are bilingual word sketches based on **m**anual selection of headwords. In this approach the user chooses the two words, usually translation equivalents from two different languages, whose word sketches they want to compare, and the corpora to be used. The two word sketches are spliced together. This takes forward something that bilingual lexicographers have been doing since word sketches were first developed: opening two browser windows side by side, with one word sketch in each.
A bilingual sketch for English *brown* and Portuguese *marrom* is shown in Figure 4.

Figure 4: Bim word sketch for *marrom*/*brown*.

4.1 Alignment of grammatical relations

The lexicographer would like to see collocations and their translation equivalents aligned. This is possible to some extent, and is attempted in bip and bic sketches, but is difficult and error-prone. Also the lexicographer would often like more control, and finds it straightforward to match, eg, brown *leather* and *couro* marrom where they are in columns next to each other. So in bim sketches we set ourselves the more limited ambition of matching up columns, so that collocates for corresponding grammatical relations are shown next to each other. (In word sketches, 'grammatical relations' or 'gramrels' are the relations such as *object*, *object_of*, *modifier* that are specified in the sketch grammar and are showed at the heads of the columns in a word sketch.)

Where the gramrels have the same names for two languages, this is trivial: we simply show same-name gramrels next to each other. However sketch grammars for different languages are usually prepared independently of each other; the grammar of different languages is different, underlying part-of-speech taggers may use different conceptualisations and word classes; and gramrels will often be given names in the language that the word sketch is for: the French equivalent of *object_of* is called *objet_de*. So matching gramrels with identical names is not the standard case. We need a mapping between the gramrels of the two languages.

A mapping for each set of names to some 'master' set is preferable to a different mapping for each pair. In our current setting, we use the English names as the master. Both 1:1 and m:n mappings are possible.

The mappings are defined in the sketch grammar using a newly introduced processing directive *UNIMAP. The following definition from a French sketch grammar

```
*DUAL
=objet/objet_de
*UNIMAP object/object_of
...
```

says that *objet* should be joined with *object* and *objet_de* should be joined with *object_of* (or the gramrels paired with English *object* and *object_of* in other languages). The algorithm for finding a target language (TL) gramrel to display next to a source language gramrel X is:

- if there is one or more TL gramrel with a UNIMAP value matching the UNIMAP value of X, select that one/them
- else if there is a TL gramrel of the same name, select that one
- else, nothing is aligned with X.

Left-over, unaligned TL gramrels are shown after SL and aligned gramrels.

4.2 An inline form for finding missing items

For a user, some matching pairs are immediately evident from the bim sketch (*brown leather, couro marrom*) but others are not. We do not find anything equivalent to *brown rice* on the Portuguese side of Figure 5. The user then wants to know "how do you say brown rice in Portuguese?" To meet this need, a new function has been added: the user may click on *rice* to reveal a text-input box where they can input the missing target-language equivalent (here, *arroz*). A new bim word sketch appears, as illustrated in the same Figure 5. This feature helps the user find the missing translation: here, *arroz integral* for *brown rice*.

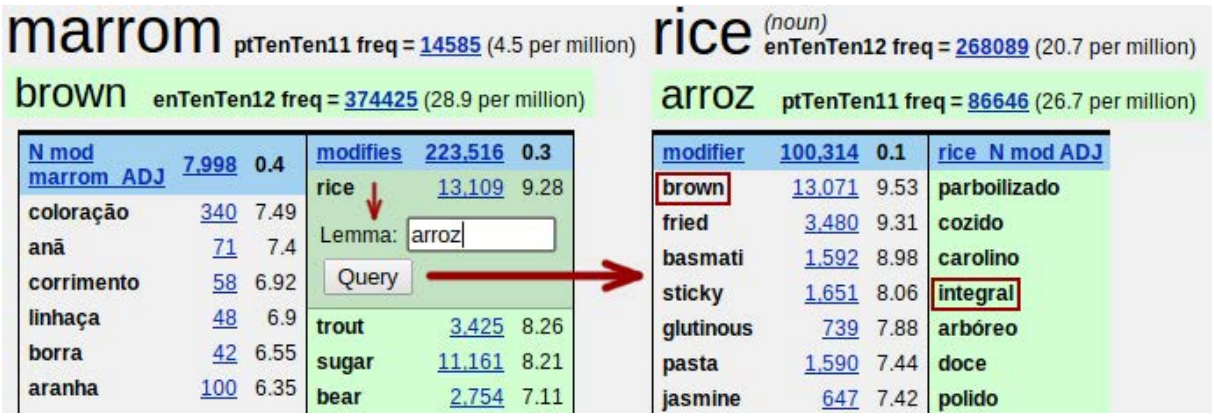

Figure 5: "Finding the missing translation" functionality.

5 Observations

Bics, bips and bims were developed in 2013.[4] In the course of presenting and beginning to use them, we made several observations:

- The bilingual dictionaries created from EUROPARL were of good quality. Most of the time, the top candidate translation was valid.
- For most collocations on the L1 word sketch, we did not find any strong candidates for L2 translation equivalents. The *declaration* examples above were selected because, there, the algorithm did find a translation candidate for many of the L1 collocations. More often, there were very few translations offered. This affected both bics and bips.
- Bims often worked well, but the access method did not: users were required to select, first, the L2; then, the L2 corpus from which the L2 word sketch should be drawn, and then, the L2 lemma. Unless they made all the right choices, they were going to be disappointed.

4 They were presented at the e-Lexicography conference in Tallinn in October 2013.

6 The *translate* button

Most users just want to select the target language, and do not want to think about 'which corpus'. The system developers are the people who know which corpus has the best word sketches for a language, so they should make that choice.

Many users would also rather not have to think of, and input, the L2 lemma. For all EU language pairs, the EUROPARL dictionary offers candidate translations, so here again, more can be done by the system leaving less work for the user. The user can be offered a BIM sketch in which the L2 word is the top candidate from the EUROPARL bilingual dictionary (with second and other candidates also offered.)

With these choices made, we can add a 'translate' button to the word sketch. The 'translate' button gives the user a choice of languages. The languages that the user can choose between are all of those where there is:

- a bilingual dictionary between L1 and the language
- high-quality word sketches for the language
- as at April 2014, there are twenty such languages
- a UNIMAP mapping

Then, when the user has selected the L2, they see the bilingual sketch directly, as in Figures 1 and 6:

amore *(noun)* itTenTen10 freq = 634049 (206.1 per million)

Liebe *(noun)* deTenTen10 freq = 294973 (103.7 per million)

n_modifier	3,656	0.6	modifier	281,129	0.5	preN_V	120,341	2.0	VerbY+Subst...	24,736	0.5
sponsale	129	9.84	wahr	2,537	6.76	giurare	430	6.71	gestehen	750	7.54
lesbico	116	9.39	bedingungslos	583	5.93	dichiarare	1,080	6.7	erwidern	139	6.1
oblativo	41	8.46	ewig	870	5.86	donare	573	6.66	besingen	48	5.38
karmico	34	8.0	dein	3,017	5.71	chiamare	1,754	6.62	entdecken	836	5.03
leopardiano	20	7.15	unerfüllt	465	5.7	Un	1,160	6.49	beteuern	43	4.87
lesbo	16	6.7	unglücklich	539	5.69	cantare	601	6.49	schwören	59	4.45
lesbici	10	6.29	innig	468	5.65	nascere	1,388	6.29	schenken	194	4.27
fato	31	6.21	verboten	490	5.58	provare	901	6.27	entgegenbringen	24	4.22
anoressico	9	5.87	romantisch	552	5.45	vivere	1,643	6.21	wünschen	634	4.22
cannibale	13	5.85	heimlich	476	5.4	confessare	311	6.18	symbolisieren	45	4.14
conta	23	5.7	groß	12,073	5.38	ricambiare	277	6.13	empfinden	200	4.12
cinefilo	9	5.18	leidenschaftlich	423	5.35	sbocciare	260	6.08	predigen	38	4.09
court	12	5.05	viel	6,631	5.33	trasmettere	584	6.05	grüßen	85	4.08

Figure 6: Bilingual word sketch for *amore*/*Liebe*.

The functionality is currently available for all combinations of English, French, German, Italian and Spanish.

7 Current and future work

- **Aligning collocates within BIMs**

  A feature of bics and bips that bims were lacking, was the alignment of translation-equivalent collocations. Although, often, no alignments were found, where an alignment was found, it was usually valid and helpful. So, we can further enrich bim-style sketches, by re-ordering the collocates in the L2 word sketch table so that matched L1, L2 pairs are next to each other.

This is currently in progress.

- **More languages**

  For monolingual or bilingual word sketches to work well, there are a number of prerequisites; first a very large corpus, then processing tools including a tokeniser, lemmatiser, part-of-speech tagger and sketch grammar. All components are currently in place for all major world languages, all EU languages, and a number of others; languages where we intend to get all components working well in the near future include Icelandic, Malay, Bahasa Indonesia and Burmese.[5]

- **More bilingual dictionaries and language pairs**

  We are looking into parallel resources for other language pairs including English and the non-EU major world languages, in particular Arabic-English, Chinese-English, Japanese-English, Russian-English.

- **Evaluating EUROPARL-based bilingual dictionaries**

  For Czech-English we are currently assessing our induced dictionary by comparison with a publisher's dictionary. We shall also compare the Dice algorithm with the Giza++ algorithm for dictionary induction. We shall then extend the evaluation to other language pairs.

8 References

Grefenstette, G. 1(999). The World Wide Web as a Resource for Example-Based Machine Translation Tasks. *Proceedings of Aslib Conference on Translating and the Computer 21.* London.

Kilgarriff, A., Avinesh, P. V. S., & Pomikálek, J. (2011). Comparable Corpora BootCaT. *Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of eLex*, 122-128.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. *Proc. Euralex.* Lorient, France.

Koehn P. (2005). EuroParl: A Parallel Corpus for Statistical Machine Translation. *Proc. Machine Translation Summit*.

Och, F. & Ney, H. (2000).. Improved Statistical Alignment Models. *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 440-447, Hong Kong, China.

Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P., editors (2013). *Building and Using Comparable Corpora.* Springer.

5 In most cases the work has been a collaboration with linguists of the language in question.