The Translation Equivalents Database (Treq) as a Lexicographer's Aid

Michal Škrabal, Martin Vavřín

Institute of the Czech National Corpus, Charles University, Czech Republic
E-mail: michal.skrabal@ff.cuni.cz, martin.vavrin@ff.cuni.cz

Abstract

The aim of this paper is to introduce a tool that has recently been developed at the Institute of the Czech National Corpus, the Treq (**Tr**anslation **Eq**uivalents) database, and to explore its possible uses, especially in the field of lexicography. Equivalent candidates offered by Treq can also be considered as potential equivalents in a bilingual dictionary (we will focus on the Latvian–Czech combination in this paper). Lexicographers instantly receive a list of candidates for target language counterparts and their frequencies (expressed both in absolute numbers and percentages) that suggest the probability that a given candidate is functionally equivalent. A significant advantage is the possibility to click on any one of these candidates and immediately verify their individual occurrences in a given context; and thus more easily distinguish the relevant translation candidates from the misleading ones. This utility, which is based on data stored in the InterCorp parallel corpus, is continually being upgraded and enriched with new functions (the recent integration of multi-word units, adding English as the primary language of the dictionaries, an improved interface, etc.), and the accuracy of the results is growing as the volume of data keeps increasing.

Keywords: InterCorp; Treq; translation equivalents; alignment; Latvian–Czech dictionary

1. Introduction

The aim of this paper is to introduce one of the tools that has been developed recently at the Institute of the Czech National Corpus (ICNC) and which could be especially helpful to lexicographers: namely, the Treq translation equivalents database[1]. It is based on data stored in the InterCorp parallel corpus (always its latest version, currently v9).

2. InterCorp

InterCorp is a large parallel synchronic corpus under continuous construction at the ICNC since 2005. The corpus has been growing systematically every year in the recent past and, since 2013 (version 6), even obsolete versions of the corpus will remain available via our corpus query interface, KonText, in order to preserve the possibility of replicating previous research. InterCorp is composed of several parts, the most

[1] Available online at http://treq.korpus.cz/.

important and valuable 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 only processed automatically[2], not manually. These include the following types of texts:

- journalistic articles and news published by Project Syndicate and VoxEurop (formerly PressEurop);
- legal texts of the European Union from the Acquis Communautaire corpus;
- proceedings of the European Parliament dated 2007–2011 from the Europarl corpus;
- film subtitles from the Open Subtitles 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. You can therefore find languages which have up to 31 million running words in the core (German) and corpora of individual languages can range in size up to 120 million running words (English), but there are also corpora which have no text in the core (i.e., no manually processed texts) and restrict themselves to collections only (e.g., Vietnamese with a total size of nearly 1.5 million words, consisting only of film subtitles, etc.). Texts in more than half of the languages are provided with morphological annotation (23 out of 39) and lemmatized (20 out of 39). The total size of InterCorp v9 is more than 1.2 billion running words / 1.5 billion tokens[3].

3. Data preparation[4]

First, when preparing data for Treq, only sentences that are aligned[5] 1:1 are selected from the entire InterCorp corpus. We restrict ourselves to this simple alignment because it tends to be more reliable; especially in the case of automatically aligned

[2] For the list of used tools, see http://wiki.korpus.cz/doku.php/en:cnk:intercorp:verze9#acknowledgements.

[3] 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. For general information about the InterCorp project, see Čermák & Rosen (2012) or Rosen (2016).

[4] Cf., e.g. the process of compiling "statistical translation dictionaries" described in Kovář et al. (2016: 343n).

[5] The core component of InterCorp is aligned with the InterText tool (Vondřička, 2014) and this alignment is subsequently manually checked and corrected, mostly in three stages (for details, see Rosen & Vavřín, 2012: 2448). Collections are aligned only by the Hunalign aligner (Varga et al., 2015; see also http://mokk.bme.hu/en/resources/hunalign/), with no correction following. Basic assessment of the quality of our automatic segmentation and alignment can be found in Rosen & Vavřín (2012: 2450).

texts, potential errors can be prevented[6].

The next step is to perform an automatic word-to-word alignment using the GIZA++ tool (Och & Ney, 2003)[7]. In older versions of Treq, a method called *intersection* was used, creating only such alignments where one word in the source language corresponds to one word in the target language, e.g.:

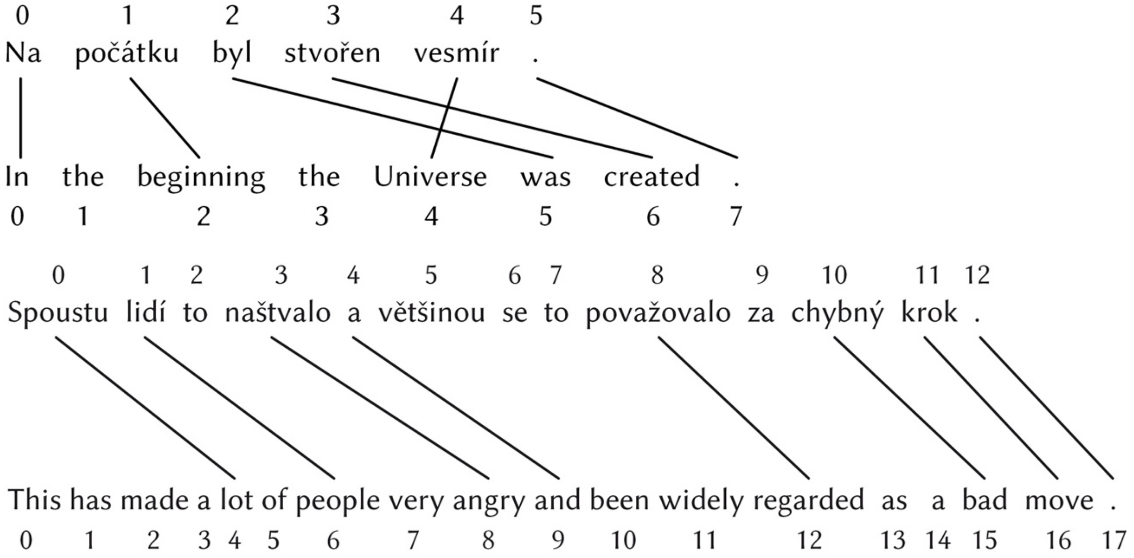

0 1 2 3 4 5
Na počátku byl stvořen vesmír .
In the beginning the Universe was created .
0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7 8 9 10 11 12
Spoustu lidí to naštvalo a většinou se to považovalo za chybný krok .
This has made a lot of people very angry and been widely regarded as a bad move .
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Figure 1: Aligning words using the *intersection* method

That is, 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. (cf. Rosen, Adamová & Vavřín, 2014; Kaczmarska & Rosen, 2015: 164–165).

Starting with release 2.0, apart from this simple alignment method, the so-called *grow-diag-final-and* method has also been used, as it allows the creation of more complicated alignments containing more than one word on both sides of the translation[8]. These multi-word units are not necessarily well-defined entities from a linguistic point of view: some may correspond to what a linguist would analyse as multi-word expressions, some may not.

[6] In the future, however, we would like to experiment also with a non-1:1 alignment (cf. Kovář et al., 2016: 350–351). Other possible plans are outlined in the conclusion of this paper.

[7] 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.

[8] Individual GIZA++ word alignment methods are described and compared by, e.g. Mareček (2009) or Girgzdis et al. (2014). In both papers, the *grow-diag-final-and* method has been evaluated as the most precise and efficient one, therefore it has been adopted also for our purpose.

Such an alignment may look like this:

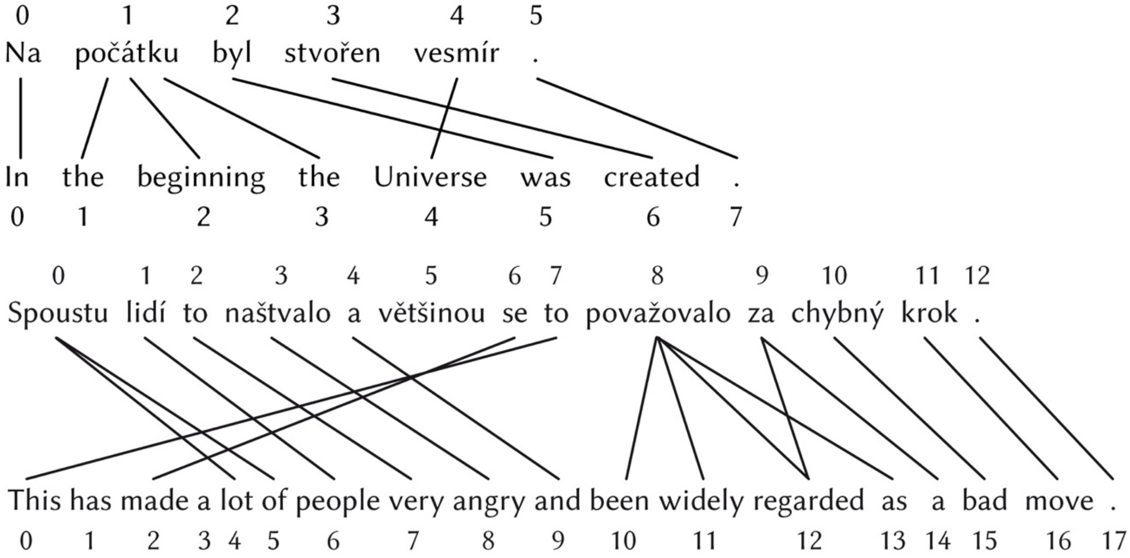

0 1 2 3 4 5
Na počátku byl stvořen vesmír .
In the beginning the Universe was created .
0 1 2 3 4 5 6 7
0 1 2 3 4 5 6 7 8 9 10 11 12
Spoustu lidí to naštvalo a většinou se to považovalo za chybný krok .
This has made a lot of people very angry and been widely regarded as a bad move .
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

Figure 2: Aligning words using the *grow-diag-final-and* method

(Note the difference: the second word in the target language (0) now corresponds not only to the third (2), but also the second and fourth (1, 3) 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 words are then sorted and summarized. The result of this automatic excerption is not revised in any way and is provided to users as a list of found equivalents of the given expression, supplemented with absolute and relative frequencies of aligned pairs.

Table 1 indicates in what proportion the frequencies found in KonText are similar to those displayed by Treq. It also specifies the different data types at each stage of their processing for Treq, considering the InterCorp v9 English component (multi-word variant).

Step by step, you can see the gradual loss of data that are used in the resulting dictionary. In the first step, we only use a 1:1 sentence alignment; thus 20.7% of sentences are lost. Subsequently, both one- and multi-word equivalents are selected based on an alignment made by the GIZA++ tool. However, the relationship between the size of the original corpus and the number of extracted equivalents cannot be clearly predicted, especially in multi-word equivalents where various combinations of the same words arise (see bold pairs below). For example, an alphabetical list of Czech–English couples extracted from the second example sentence above would look like this:

a – and

chybný – bad
krok – move
lidí – people
naštvalo – angry
považovalo – been widely regarded as
považovalo za – been widely regarded as
považovalo za – regarded as
se – made
Spoustu – lot of
to – this
to – very
za – regarded as
. – .

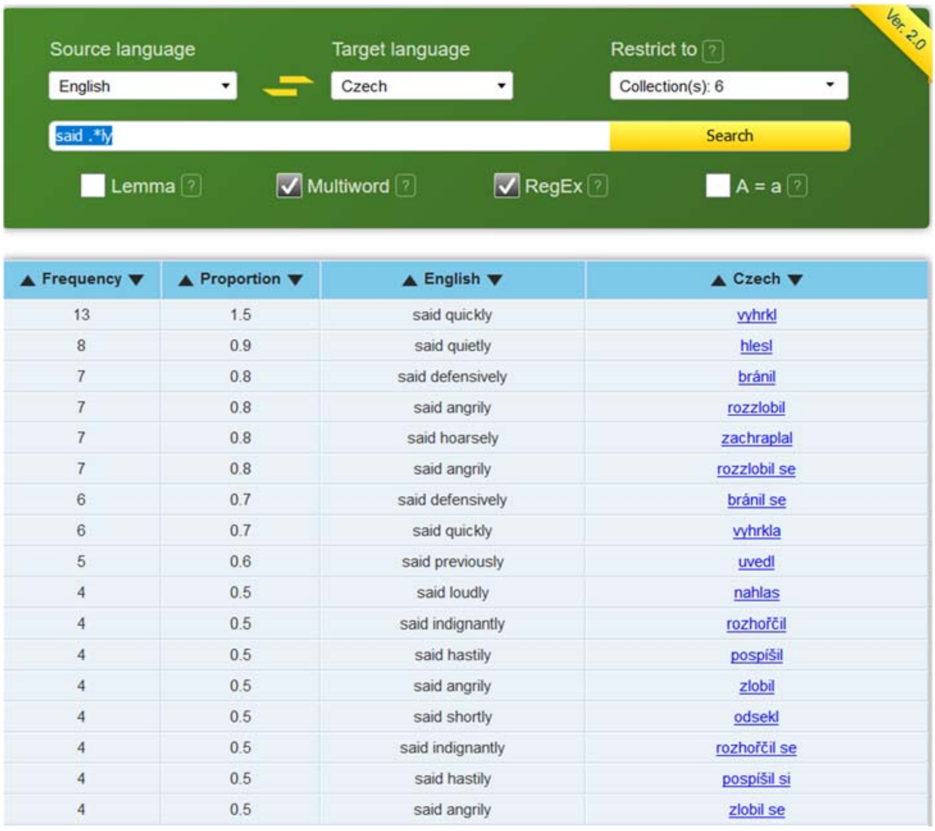
Processing phase	Output data		Count (in thousands)						
			Core	Sub.	Acq.	Eu.	Vox.	Synd.	Total
0. Input	Tokens (in English)		25 149	66 790	29 626	17 384	3 123	4 387	146 458
	Sentences (in English)		1 510	9 211	1 426	681	152	190	13 171
1. Sentence alignment (1:1)	Aligned sentences	lemmas	1 267	6 955	1 251	656	127	180	10 437
		word forms	1 267	6 955	1 254	656	127	180	10 440
2. Word alignment	Equivalents identified	lemmas	15 785	41 189	19 344	12 812	1 670	3 352	94 153
		word forms	15 538	41 445	19 656	12 899	1 598	3 344	94 479
3. Dictionary compilation	Dictionary entries	lemmas	3 235	6 697	1 441	1 213	547	550	13 682
		word forms	4 639	9 276	2 056	1 946	670	873	19 460
4. Dictionary cleanup	Dictionary entries	lemmas	2 775	5 375	1 133	1 061	461	458	11 263
		word forms	3 966	7 146	1 722	1 760	566	750	15 909

Table 1: Data processing for a Czech-English dictionary (Sub.=Subtitles, Acq.=Acquis Communautaire, Eu.=Europarl, Vox.=VoxEurop, Synd.=Project Syndicate)

In the third step, lines that are the same on both sides of the alignment are added together throughout the text. This will give us the list and the frequency of the equivalents. Finally, in the last step, we exclude all the counterparts containing the punctuation to get the final version of the dictionary. For all language pairs where the lemmatization is available on both sides of the alignment, we apply the same procedure to the lemmatized form of data (*na počátek být stvořit vesmír – in the beginning the universe be create*).

4. Interface

Access to the extracted data is then mediated by the Treq online search interface.



Source language: English
Target language: Czech
Restrict to ?: Collection(s): 6

said .*ly
Search

☐ Lemma ?
☑ Multiword ?
☑ RegEx ?
☐ A = a ?

▲ Frequency ▼	▲ Proportion ▼	▲ English ▼	▲ Czech ▼
13	1.5	said quickly	vyhrkl
8	0.9	said quietly	hlesl
7	0.8	said defensively	bránil
7	0.8	said angrily	rozzlobil
7	0.8	said hoarsely	zachraplal
7	0.8	said angrily	rozzlobil se
6	0.7	said defensively	bránil se
6	0.7	said quickly	vyhrkla
5	0.6	said previously	uvedl
4	0.5	said loudly	nahlas
4	0.5	said indignantly	rozhořčil
4	0.5	said hastily	pospíšil
4	0.5	said angrily	zlobil
4	0.5	said shortly	odsekl
4	0.5	said indignantly	rozhořčil se
4	0.5	said hastily	pospíšil si
4	0.5	said angrily	zlobil se

Figure 3: Advanced searching (via RegEx and multi-word units) in the English–Czech section[9]

By default, found counterparts of the searched expression are ranked in descending order of frequency of these equivalent pairs. Their relative frequency is the user's primary guide: the more often the equivalent of the search term occurred compared to other equivalents, the higher the probability that it is plausible. For large-sized and genre-varied corpora, it is advisable to indicate the frequency of equivalent pairs separately for distinct types of texts (see above Section 1) via the *Restrict to* option.

Starting with version 2, it became possible to enter multi-word expressions into the query window (in both directions, of course), yielding both one- and multi-word units as results, in compliance with user preferences. With non-1:1 word alignments, it is

[9] We have adopted this example from Dr. Lenka Fárová (unpublished presentation). It does a good job of showing a non-symmetric nature of equivalent reporting verbs in English and Czech.

now possible, e.g., in the English–Czech language combination, to search for 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*). Moreover, current results more faithfully correspond to the language reality as the equivalence between lexemes in the source and target language cannot, understandably, be limited to an "ideal" 1:1 ratio.

With the implementation of multi-word units, the need to incorporate a query language that would allow the use of wild cards has become urgent[10]: up to now, Treq has only been searching for an exact string of characters. Furthermore, a second primary language (besides Czech), namely English, has been added. And, in addition to the existing bidirectional Czech-X lexicons, bidirectional English-X lexicons have also been generated from the InterCorp data. Thus, the possibility of using Treq is opened up to a much wider audience now as users are no longer limited by the need to master Czech. Theoretically, in the future, the primary language can be extended to any one represented in InterCorp; in this respect, it is necessary to take into account the interests and needs of users.

5. The possible use of Treq in lexicography

(Latvian–Czech dictionary case)

Treq is a relatively new application (its initial version, 0.1 alpha, was released in September 2014[11]), but it is quickly gaining popularity among users, especially for its simplicity and straightforwardness[12]. Possible uses of Treq range from simple, one-shot probes while searching for an equivalent expression for a target language, to more sophisticated and elaborate corpus-assisted translations (Škrabal & Vavřín, 2017: 251–257). However, the equivalents offered by Treq can also 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 along with 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 to immediately verify its individual occurrences in the context, and thus more easily distinguish relevant translation candidates from misleading ones.

[10] Treq is based on the database system MySQL, which uses Henry Spencer's regular expression library compliant to the POSIX.2 standard (see e.g., https://garyhouston.github.io/regex/).

[11] Detailed information about individual versions can be found in the Version Info at: https://treq.korpus.cz.

[12] During 2016, over 719 thousand user interactions were registered at the www.korpus.cz portal. The tool used most often was KonText (with more than 85% of the total), followed by the Treq database (more than 70,000 queries, i.e., almost 200 per day, which represents close to 10% of the total number of queries entered).

The extraction of data from parallel corpora for lexicographical purposes is a logical process that is inherent in the very nature of these data. Partial attempts in this regard have also been undertaken in Czech lexicography, e.g., in the case of English (e.g., Čmejrek, 1998; Čmejrek & Cuřín, 2001; Popelka, 2011), Croatian (Jirásek, 2011), or Lithuanian (Skoumalová, 2008). These authors agree that dictionaries automatically extracted from a parallel corpus are merely the starting point for subsequent lexicographical work; nevertheless, they can relieve much of the burden placed on the lexicographer. This is also confirmed by our own experience as Treq is being used—*inter alia*—for the construction of a Latvian–Czech dictionary (Škrabal, 2016a). It is obvious that the extent to which the retrieved data can be utilised in this way depends primarily on the amount of data for the respective language combination[13].

Currently, the Latvian component of InterCorp (release 9) has a total of over 40.6 million words: the initial manually aligned belletristic core (currently 1,666,000 words) was, as for many other languages in InterCorp, extended by a collection of automatically aligned texts from the Acquis Communautaire corpus (24,667,000 words), Europarl corpus (13,895,000 words) and the OpenSubtitles database (381,000 words).

Let us compare these figures to the situation in the early phase of compiling the Latvian–Czech Dictionary, namely to InterCorp version 3.1 (released in May 2011). The Latvian–Czech component then consisted of parallel fiction texts only (20 in the Czech original, 7 in the Latvian original, and 6 in other languages), numbering slightly more than 1 million running words which were neither lemmatized nor tagged. These data were tentatively processed by the NATools workbench[14] (cf. Skoumalová, 2008) and a simple dictionary (or rather glossary) was compiled. We will inspect the lemma *biedrs* (for individual senses, see below)[15].

biedra [Gen.sg.] (13): 0, ***kamarádův***, *všudy*, *uvěřitelný*, ***kamarád***, *oddělení*, *rozchod*

biedram [Dat.sg.] (16): ***kamarád***, ***soudruh***, *čerstvý*, *budižkničemu*, *trmácet*

[13] Cf. Jirásek's (2011: 55) experience from the Croatian–Czech part of InterCorp: "It turned out that if we do not want to stay at the level of pocket dictionaries, we need a parallel corpus of at least 10 million running words. Such a size of corpus allows us to reliably process equivalents for a medium-sized dictionary. For a larger dictionary, however, it can only serve as an orientation aid, not the main source of equivalents." By a medium-sized vocabulary, is meant one containing approximately 20,000 headwords, representing only "typical and predominant meanings in everyday communication". The larger-sized dictionary should contain about 50 thousand headwords (ibid.: 45).

[14] http://linguateca.di.uminho.pt/natools/

[15] Individual grammatical forms are given with their absolute frequencies in the then corpus, followed by Czech equivalent candidates (as lemmas) ordered by plausibility, as estimated by the frequency of aligned pairs. The plausible candidates for dictionary equivalents are in bold, those with limited application (in collocations mostly) are marked by an asterisk (*), and 0 indicates null equivalents.

> *biedri* [Nom.pl./Voc.sg.] (56): ***soudruh***, ***kamarád***
>
> *biedriem* [Dat./Ins.pl.] (16): ***kamarád***, ***druh***, *trhnout*, ***spolubojovník****, *přeletět*
>
> *biedrs* [Nom.sg.] (52): ***soudruh***, ***kamarád***, ***člen***, ***společník****
>
> *biedru* [Acc./Instr.sg./Gen.pl.] (50): ***člen***, ***kamarád***, 0, ***druh***, ***soudruh***
>
> *biedrus* [Acc.pl.] (13): ***soudruh***, ***kamarád***, *na*, *povzbuzovat*, *brabec*, 0

Nowadays, thanks to the Treq tool, leveraging InterCorp data is as simple for the lexicographer as entering the lemma *biedrs* into the query window, and results can be seen immediately.

> ***člen*** (483 out of 755, i.e., 64%), ***soudruh*** (81), ***kamarád*** (49), *nečlen* (19), *producent* (16), ***kolega*** (12), ***druh*** (11), ***spolužák**** (10), ***přítel*** (7), *poslanec* (7), ***partner**** (7), ***společník**** (6), *členství* (5), ***spolubojovník**** (4), …

It should be noted that these results are useful only in the advanced phase of the lexicographic work on the relevant headword, preceded by an analysis of the corpus data[16] and, in the case of polysemous headwords, drafting the initial sketch of the sense structure. This can often differ from the existing lexicographical description, especially if it is not corpus-based, which is also the case of the chosen lexeme *biedrs.* Thus, the sense division in the newest Latvian monolingual dictionary (MLVV): 1. 'fellow, friend, colleague'; 2. 'member'; 3. 'comrade' had to be rejected for our purpose. On the basis of a manual analysis of corpus data (776 occurrences of the lemma in LVK2013), an overlooked sense[17] (yet, incidentally more frequent than the third one, historically-marked) was discovered; the rank of the first two senses was adjusted by frequency as well into this resulting semantic framework (cf. also Škrabal, 2016b):

> 1. 'member' (497 hits in LVK2013, i.e., 64%); 2. 'fellow, friend, colleague' (204 hits, i.e., 26%); 3. 'deputy' (50 hits, i.e., 6%); 4. 'comrade' (25 hits, i.e., 3%).

Only on the background of such a semantic skeleton did we examine the offered translation candidates in terms of the adequacy of the expression in the source language, i.e., we compared the contexts in which the expressions occur in both

[16] A list of corpora used during the work on the Latvian–Czech dictionary includes, besides InterCorp, also the following three:

- representative Latvian corpus *Līdzsvarots mūsdienu latviešu valodas tekstu korpuss 2013* (LVK2013, 5.5 million tokens, lemmatized, tagged)
- *Latviešu valodas tīmekļa korpuss* (LVTK) compiled from Latvian web pages (over 122 million tokens, non-lemmatized, only partially tagged)
- *lvTenTen corpus* as a member of the TenTen corpora family (Jakubíček et al., 2013) accessible via Sketch Engine (658 million tokens, lemmatized, tagged).

[17] This sense is not a new one, just an updated one from the inter-war period.

languages (via the KonText interface).

By simply modifying the query above into *.*biedrs* (and ticking the *RegEx* option) we will get a considerable amount of compounds with the lemma as its component. These can serve in two ways: either as candidates for separate headwords (e.g., *ceļabiedrs*) or, if written separately (as often happens in Latvian, e.g., *ceļa biedrs*), as potential collocations under the respective headword. Regular expressions can thus provide the lexicographer with possible translation equivalents not only for a single word, but even for a word list.

> *ceļabiedrs* [‘fellow traveller’]: ***spolucestující*** (3), ***společník*** (3), *naštvaný* (1), *průvodní* (1), *spolubojovník* (1), *sužovat* (1), *ušetřený* (1)
>
> *cīņasbiedrs* [‘comrade-in-arms’]: ***spolubojovník*** (1)
>
> *darbabiedrs* [‘colleague, co-worker’]: ***kolega*** (7), ***spolupracovník*** (5)
>
> *domubiedrs* [‘person who holds the same views’]: *podobný* (1), *rodina* (1), ***spojenec*** (1)
>
> *dzīvesbiedrs* [‘spouse, mate’]: *manželka* (50), ***manžel*** (43), ***partner*** (5), *manželský* (2), *držitelův* (1), *Lullingové* (1), *pára* (1), *tabule* (1)
>
> *galdabiedrs* [lit. ‘table-mate’]: *bodávat* (1), ***kumpán*** (1), *stolní* (1)
>
> *karabiedrs* [‘comrade-in-arms’]: ***spolubojovník*** (1), *válečný* (1), *zlíbit* (1)
>
> *klasesbiedrs* [‘classmate’]: ***spolužák*** (41)
>
> *laikabiedrs* [‘contemporary’]: ***současník*** (6), ***pamětník*** (2), ***vrstevník*** (2), *doba* (1), *Gruzie* (1), *spoluobčan* (1), *vyprávění* (1)
>
> *līdzbiedrs* [lit. ‘co-mate’]: *learning* (3), *bližní* (1), ***spolupracovník**** (1), ***vrstevník**** (1), *záhada* (1)
>
> *rotaļbiedrs* [lit. ‘toy-mate’]: ***kamarád**** (1)
>
> *skolasbiedrs* [‘schoolmate’]: ***spolužák*** (37), *kamarád* (1), *spolužákův* (1), *včerejší* (1)[18]

[18] There were only the following compounds with their translation candidates in the data extracted by the NATools workbench:

darbabiedrs: ***kolega***, *se*, ***spolupracovník***, *známý*

karabiedrs: *vyzvědět*

klasesbiedrs: ***spolužák***, *muset*, *zařídit*

laikabiedrs: *můj*, ***pamětník***, *hodně*, *doba*, *nic*, *místo*, *většina*, *průběh*, ***současník***, *Haškův*

skolasbiedrs: *spolužákův*, *recese*, ***spolužák***, 0, *leccos*, *sejít*, *vůbec*, *kamarád*

Finally, after ticking the box *Multiword*, we can extend our list of multi-word expressions and their counterparts with these relevant non-1:1 pairs.

dzīvesbiedrs: *manžel nebo manželka* (22)

arodbiedrības biedrs: *odborář* (16)

konservatīvās partijas biedrs: *konzervativec* (5)

sarunu biedrs: *společník* (5), *protějšek* (3), *partner* (3)

This probe, as well as others carried out while testing the new Treq version, illustratively indicates that, despite its non-representative nature, size, and composition of texts[19], the Latvian–Czech component of the parallel corpus InterCorp, or its extension Treq, respectively, is a valuable source among the sources used to compile a Latvian–Czech dictionary.[20] This is because it is the only one that directly offers Czech equivalents of Latvian lexemes to such an extent. Unlike other similar projects[21] based on parallel corpus data, InterCorp contains a considerable share of original and translated fiction which has been manually checked and therefore provides more precise results. Another advantage, compared to other tools, is Treq's speed, user-friendliness and direct access to parallel concordances via the KonText interface (with its advanced functionality).

Bilingual word sketches (Kovář et al., 2016) are another tool which could be of significant help in the future, along with the Translate button tool (Baisa et al., 2014); but unfortunately, they are not available now for this language combination.

6. Outlook

Further improvements in the results of Treq yields can be expected along with the increasing volume of data and genre variety of the texts used and a gradual improvement in automatic word-aligning tools. At the moment, InterCorp is the largest parallel corpus available for many Czech-X language combinations, including

[19] More precisely: the minimum proportion of Latvian originals that would be ideal for our purposes. In the belletristic core, there are only four novels, one memoir, one book of fairy tales and one shorter essay with the source language being Latvian, while in the Europarl collection there are 268 transcripts from 16 different authors. The total size of such a subcorpus is 387,544 tokens (incl. punctuation), i.e., less than 1% of the total volume of data in the Latvian part of the InterCorp (in the core, the ratio of Latvian originals is about 20%).

[20] Cf. Nikuļceva's (2006) situation when she was writing her Czech–Latvian dictionary a decade and a half ago: there was no Czech–Latvian parallel corpus at all, not to say a Treq-like tool, just a synchronic corpus of Czech SYN2000 (100 million tokens).

[21] Including, e.g., *Opus* (http://opus.lingfil.uu.se/), *Glosbe* (https://glosbe.com/), *Linguee* (www.linguee.com/), *Europarl* (http://www.statmt.org/europarl/) etc., or a parallel corpus of fiction texts in Slavic and other languages *ParaSol* (http://parasolcorpus.org/).

Czech–Latvian[22]. Generally, this relates to a greater effort in the building of parallel corpora in comparison to monolingual ones.

From the example of the polysemous lexeme above, it is apparent that Treq only offers potential translation equivalents, performing no word sense disambiguation. Therefore, it would be desirable to try to align words while paying attention to morphosyntactic and/or syntactic-semantic categories. 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.

7. 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. We would like to thank our colleagues, Michal Křen and Alexander Rosen, for the valuable comments on the article, the latter also for his contribution to the design and implementation of the entire Treq project. In addition, we owe our gratitude to Elżbieta Kaczmarska for giving an impetus to the development of this instrument, to Ondřej Bojar and David Mareček for technical assistance and to Jan Kocek for helping us with the graphic processing of data. Last but not least, we also wish to thank two anonymous reviewers for their thought-provoking comments.

8. References

Baisa, V., Jakubíček, M., Kilgarriff, A., Kovář, V. & Rychlý, P. (2014). *Bilingual word sketches: the translate button.* In A. Abel, C. Vettori & N. Ralli (eds.) *Proceedings of the XVI EURALEX International Congress: The User in Focus.* Bolzano: Institute for Specialised Communication and Multilingualism, pp. 505–513.

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

Čmejrek, M. (1998). *Automatická extrakce dvojjazyčného pravděpodobnostního slovníku z paralelních textů.* Master's thesis. Praha: Ústav formální a aplikované lingvistiky MFF UK.

Čmejrek, M. & Cuřín, J. (2001). Automatic Extraction of Terminological Lexicon from Czech-English Parallel Texts. In *International Journal of Corpus Linguistics Special Issue 2001*, pp. 1–12.

Girgzdis, V., Kale, M., Vaicekauskis, M., Zarina, I. & Skadiņa, I. (2014). Tracing

[22] At least in the traditional sense, as opposed to web-crawled corpora or easily accessible collections of parallel texts online, cf. also Latvian parallel corpora offered via Sketch Engine, including corpus EUR-Lex Latvian 2/2016 with over than 491 million tokens (not tagged yet).

Mistakes and Finding Gaps in Automatic Word Alignments for Latvian-English Translation. In A. Utka et al. (eds.) *Human Language Technologies – The Baltic Perspective. Proceedings of the Sixth International Conference Baltic HLT 2014.* Amsterdam: IOV Press BV, pp. 87–94.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In *7th International Corpus Linguistics Conference CL 2013.* Lancaster: UCREL, pp. 125–127.

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.

Kaczmarska, E. & Rosen, A. (2015). Jak najít optimální překlad polysémních jednotek – porovnání metod formální analýzy paralelních textů. *Časopis pro moderní filologii* 97, 2, pp. 157–168.

Kovář, V., Baisa, V. & Jakubíček, M. (2016). Sketch Engine for Bilingual Lexicography. *International Journal of Lexicography* 29, 3, pp. 339–352.

Mareček, D. (2009). Using tectogrammatical alignment in phrase-based machine translation. In J. Šafránková & J. Pavlů (eds.) *WDS 2009 Proceedings of Contributed Papers.* Praha: Matfyzpress, pp. 22–27.

MLVV = *Mūsdienu latviešu valodas vārdnīca* [online]. Accessed at: http://www.tezaurs.lv/mlvv/. (23 May 2017)

Nikuļceva, S. (2006). *Česko-lotyšský slovník. Čehu-latviešu vārdnīca.* Praha: Leda.

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

Popelka, J. (2011). *Automatické vytváření slovníků z paralelních korpusů.* Master's thesis. Praha: Ústav formální a aplikované lingvistiky MFF UK.

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.

Rosen, A., Adamová, M. & Vavřín, M. (2014). Extrakce lexikálních ekvivalentů z paralelního korpusu. In *Korpusová lingvistika Praha 2014. 20 let mapování češtiny. Abstrakty.* Praha: Ústav Českého národního korpusu, pp. 177–179.

Rosen, A. & Vavřín, M. (2012). Building a multilingual parallel corpus for human users. In N. Calzolari et al. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul: European Language Resources Association (ELRA), pp. 2447–2452.

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

Škrabal, M. (2016a). *Srovnávací aspekty lotyšského a českého lexikonu: Materiály k sestavení lotyšsko-českého slovníku.* PhD. thesis. Praha: Ústav Českého národního korpusu FF UK.

Škrabal, M. (2016b). Straddling the boundaries of traditional and corpus-based lexicography: A Latvian-Czech dictionary. In T. Margalitadze & G. Meladze (eds.) *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity.* Tbilisi: Ivane Javakhishvili Tbilisi State University, pp. 910–914.

Škrabal, M. & Vavřín, M. (2017). Databáze překladových ekvivalentů Treq. *Časopis pro moderní filologii* 99, 2, pp. 245–260.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. In G. Angelova et al. (eds.) *Proceedings of the RANLP 2005*, pp. 590–596.

Vondřička, P. (2014). Aligning parallel texts with InterText. In N. Calzolari et al. (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavík: European Language Resources Association (ELRA), pp. 1875–1879.