Treq: A Translation Equivalents Database (demo)

Michal Škrabal, Martin Vavřín

[Institute of the Czech National Corpus](http://ucnk.ff.cuni.cz), Charles University in Prague, Czech Republic

E-mail: michal.skrabal@ff.cuni.cz, martin.vavrin@ff.cuni.cz

1. Introduction

The aim of this paper is to introduce one of the tools developed at the Institute of the Czech National Corpus (ICNC), which stands halfway between corpora and dictionaries: namely the Treq translation equivalents database. It is based on data stored in the InterCorp parallel corpus (always its latest version, currently v9).

1. InterCorp

InterCorp is a large parallel synchronic corpus under continuous construction at the ICNC since 2005. In the recent past, the corpus has been growing systematically every year, and since 2013 (version 6), even obsolete versions of the corpus are available indefinitely via our corpus query interface, KonText, in order to preserve the possibility to replicate previous research. InterCorp is composed of several parts, the most important of which is arguably the so-called *core –* literary texts with manually corrected OCR and sentence alignment. In addition to the core, there are several *collections*, consisting of texts which were processed only automatically, not manually. These include the following types of texts:

* journalistic articles and news published by [Project Syndicate](http://www.project-syndicate.org/) and [VoxEurop](http://www.voxeurop.eu);
* legal texts of the European Union from the [Acquis Communautaire](http://langtech.jrc.it/JRC-Acquis.html) corpus;
* proceedings of the European Parliament dated 2007–2011 from the [Europarl](http://www.statmt.org/europarl/) corpus;
* film subtitles from the [Open Subtitles](http://www.opensubtitles.org/) database.

InterCorp v9 contains, besides Czech as the pivot language (for every text in InterCorp, there *must* be a single Czech version, either the original or a translation), another 39 languages ​​that are, however, unevenly represented. Therefore, you can find languages which have up to 31 million running words in the core (German) and corpora of individual languages can range up to 120 million running words in size (English), but there are also corpora which have no texts in the core (i.e. no manually processed texts) at all and restrict themselves to collections (e.g. Vietnamese or Icelandic)[[1]](#footnote-1).

1. Data preparation

First of all, when preparing data for Treq, only sentences that are aligned 1:1 are selected from the entire corpus. We restrict ourselves to this simple alignment because it tends to be more reliable; especially in the case of automatically aligned texts, potential errors can be prevented. The next step is to perform an automatic word-to-word alignment using GIZA++ (Och & Ney 2003)[[2]](#footnote-2).

The current version of Treq aligns parallel texts using a method called *intersection*, thus creating only such alignments where one word in the source language corresponds to one word in the target language, e.g.:

0-0 1-2 2-3 4-4 5-5 7-6 8-9 10-10 11-11

which means that the first word in the source language (0) corresponds to the first word in the target language (0), the second word (1) corresponds to the third one (2) etc.

For the upcoming release, we decided to use – apart from this simple alignment method – the *grow-diag-final-and* method, which allows us to create more complicated alignments of more than one word on both sides of the translation. Such an alignment may look like this:

0-0 0-1 1-2 2-3 4-4 5-5 6-6 7-6 8-7 8-8 8-9 9-10 10-10 11-11

(Note the difference: the first word in the target language (0) now corresponds not only to the first (0), but also the second (1) word in the target language.)

From such an alignment we choose, using a simple script, the largest possible number of combinations of words that this alignment allows. In both cases, the aligned pairs of (multiple) words are then sorted and summarized. The result of this automatic excerption is not revised in any way. However, the relative frequency of the corresponding pairs may serve as an indicator of the relevance of the equivalents. The more often the equivalent of the word or multi-word unit occurs in comparison with other equivalents, the greater the likelihood that it is a plausible translation.

1. Possible uses of Treq

Treq is a relatively new application (its initial version, 0.1 alpha, was released in September 2014[[3]](#footnote-3)), but it is quickly gaining popularity among users, especially for its simplicity and straightforwardness. Apart from their most typical use – when seeking a specific equivalent expression for a target language – the equivalents offered by Treq can be considered as potential dictionary equivalents. This is a handy tool for lexicographers, as they instantly get a list of candidates for target language counterparts and their frequencies (expressed both in absolute numbers and percentages), which suggests the probability that a given candidate is functionally equivalent. A significant advantage is the possibility to click on any of them and immediately verify its individual occurrences in the context – and thus more easily distinguish relevant candidates from the misleading ones.

The similar principle, albeit with the help of other instruments, has already been applied in the context of Czech lexicography, e.g. in the case of Croatian (using a Microsoft Access database application; Jirásek 2011) and Lithuanian (the NATools tool used initially for extracting terminology dictionaries; Skoumalová 2008). Both authors agree that dictionaries automatically extracted from a parallel corpus are merely the starting point for subsequent lexicographical work; nevertheless, they can relieve a lot of the burden placed on the lexicographer. This is also confirmed by our own experience, as Treq is being used in the construction of the Latvian-Czech dictionary (Škrabal 2016). Extraction of bilingual dictionaries from parallel and comparable corpora is summarised by Sharoff et al. (2013), cf. also J. Tiedemann’s master’s thesis (Tiedemann 2000).

1. New version of Treq and future prospects

The new version of Treq will allow entering multi-word units into the query window (of course in both directions), yielding both one- and multi-word expressions as results. Until now, it was not possible, e.g. in the English-Czech language combination, to search for many classes of expressions, such as phrasal verbs, discourse markers, phrases in a general sense, etc. (in the direction from English to Czech), and in the opposite direction, e.g. reflexive verbs (which are formed in Czech using a separate reflexive morpheme, *se*/*si*).

Another planned novelty is adding a second source language (currently Czech), namely English. In addition to the bidirectional Czech-X lexicons, bidirectional English-X lexicons will be also generated from InterCorp data. Thus, the possibility to use Treq will open up to a much wider audience, users will no longer be limited by the need to master Czech.

Furthermore, a number of small interface improvements will be implemented as well: e.g. a choice between single- and multi-word equivalent or a case-(in)sensitive switch. With multi-word expressions, it was also necessary to incorporate some form of query language that allows users to take advantage of wild cards.

Further improvements in the results Treq yields can be expected along with the increasing volume of data and genre variety of the texts used. For large and genre varied corpora, it is certainly reasonable to state the frequency of equivalent pairs separately for different types of texts (in Treq: fiction, journalism, legal texts, subtitles); however, the current version of Treq already allows this (field *Restrict* *to*).

We would also like to explore other options of aligning multi-word units, e.g. to start by searching the text for multi-word units using specialized tools and then seek alignment for individual words already within the identified multi-word units. It would also be desirable to try to align words while paying attention to morphosyntactic or syntactic-semantic categories.

1. Acknowledgements

This paper resulted from the implementation of the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

1. References

Čermák, F. & Rosen, A. (2012): The case of InterCorp, a multilingual parallel corpus, *International Journal of Corpus Linguistics* 13, 3, pp. 411–427.

Jirásek, K. (2011). Využití paralelního korpusu InterCorp k získávání ekvivalentů pro chorvatsko-český slovník. In F. Čermák (ed.), *Korpusová lingvistika Praha 2011: 1 – InterCorp.* Praha: NLN, pp. 45–55.

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29, 1, pp. 19–51.

Rosen, A. (2016). InterCorp – a look behind the façade of a parallel corpus. In: E. Gruszczyńska & A. Leńko-Szymańska (eds.): *Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora*. Warszawa: Instytut Lingwistyki Stosowanej, pp. 21–40.

Sharoff, S. & Rapp, R. & Zweigenbaum, P. & Fung, P. (eds.) (2013). *Building and Using Comparable Corpora*. Springer-Verlag: Berlin.

Skoumalová, H. (2008). Extracting dictionaries from parallel corpora. In *Proceedings of The Third Baltic Conference on Human Language Technologies.* Kaunas: Vytautas Magnus University, pp. 297–301.

Škrabal, M. (2016). *Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku.* [*Comparative aspects of Latvian and Czech lexicons: Materials for assembling a Latvian-Czech dictionary.*]PhD thesis. Charles University, Prague.

Tiedemann, J. (2000). *Automatical Lexicon Extraction from Aligned Bilingual Corpora.* Master’s thesis, Otto-von-Guericke-Universität, Magdeburg.

1. For information about the exact composition of the corpus and the size of its components, see [http://wiki.korpus.cz/doku.php/en:cnk:intercorp](http://wiki.korpus.cz/doku.php/en%3Acnk%3Aintercorp). For general information about the InterCorp project, see Čermák & Rosen 2012 or Rosen 2016. [↑](#footnote-ref-1)
2. For details about our setup, see <https://github.com/moses-smt/mgiza/tree/master/mgizapp>. An auxiliary script created by Ondřej Bojar (<http://www1.cuni.cz/~obo/>) was also used. [↑](#footnote-ref-2)
3. Detailed information about individual versions can be found in Version Info at: [https://treq.korpus.cz](https://treq.korpus.cz/index.php). [↑](#footnote-ref-3)