Mikhailov, Mikhail. (2021). Mind the Source Data! Translation Equivalents and Translation Stimuli from Parallel Corpora. In: Vincent X. Wang, Lily Lim, Defeng Li (eds). *New Perspectives on Corpus Translation Studies*. Springer, pp. 259-279.

MIND THE SOURCE DATA! TRANSLATION EQUIVALENTS AND TRANSLATION STIMULI FROM PARALLEL CORPORA

Mikhail Mikhailov, Languages Unit, Tampere University, mikhail.mikhailov@tuni.fi

Abstract

Statements like 'Word X of language A is translated with word Y of language B' are incorrect, although they are quite common: words cannot be translated, as translation takes place on the level of sentences or higher. A better term for the correspondence between lexical items of source texts and their matches in target texts would be translation equivalence (Teq). In addition to Teq, there exists a reverse relation – translation stimulation (Tst), which is a correspondence between the lexical items of target texts and their matches (=stimuli) in source texts.

Translation equivalents and translation stimuli must be studied separately and based on natural direct translations. It is not advisable to use pseudo-parallel texts, i.e. aligned pairs of translations from a 'hub' language, because such data do not reflect real translation processes. Both Teq and Tst are lexical functions, and they are not applicable to function words like prepositions, conjunctions, or particles, although it is technically possible to find Teq and Tst candidates for such words as well. The process of choosing function words when translating does not proceed in the same way as choosing lexical units: first a relevant construction is chosen, and next it is filled with relevant function words.

In this chapter, the difference between Teq and Tst will be shown in examples from Russian-Finnish and Finnish-Russian parallel corpora. The use of Teq and Tst for translation studies and contrastive semantic research will be discussed, along with the importance of paying attention to the nature of the texts when analysing corpus findings.

Keywords: parallel corpora, translation equivalents, interlingual correspondences, corpora in lexicography, word alignment

1 Introduction

Electronic corpora are used nowadays in almost every field of linguistic research, and they are especially popular in lexicography (see e.g. Ooi 1998, Krishnamurthy 2008, Walter 2010, Hanks 2012, Kilgarriff 2013), at least when talking about monolingual corpora and projects involving only one language. In recent years, comparable and parallel corpora have also become one of the main sources of data in contrastive and translation studies. "Translation is a source of perceived similarities across languages. Most linguists working in the field have either explicitly or implicitly made use of translation as a means of establishing cross-linguistic relationships" (Johansson 2007: 3). In spite of all this, multilingual corpora do not seem to be used on a large scale for compiling bilingual dictionaries; they remain for the time being only a secondary source of data, if they are used at all. Why is this the case?

The possibilities of extracting bilingual lists of translation equivalents from parallel corpora have been discussed since the 1990s (Tiedemann 1997, Tiedemann 1998, Čmejrek & Cuřín 2001, Danielsson 2003, Kraif 2003, Garabík & Dimitrova 2015, Čermák 2019: 99–100). Many researchers consider parallel corpora a promising source of data for multilingual lexicography (Sinclair 2001, Teubert 2001, Kenning 2010, Doval & Sánchez Nieto 2019, Zakharov & Bogdanova 2020). At the same time, one must admit that this resource presents far more challenges compared to using corpora for compiling monolingual dictionaries (Mikhailov & Cooper 2016: 149-154, Salkie 2008, Salkie 2002, Perdek 2012, Kubicka 2019, Tarp 2020), and therefore comparable corpora are often considered a more realistic alternative (see e.g. Gamallo 2019).

The crucial problem of parallel corpora is that they are much smaller in size than monolingual corpora, and they will never be very large. While the TenTen corpora at Sketch Engine have passed the milestone of 10 G words, even the largest parallel corpora are only approaching the range of 1 G words for some common pairs of languages. Europarl, a parallel corpus of European Parliament debates, contains data in 21 languages of the EU, and it currently has the size of about 50 M tokens per language (Koehn 2005, Tiedemann 2012, https://opus.nlpl.eu/Europarl.php). The UN Parallel Corpus has about 500 M tokens per each of the six languages of the United Nations (ar, en, fr, es, ru, zh) (Ziemski et al. 2016). The ParaCrawl project is crawling parallel tests from the web and has succeeded in collecting data for over 40 language pairs. The largest ParaCrawl corpora are the French-English corpus, with over 1 G tokens, and German-English and Spanish-English corpora, which have close to 1 G tokens (Bañon et al. 2020).

The reason for the relatively modest sizes is that although almost all types of texts are occasionally translated, only a limited number of genres are translated on a regular basis. These are news, technical instructions and user manuals, tourist brochures, political speeches, legal texts (remember

that the famous Rosetta stone had a text of a decree by Ptolemy V inscribed in Ancient Egyptian and Ancient Greek as parallel texts), religious texts (e.g. the Bible), and fiction. Even these sources of data are not as inexhaustible as monolingual texts. Only a small proportion of fiction books is translated, and only documentation for imported products is translated. Likewise, only news from international news agencies are regularly translated. Many other text types – private letters, local news, financial documents, textbooks for schools – are not translated under normal circumstances, unless a special need arises (e.g. evidence for a trial at a court of law). Documents, contracts, agreements, and the international letter exchange of state bodies and international companies are often translated, but most of these documents are not available to the general public. Thus, the amount of natural parallel texts is always incomparable to the amount of monolingual texts circulating in the community. For world languages and for languages with great numbers of speakers, the amount of parallel texts is much larger than for languages of lesser diffusion, and it is clear that for pairs of geographically distant minority languages (e.g. Gaelic-Irish and Kunama, Uyghur and Maltese) natural parallel texts are practically non-existent. Apart from the issue of the availability of the data, aligning parallel texts presents a serious technical challenge that slows down the whole process of compiling a parallel corpus. Large projects use fully automated aligning with some percentage of inevitable misalignments (see e.g. Koehn 2005, Bañón et al. 2020). Because of these issues, bilingual parallel corpora cannot be as large as monolingual corpora. Furthermore, parallel corpora are not available for every language pair, every text type, and every topic.

Emilia Kubicka notes that "scholars dealing with translation studies have repeatedly pointed out the gap between traditional bilingual dictionaries and actual textual reality, and called for the creation of translation dictionaries which reflect the actual linguistic equivalents used by translators" (Kubicka 2019: 75-76). At the same time, it is important to understand that a bilingual dictionary must supply equivalents for any word of any register, even if texts in which some of these words typically occur are seldom or never translated. Unfortunately, parallel corpora would not provide data for all words because of their limited size and restrictions in structure. For this reason, unlike monolingual corpora for monolingual lexicography, parallel corpora will never become a dominating source of data for multilingual lexicography. They will always be an additional resource, to be checked out using monolingual data.

At this point, a salient question arises. In some cases we can suggest that a word *x* from a text in the language *A* has an equivalent *y* in our native language without consulting dictionaries or parallel corpora. How do we manage to do it? Obviously, we do not have an "internal parallel corpus". What we might have in our brains are phrases in our native language that might be used in similar contexts or situations, i.e. a kind of "internal comparable corpus". This means that comparable

corpora have better perspectives as a source of interlingual equivalents compared to parallel corpora. Unlike parallel texts, comparable texts can be found for any text type and for almost any topic. However, comparable corpora cannot be aligned and therefore there is no straightforward way of searching for lexical correspondences. Although researchers actively develop methods of extracting interlingual equivalents from comparable corpora (Delpech 2014, Grabovski 2018, Terryn et al. 2020) such tools are not yet widely available. At the current state of the technologies, comparable corpora are mostly used for reference purposes, e.g. to check out translation equivalents found in a parallel corpus or a dictionary.

In spite of its limited usability as a tool for the lexicographer, the parallel corpus can still be a very useful source of data for contrastive and typological studies. It is much more convincing to study authentic examples rather than the eternal *John killed Mary* or *The cat is on the mat* with do-it-yourself translations into other languages. In his book, Stig Johansson shows multiple case studies from different areas of contrastive studies that benefit from the use of parallel corpora: times of the day, *love/hate*, *to spend time*, *to seem*, *well*, etc. A parallel corpus makes it possible to compare frequencies and thus to detect translationese, to find equivalents used by translators and evaluate their popularity and usability (Johansson 2007). Authentic examples from published translations offer new opportunities for the development of this direction in linguistics, but like any research data, parallel texts require accuracy in use. One must keep in mind, however, that those 'naturally born' authentic examples, as opposed to artificial examples from the top of a linguist’s head, do not appear in the texts for the sake of becoming an illustration of a certain linguistic phenomenon in a scholarly publication, but are instead a result of natural communication activities. The translator does not try to convey a meaning of repeated or interrupted action, the indefiniteness of the object, diminutives, etc. *per se* from the source text: the translator's mission is to transmit a message in another language.

Statements like 'Word *x* of language *A* is translated with word *y* of language *B*' are not quite correct from a linguistic perspective (a detailed explanation of this issue will be provided in the beginning of Section 2). In spite of this, we can sometimes read such statements in linguistic literature (see e.g. Ramón & Labrador 2008, Dobrovol'skij & Pöppel 2016, Pöppel 2018, Zalizniak et al. 2018, Claire Brierley & Hanem El-Farahaty 2019). Of course, most of the authors use the term "translation" as a shortened version of "the item that appears as a representative of the word *x* when translating segments containing *x* into another language", and they understand the difference between translating and choosing a suitable lexical element when translating. Josep Marco uses three terms for this phenomenon: translation, translation solution, and translation correspondence (Marco 2019). In any case, the term "translation" used for interlingual lexical correspondences is

confusing. It downgrades the translation process to a mechanical substitution of elements where a parallel text is considered a set of pairs of matching sentences and not translations performed by a human with certain skills and training at a certain moment of time in a certain place and for a certain audience.

In this chapter, the interlingual lexical correspondences will be discussed from the viewpoint of the translation process. The following issues will be addressed:

- To what extent do translation equivalents from parallel corpora correlate with equivalents from bilingual dictionaries?
- How important is the direction of a parallel corpus for looking up translation equivalents?
- Do words of all grammatical classes have translation equivalents?

The data used in the study will be the Russian-Finnish and Finnish-Russian parallel corpora of fiction texts, ParRus and ParFin. Both corpora are composed of full texts and include works by different authors and translations by different translators. For some works, more than one translation is available. Works from different historical periods are included. Corpora of fiction texts represent language for general purposes, and these data are therefore suited to our study. ParRus and ParFin are different in size and are not identical in composition because of the natural asymmetry of literary translation activities in these two very different cultures. As a result, the two corpora do not form a bidirectional corpus, but they can still be used for comparing Russian-Finnish and Finnish-Russian data. More detailed information on the composition of ParRus and ParFin can be found in Mikhailov & Härme (2015) and Härme & Mikhailov (2016).

2 Translation vs translation equivalent

The term "translation" is overused in linguistic literature. This term often appears in contexts like "Word *x* is translated with the word *y*" or "Word *x* is not translated", etc. Strictly speaking, the expression "translation of the word *x* to language A" is not correct, because translation is "conversion of writing or speech from one language to another" (Danesi 2000, s.v. translation), i.e. only communicative-level units can be called translations, and the lowest appropriate unit would be an utterance. Dorothy Kenny (2011) examines the concept of the translation unit from different points of view and shows that it is not connected to single words in the text, but rather at least to phrases or patterns. For intertextual interlingual matches of lower levels (word, grammatical form, morpheme), it is better to use other terms, for example, "translation correspondence", "translation equivalent", "lexical correspondence", etc. (cf. Kraif 2002).

To study correspondences between source and target texts, two functions, Tr (translation) and Teq (translation equivalence), can be defined. To make the explanation more simple, fictional examples will be used.

Tr(m, sl, tl): translation Tr of the message *m* from the language *sl* to the language *tl.*

> Tr("John killed Mary", en, ru) -> {"Džon ubil Mèri", "Džon pogubil Mèri", "Džon zagubil Mèri", “Džon – ubijca Mèri”, …}

Teq(u, sl, tl): translation equivalent Teq of the lexical unit *u* of the language *sl* in the language *tl*.

> Teq("John", en, ru) -> {"Džon", "Ioann", "Ivan", …}

Obviously, Teq is a reoccurring lexical correspondence, and it does not cover all possible word alignments that can be discovered in parallel texts. Teqs should be more or less compatible semantically. For example, Russian words *on* ‘he’ or *čelovek* ‘person’ should not be included in the list of Russian Teqs of the English personal name *John*, although they might be used for translating messages containing the word *John*.

It is quite obvious to a linguist that when translating message *m* between languages *la* and *lb*:

> Tr(m, la, lb) ≠ Tr(Tr(m, la, lb), lb, la)

This means that the back translation of a message is not likely to reproduce the same message.[1] The Teq function is also irreversible, i.e.:

> Teq(u, la, lb) ≠ Teq(Teq(u, la, lb), lb, la)

It is very important to understand that translations have a direction from source language to target language. Consequently, parallel corpora also have a direction: they can be uni- or bidirectional. If a corpus is bidirectional, it is necessary to define subcorpora including texts with required directions of translation.

In addition to "natural" parallel texts, where original source texts are paired with their direct translations, there are indirect translations, where translation is performed via a third language. This happens sometimes with translations of fiction when it is difficult to find a translator with the required pair of languages (or for other reasons). For example, all works by Chinghiz Aitmatov, a renowned Kyrgyz author of the Soviet period, were translated into Finnish from Russian, including his early works, which were originally written in the Kyrgyz language. In multilingual environments, it is possible to obtain pseudo-parallel texts, where both paired texts are translations from a third language. For example, most EU documents are available in all the official languages of the European Union, and it is therefore possible to obtain parallel texts for language pairs like

1 This is true even for machine translation: the result of back translation is often different from the initial source language message.

Lithuanian and Greek, Maltese and Danish, etc. However, these parallel texts will be pseudo-parallel, because in fact the texts are translated from another language, most likely, from English. It is obvious that in most cases, one should avoid using indirect translations and pseudo-parallel texts.

So, if Russian translation equivalents for Finnish words are to be found, direct translations from Finnish to Russian are required, not translations from Russian to Finnish. The latter will not yield Russian translation equivalents, but the Russian translation stimuli of Finnish words. (In everyday life, one can say *Your father is just like you*, but it is clear that this statement does not look quite natural). As for lexical correspondences acquired from pseudo-parallel texts or indirect translations; they cannot be interpreted in terms of the translation of this pair of languages. McEnery and Xiao note that the direction of translation is important for corpus-based contrastive studies (McEnery & Xiao 2007), and it is worth adding that it is equally important in lexicography.

Let us take a simple example from our data. Finnish-Russian dictionaries register for the Finnish word *sauna* 'bath' two Russian Teqs, *sauna* and *banja*, while Russian-Finnish dictionaries suggest for the Russian word *banja* 'bath' only one Finnish Teq, *sauna*.

Teq("sauna", fi, ru) -> {"sauna", "banja"}
Teq("banja", ru, fi) -> {"sauna"}

The first Russian Teq for *sauna* is a borrowing from Finnish. We can assume therefore that if we look up Russian translation equivalents for the Finnish word *sauna* in real-life translations from Finnish to Russian, we would find mostly examples with the word *sauna*, because it is a Finnish culturally-bound word and would be more appropriate for texts about Finland (as most texts in Finnish are expected to be). If we build a reverse parallel concordance for the Finnish word *sauna* in a Russian-Finnish corpus, we are likely to get both *sauna* 'sauna' and *banja* 'Russian bath'. The word *banja* would be used as a general word for any bath or to refer to the Russian traditional bath, while the word *sauna* would refer only to the Finnish bath. For this reason, one can expect that the word *banja* would be more common than the word *sauna*.

This hypothesis was not however fully confirmed in authentic material: the parallel concordances from corpora of literary texts yield slightly different results (see Table 1 and Table 2). In the Finnish-Russian corpus, the equivalent *banja* gets an unexpectedly high frequency, and only separate querying of two subcorpora – the "pre-war" = "before 1945" and "post-war" = "after 1945"[2] – makes it clear that the Finnish borrowing *sauna* means in Russian a "modern", "urban", electrical Finnish bath, and therefore in Russian translations of works by Aleksis Kivi, Juhani Aho, and other classical authors of Finnish literature, the word *sauna* is rare and the equivalent *banja* is

2 In Finland, like in many other countries of Europe, the processes of urbanisation and industrialisation accelerated after the end of World War II, and the whole way of living changed.

used instead. As for reverse concordancing in the Russian-Finnish corpus, the word *sauna* occurs on the Russian side only once, and it means ‘Finnish sauna’: all the other examples have *banja* 'Russian bath'.

Table 1 Matches for the Finnish word *sauna* in the Finnish-Russian parallel corpus

Matches	**Before 1945**	**After 1945**	**Total Result**
banja	139	67	**206**
sauna	12	140	**152**
∞[3]	18	9	**27**
Total Result	**169**	**216**	**385**

Table 2 Matches for the Finnish word *sauna* in the Russian-Finnish parallel corpus (reverse concordancing)

Matches	**F**
banja	242
sauna	1
∞	9
Total Result	**252**

This example demonstrates that the direction of the corpus matters: a search in a corpus containing translations in both directions would yield unreliable results, a search in the wrong direction is likely to lead to wrong conclusions, and the use of indirect translations and pseudo-parallel texts would distort the picture even more. In the example with the Russian equivalents for the Finnish word *sauna*, a search in Russian-Finnish texts would give us an impression that *banja* is the only Russian equivalent for the Finnish word *sauna*, which would be incorrect, and only a carefully organised search in the Finnish-Russian corpus would show that there are two translation equivalents – *sauna* and *banja* – and the choice depends on the cultural context.

3 Translation equivalent vs translation stimulus

The example from the previous section demonstrates that a reverse parallel concordance is not the same thing as a parallel concordance. A reverse parallel concordance does not tell us about translation equivalents, but about the language units of the source text that provoke the use of certain units in translation. Let us call this dependence **translational stimulus**. Translational stimulus Tst (u, sl, tl) is a function, the reverse to the function Teq. It is obvious that

$$Teq(w, la, lb) \neq Tst(w, la, lb),$$

3 The sign ∞ is used to mean 'other equivalents'.

although the resulting sets usually do have an overlap. This was just demonstrated in the example with the word *sauna*.

In order to have a closer look, let us take a more complex example – the Finnish Teq for the Russian word *volosy* ‘hair’. This time, the concordances are much longer: over 900 examples in the Russian-Finnish corpus and over 600 in the Finnish-Russian one. Fortunately, it is not necessary to read all the examples and mark equivalents manually. Smaller concordances can be handled in Excel by means of applying filters to a table and group annotation. Very large tables can be processed in R by running relatively simple scripts that match examples for substrings and assign relevant equivalents to each example.

After checking these two large parallel concordances, we have Tables 3 and 4. Surprisingly, the lists of Finnish correspondences and their rank places coincide in both tables, although the normalised frequencies (ipm = instances per million tokens) vary substantially.

Table 3 Finnish Teq for the Russian word *volosy* 'hair' (Russian-Finnish corpus)

Fi	**F**	**ipm**
hiukset 'hair'	578	161.20
tukka 'hair'	195	54.38
karvat 'bristle'	35	9.76
hius(-) 'hair'	33	9.20
kihara 'curly'	18	5.02
pää 'head'	12	3.35
jouhi 'horsehair'	5	1.39
∞	19	5.30
⌀[4]	19	5.30
Total Result	**911**	254.08

4 The sign ⌀ means the omission of the unit in the corresponding segment.

Table 4 Finnish Tst for the Russian word *volosy* 'hair' (Finnish-Russian corpus)

Fi	**F**	**ipm**
hiukset 'hair'	314	201.74
tukka 'hair'	234	150.34
karvat 'bristle'	22	14.13
hius(-) 'hair'	20	12.85
kihara 'curly'	14	8.99
pää 'head'	4	2.57
jouhi 'horsehair'	2	1.28
∞	8	5.14
⌀	21	13.49
Total Result	**639**	410.54

In the Teq list, the first equivalent, *hiukset*, outmatches all the remaining candidates, while in the Tst list, the second stimulus, *tukka*, closely follows the first equivalent. The phenomenon can be explained by the interference of the source language during translation in the Russian-Finnish data. Obviously, the Finnish translators subconsciously choose for the Russian pluralia tantum *volosy* a Finnish pluralia tantum *hiukset*, although there is another equivalent, *tukka*, which is as good, but is a singularia tantum. This case shows that if only the Teq are checked, one can possibly overlook a good suggestion. Still, it would not be a good idea to mix the two sets of data.

More substantial differences between Teq and Tst can be seen after analysing parallel concordances for the Russian verbs *pokupat'* and *kupit'*, 'to buy'. The two verbs make an aspect pair:[5] the first verb is imperfective and has the meaning of a habitual, incomplete, and repeated action of buying, while the second is a perfective verb and has the meaning of a completed action. The aspectual differences are not only grammatical, but also semantic, which results in the use of different translation equivalents, as can be seen in Tables 5 and 6. The Tst list is shorter, and the difference in frequencies is visible to the naked eye.

5 Russian verbs belong to one of two aspects: the perfective (which sees the situation as a single whole (Comrie 1976: 16)) or the imperfective (which refers to general facts, or to continuing or repeated events). Perfective verbs have two tense forms: the past and future simple. Imperfective verbs have three tense forms: the past, present, and future complex (which is formed with the auxiliary byt' 'to be' + infinitive). Gerunds of perfective verbs are in the past tense, while gerunds of imperfective verbs are in the present tense. Perfective verbs can only form past participles, while imperfective verbs form both present and past participles. The Russian language does not have a perfect aspect (which should not be confused with the Russian perfective). Verbs with close meaning belonging to different aspects form so-called aspect pairs. These paired verbs can replace each other in different contexts. Still, they are different lexemes, not forms of the same word. For details, see (RG 1980: §§1384-1387, 1490-1498).

Table 5 Finnish Teq for *pokupat' / kupit'* (Russian-Finnish corpus)

kupit'			**pokupat'**		
Fi	**F**	**ipm**	**Fi**	**F**	**ipm**
ostaa 'to buy'	657	183.24	ostaa 'to buy'	179	49.92
hankkia 'to obtain'	7	1.95	ostella 'to do shopping'	24	6.69
lahjoa 'to bribe'	6	1.67	kauppa 'store, N'	3	0.84
saada 'to get'	6	1.67	∞	5	1.39
maksaa 'to pay'	4	1.12	⌀	2	0.56
ostella 'shop, N'	3	0.84			
∞	12	3.35			
⌀	8	2.23			
Total Result	**701**	**195.51**	**Total Result**	**213**	**59.41**

Table 6 Finnish Tst for *pokupat' / kupit'* (Finnish-Russian corpus)

kupit'			**pokupat'**		
Fi	**F**	**ipm**	**Fi**	**F**	**ipm**
ostaa 'to buy'	402	258.27	ostaa 'to buy'	134	86.09
hankkia 'to obtain'	41	26.34	hankkia 'to obtain'	5	3.21
saada 'to get'	15	9.64	hakea 'to seek'	3	1.93
hakea 'to seek'	6	3.85	∞	6	3.85
ottaa 'to take'	4	2.57	⌀	7	4.50
∞	18	11.56			
⌀	16	10.28			
Total Result	**503**	**323.16**	**Total Result**	**155**	**99.58**

Again, we have to admit that the Tsts from the reverse concordances give some idea about lexical correspondences in the languages in question. As in the previous example with the noun *volosy* 'hair', some interference with the Russian originals can be noticed: among the Finnish equivalents for the Russian perfective verb *pokupat'*, the second place is occupied by the Finnish verb *ostella* 'to shop' with quite a high frequency. This verb has the additional semantics of recurring action and is more frequent in Russian translations than in non-translated Finnish, e.g. in the fiTenTen2014 corpus hosted at Sketch Engine – it has a frequency of 4.28 ipm. The list of Tsts for these verbs (Table 5) does not contain *ostella*. This list, however, provides us with two good suggestions that are not in the Teq list: *hankkia* 'obtain' and *saada* 'get'.

It is important to understand that Tsts do not reflect the real translation processes. However, unlike Teqs, Tsts are not subject to interference and can help to eliminate such lexemes. Jurkiewicz-Rohrbacher distinguishes between translation equivalents, which work only in the direction of

translation, and functional equivalents, which work both ways (Jurkiewicz-Rohrbacher 2019: 110-111). In our case, comparing Teqs and Tsts does not produce inverse correspondences, but helps to filter out the equivalents that are influenced by the source language. Tsts would therefore be useful for contrastive and typological studies. Nevertheless, the researcher should understand the difference between Teqs and Tsts, look up Teqs and Tsts separately, and purposefully use Tsts to detect asymmetry in the lexical systems of the two languages.

4 Does any word have translation equivalents?

When talking about translation equivalents, it is also important to understand whether all lexemes can have translation equivalents. In corpus linguistics, aligning parallel texts at the word level, so-called word alignment, is practiced (Tiedemann 2004, Östling & Tiedemann 2016). The purpose of such alignment is to find the maximum number of matches between the words of aligned sentences. The starting point of the algorithm is an assumption of the presence of a potential match for any token.

Let us illustrate word alignment in a simple Russian sentence, *Ja čitaju knigu s babuškoj*, and its English and Finnish translations, *I am reading a book with grandma* and *Luen kirjaa mummon kanssa* (see Figures 1 and 2).

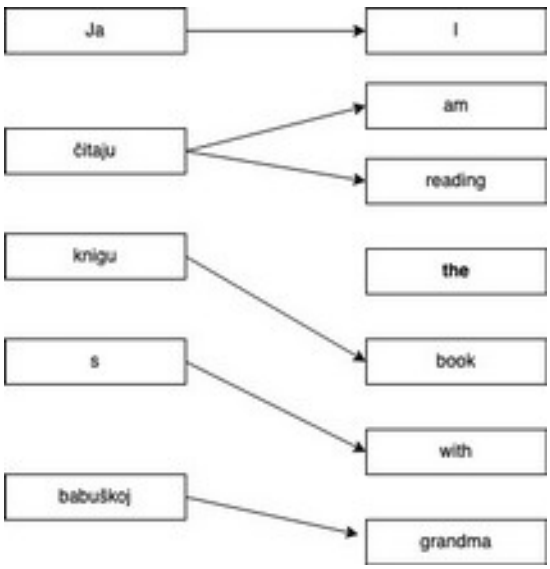

Fig. 1 Word alignment: A Russian-English example

It is clear even from these simple examples that some tokens of the source sentence have no correspondence in the translations and some may correspond to more than one token in the target text. Even for the tokens that can be aligned, there are doubts whether they are indeed "translated"

and whether ‘translation equivalent’ would be the correct term here. Are the tokens *with* and *kanssa* Teq for the Russian preposition *s* ‘with’? As we know, the choice of preposition often depends on the noun, cf. ru *Petr v škole -> Petr is at school* and *Petr v komnate -> Petr is in the room*, where the Russian preposition *v* 'in' corresponds with the English preposition *at* in the first sentence and *in* in the second sentence.

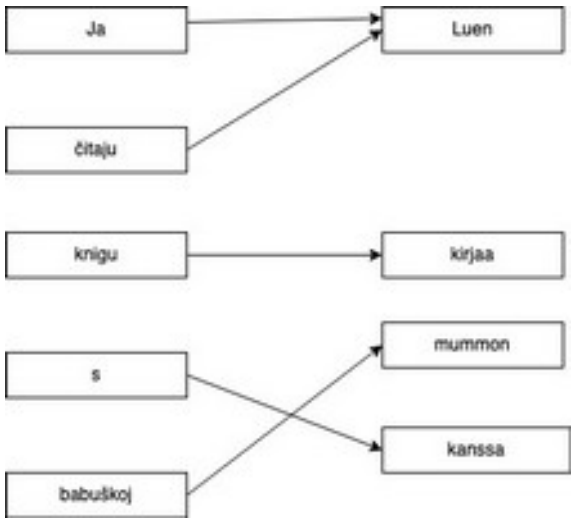

Fig. 2 Word alignment: A Russian-Finnish example

To check whether translation equivalence and translation stimulation are applicable for function words, I looked up the Finnish correspondences for the Russian conjunction *hotja* ‘although’ in the Russian-Finnish corpus. This time, the search was performed on the texts starting from the middle of the 20th century. The results of the search can be found in Table 7.

Table 7 Finnish correspondences for the word *hotja* ‘although’ (Russian-Finnish data)

Teq	**F**	**ipm**
vaikka 'although'	510	336.14
edes 'even'	49	32.3
tosin 'indeed'	26	17.14
ainakin 'at least'	24	15.82
mutta 'but'	21	13.84
joskin 'although if'	13	8.57
huolimatta 'in spite of'	6	3.95
vaan 'though'	6	3.95
kuitenkin 'still'	5	3.3
paitsi 'except'	1	0.66
∞	42	27.68
Total number of examples	**703**	**463.35**

The reverse search for translation stimuli in the Finnish-Russian corpus provides a very similar list of correspondences (Table 8). Interestingly, the conjunction *hotja* is much more frequent in translations into Russian than in original Russian texts; the difference in relative frequencies is almost triple. The frequencies of Tsts descend more smoothly than the frequencies of Teqs, where *vaikka* 'although' clearly dominates. From the statistics in Table 7, we can see that the conjunction *vaikka* 'although' is the absolute favourite: 71% of the contexts are translated into Finnish using this conjunction, and this corresponds with the recommendations of the Russian-Finnish dictionaries. The Finnish-Russian data (Table 8) also have *vaikka* as the main correspondence for *hotja* with 68% of all examples. However, in this data *mutta* 'but', *edes* 'even', *ainakin* 'at least', and *kuitenkin* 'still' are more visible and have much higher frequencies than in Table 7.

Table 8 Finnish correspondences for the word *hotja* ‘although’ (Finnish-Russian data)

Tst	**F**	**ipm**
vaikka 'although'	1003	850.82
mutta 'but'	119	100.95
edes 'even'	55	46.66
ainakin 'at least'	46	39.02
kuitenkin 'still'	31	26.3
vaan 'though'	19	16.12
tosin 'indeed'	13	11.03
huolimatta 'in spite of'	8	6.79
joskin 'although if'	7	5.94
paitsi 'except'	7	5.94
∞	165	139.97
Total number of examples	**1473**	**1249.51**

The remaining part of the lists contrasts the Teq statistics for the content words in the previous section: many of the matches are not only unlikely to appear in bilingual dictionaries, but are not even conjunctions.

To get a better understanding of what is going on, let us have a look at a few examples:

(1) К чему этот насмешливый тон? Причем тут "наследники"? **Хотя** жена действительно …
(Пастернак Б.Л., Доктор Живаго) ('What is this mocking tone for? What do the "heirs" have to do with this? **Although** the wife indeed...')
Miksi tuollainen pilkallinen sävy? Mitä tekemistä tässä on perillisillä? **Tosin** vaimo todellakin ... (transl. J. Konkka.) ('Why such a mocking tone? What do the "heirs" have to do with this? **Really** the wife indeed...")

(2) Вы **хотя** бы отдаленно представляете себе, о чем говорите? (Маринина А., За все надо платить)
('Do you understand **at least** approximately, what you are talking about?')
Onko teillä harmainta**kaan** käsitystä siitä mitä te puhutte? (transl. O. Kuukasjärvi)
('Do you have **any** slight idea of what you are talking about?')

(3) Он все-таки **хотя** и очень милый, но странный. (Улицкая Л., Сквозная линия)
('**Although** he is nice, still he is strange')
Kaikesta rakastettavuudestaan **huolimatta** hän oli kovin omituinen mies. (transl. A. Pikkupeura)
('**In spite of** all his loveability, he is a very strange man')

In example (1), the structure of the translation is more or less similar to that of the source text, but in examples (2) and (3), the translators changed the syntax and the correspondences for *hotja* are not easy to find.

We get an even more contradictory picture for the Finnish correspondences of the Russian particle *nu* ‘well, so’ (Table 9).

Table 9 Finnish correspondences for the word *nu* ‘well, so’ (Russian-Finnish data)

Teq	**F**	**ipm**
no 'well'	1742	1148.16
niin 'so'	155	102.16
mutta 'but'	95	62.62
entä 'and'	71	46.8
ja 'and'	71	46.8
nyt 'now'	67	44.16
mikä/mitä 'what'	52	34.27
hyvä 'good'	48	31.64
voi 'oh'	41	27.02
sitten 'than'	35	23.07
kyllä 'yes'	25	16.48
vaikka 'although'	23	15.16
jo 'already'	21	13.84
siinä 'there'	20	13.18
oikein 'really'	16	10.55
vain 'only'	15	9.89
hei 'hi'	12	7.91
totta 'true'	12	7.91
ihan 'really'	5	3.3
∞	213	140.39
Total number of examples	**2739**	**1805.29**

The length of the list speaks for itself, as it demonstrates that there are no exact correspondences (cf. Salkie 2002) for the Russian particle *nu* in Finnish texts. The dominating *no* 'well' covers only about 30% of cases, and it is mainly used when translating sentences with *nu* in the initial position. The remaining Teq are all so different that it is even hard to imagine how all these Finnish words could correspond to the same Russian word.

The inverse parallel concordance from the Finnish-Russian data quite expectedly also yields a long vague list of correspondences (see Table 10). It is worth noting that this time particle *nu* is much more frequent in the texts originally written in Russian.

Table 10 Finnish correspondences for the word *nu* ‘well, so’ (Finnish-Russian data)

Tst	**F**	**ipm**
no 'well'	668	566.65
niin 'so'	84	71.26
mutta 'but'	45	38.17
nyt 'now'	42	35.63
ja 'and'	35	29.69
sitten 'than'	30	25.45
voi 'oh'	28	23.75
entä 'and'	27	22.9
mikä/mitä 'what'	27	22.9
jo 'already'	25	21.21
kyllä 'yes'	21	17.81
hyvä 'good'	15	12.72
vaikka 'although'	12	10.18
siinä 'there'	8	6.79
ihan 'really'	7	5.94
hei 'hi'	5	4.24
vain 'only'	5	4.24
totta 'true'	4	3.39
oikein 'really'	3	2.54
∞	109	92.46
Total number of examples	**1200**	**1017.93**

Checking some contexts with *nu* from the Russian-Finnish data again demonstrates changes in the syntax of the translations.

(4) **Ну** да где тут думать, поезд-то уж близко, думать некогда. (Пастернак Б.Л., Доктор Живаго)
('**So** when would you think, the train is already close, no time to think')
Vaikka eihän siinä ollut ajattelemisen aikaa, juna oli jo lähellä. (transl. Juhani Konkka)
('**Anyway** there was no time for thinking, the train was already close')

(5) **Ну**, скажем, в театр? (Булгаков М.А., Театральный роман)
('**Well**, for example to a theatre?')
Sanotaan **nyt** vaikka teatteriin? (transl. Esa Adrian)
('Shall one say **now** for example to a theatre?')

(6) Дядя Толя книжку принес старинную. Называется "Заветные сказки". Старинные сказки русские, необработанные. Там такие тексты, **ну** точно как бабушка выдает. (П. Санаев. Похороните меня за плинтусом)
('Uncle Tolja has brought a book, an old one. It is called "The Secret Tales". Old Russian fairy tales, unabridged. There are such texts there, **well**, exactly like those grandma does.')
Tolja-setä toi ikivanhan kiijan. Sen nimi on Perinnesatuja. Siinä on vanhoja venäläisiä satuja, muokkaamattomia. Siellä on sellaisia tekstejä, **ihan** niin kuin mummo pudottelee. (transl. Kirsti Era)
('Uncle Tolja brought a very old book. It is called Traditional tales. There are old Russian tales there, unchanged. There are such texts there, **well**, exactly like grandma gives out'.)

The explanation is simple: *nu* is a discourse word, and as such it does not even have its own meaning but is rather used to underline or emphasise certain elements of the utterance where it is

used and for linking the current sentence to previous sentences. Such marker words function in different languages in very different ways, and there is no direct correspondence between them. There might be many different ways to map the message of an utterance of the source into an utterance of the target text.

When searching for Teqs for cohesion words, one often has to act by the method of exclusion, that is, to start with determining Teqs for content words – nouns, verbs, adjectives, and adverbs – and only at the next stage try to find matches for the remaining tokens (cf. automated word aligning techniques, see e.g. Tiedemann 2004). In fact, these words are not dictated by the tokens of the source text, but rather by the syntactic constructions and communicative functions of utterances. Therefore, establishing links with the source text is just a convention; the translator hardly cares about expressing the concrete lexemes like *nu* or *hotja* in translation, although he/she is likely taking pains to express the meanings of uncertainty or concession that are present in the utterance to translate.

To sum it up, although Teq and Tst searches for a function word might return some frequently reoccurring matches, as happened in the cases above, they are not very helpful for practical use as opposed to Teq and Tst searches of content words: nouns, verbs, adjectives, and adverbs.

5 Conclusions

The examples given in this chapter demonstrate that findings from parallel corpora are not identical to equivalents registered in bilingual dictionaries. Parallel corpora may suggest good solutions not listed in dictionaries, and it is possible to check which equivalents are most frequently used for translating. At the same time, parallel corpora sometimes demonstrate the influence of dictionaries on translators and in this way form a vicious circle (cf. e.g. Perdek 2012, Mikhailov 2020). Despite these reservations, the community has already noticed the usefulness of these data and many lexicographical services – GlosBe, Linguee, and the like – provide in addition to dictionary entries concordances from parallel corpora.

The two reverse functions – Teq (translation equivalent) and Tst (translational stimulus) – that were introduced in this chapter give a better understanding of lexical correspondences in parallel texts. Only the former reflect real translation processes, as the other is an *a posteriori* link leading backwards from the target to the source text. Nevertheless, it can be useful for checking out natural translation equivalents and detecting those that are "infected" with source language interference.

The adequate direction of translation and the exclusion of pseudo-parallel texts play an important role in all cases. Only the correctly chosen data will provide correct results that have theoretical and practical value. One might say that this has nothing to do with specialist texts that are dealing with

technical, economic, or legal issues: special terms are the same in any language. This is not quite true. Different languages have different traditions in terminological issues as well, which might result in multiple interlingual correspondences and substantial differences in frequencies depending on the direction of translation. It is probable that ignoring the direction of translation in the data used for developing MT systems might affect the quality of translation.

The examples given in the chapter show that the functions Teq and Tst work only with content words, i.e. nouns, verbs, adjectives, and adverbs. For these word classes, one can get useful information on interlingual correspondence for lexemes.

Cohesion words (conjunctions, prepositions, particles) of the translation are not dictated by the source text; they appear in the target text for the purpose of joining the content words into meaningful entities, and they are adjusted at the editing stage in accordance with the language and style norms of the target language. Therefore, if we talk about translation equivalents, there would be no Teqs for specific particles, prepositions, or conjunctions, but rather for the constructions they are used in.

For example, the English preposition *with* does not have any Teq in other languages, but the construction 'with + Noun' does. In Russian, it would be 'preposition *s* + noun in the Instrumental case', in Finnish 'noun in the Genitive case + postposition *kanssa*' or 'noun in the Comitative case'. In addition to these direct correspondences, other translation equivalents are possible.

When working with constructions, one would need to highlight sets of formal features of a certain construction, then get examples from a corpus, and only after that look up an appropriate construction in another language. Hence, the whole procedure would be different.

Translation equivalence on the level of constructions can also be very helpful with terms and phraseological units. A construction grammar (Fried & Östman 2004) would be a useful instrument to explain relations between multiword elements.

References

Bañón, Marta and Pinzhen Chenz and Barry Haddowz et al. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic*s, Online, 4555–4567.

Baños, Rocío. 2013. ‘That is so cool’: investigating the translation of adverbial intensifiers in English-Spanish dubbing through a parallel corpus of sitcoms. *Perspectives*, 21:4, 526-542, DOI: 10.1080/0907676X.2013.831924

Čermák, Petr. 2019. InterCorp. A parallel corpus of 40 languages. In *Parallel Corpora: Creation and Applications,* eds Irene Doval and Maria Teresa Sánchez Nieto, 93–102. Amsterdam/Philadelphia: John Benjamins.

Claire Brierley, and Hanem El-Farahaty. 2019. An interdisciplinary corpus-based analysis of the translation of كرامة (karāma, 'dignity') and its collocates in Arabic-English constitutions. *JosTrans*, 32: 121-145. https://jostrans.org/issue32/art_brierley.pdf

Čmejrek, Martin and Jan Cuřín. 2001. Automatic Extraction of Terminological Translation Lexicon from Czech-English Parallel Texts. *International Journal of Corpus Linguistics,* 6 (Special Issue): 1–12.

Comrie, Bernard. 1976. *Aspect. An Introduction to the Study of Verbal Aspect and Related Problems.* Cambridge: Cambridge University Press.

Danesi, Marcel. 2000. *Encyclopedic Dictionary of Semiotics, Media, and Communication*. Toronto Studies in Semiotics. Toronto: University of Toronto Press.

Danielsson, Pernilla. 2003. Automatic extraction of meaningful units from corpora: a corpus-driven approach using the word stroke. *International Journal of Corpus Linguistics*, 8(1): 109-27.

Delpech, Estelle Maryline. 2014. *Comparable Corpora and Computer-Assisted Translation*. London, Hoboken: John Wiley & Sons.

Doval, Irene, and Maria Teresa Sánchez Nieto. 2019. Parallel corpora in focus: an account of current achievements and challenges. In *Parallel Corpora: Creation and Applications,* eds Irene Doval and Maria Teresa Sánchez Nieto, 1–15. Amsterdam/Philadelphia: John Benjamins.

Dobrovol'skij Dmitrij, and Ludmila Pöppel. 2016. Discursive Constructions in the Russian-Swedish Dictionary Database: A Case Study of *v tom-to i N*. In *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity,* eds Tinatin Margalitadze, George Meladze, Ivane Javakhishvili, 668–677. Tbilisi: Tbilisi University Press.

Fried, Mirjam & Jan-Ola Östman. 2004. Construction Grammar. A thumbnail sketch. In *Construction Grammar in a Cross-Language Perspective*, eds M. Fried, M., & J.-O. Östman, 11-86. John Benjamins Publishing Company, Philadelphia.

Gamallo, Pablo. 2019. Strategies for building high quality bilingual lexicons from comparable corpora. In *Parallel Corpora: Creation and Applications*, eds Irene Doval and Maria Teresa Sánchez Nieto, 251–266. Amsterdam/Philadelphia: John Benjamins.

Garabík, Radovan & Ludmila Dimitrova. 2015. Extraction and presentation of bilingual correspondences from Slovak-Bulgarian parallel corpus. *Cognitive Studies | Études Cognitives*, 15: 327–334 . DOI: 10.11649/cs.2015.022

Grabowski, Łukasz. 2018. On identification of bilingual lexical bundles for translation purposes: the case of an English-Polish comparable corpus of patient information leaflets. In *Multiword Units in Machine Translation and Translation Technology,* eds. Ruslan Mitkov, Johanna Monti, Gloria Corpas Pastor, Violeta Seretan. John Benjamins, 182-199. DOI: 10.1075/cilt.341.09gra

Hanks, Patrick. 2012. The corpus revolution in lexicography. *International Journal of Lexicography*, Vol. 25 No. 4, pp. 398-436 . doi:10.1093/ijl/ecs026

Härme, Juho and Mikhail Mikhailov. 2016. From Russian to Finnish and back: compiling Russian–Finnish–Russian parallel corpora. In *Translation from / into languages of limited*

diffusion 3, ed. Lubica Medvecká, 139–147. Bratislava: The Slovak society of Translators of Scientific and Technical literature.

Johansson, Stig. 2007. *Seeing through multilingual corpora on the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins.

Jurkiewicz-Rohrbacher, Edyta. 2019. *Polish verbal aspect and its Finnish statistical correlates in the light of a parallel corpus.* Ph.D. dissertation. Helsinki: University of Helsinki.

Kenning, Marie-Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In *The Routledge Handbook of Corpus Linguistics,* eds M. McCarthy & A. O'Keeffe. London ; New York, NY : Routledge, 487-500.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation: A Corpus Based Approach*. London and Manchester: St. Jerome Publishing.

Kenny, Dorothy. 2011. Translation units and corpora. In *Corpus-based translation studies: research and applications,* eds Wallmach, K., Kruger, A., & Munday, J., 76-102. London: Continuum.

Kilgarriff, Adam. 2013. Using corpora as data sources for dictionaries. In *The Bloomsbury Companion to Lexicography*, ed. Howard Jackson. Bloomsbury, London, 77–96. http://kilgarriff.co.uk/Publications/Kilg_30aug2012.doc?format=raw

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit* (MT Summit X), Phuket, Thailand.

Kraif, Olivier. 2002. Translation alignment and lexical correspondences. In *Lexis in Contrast: Corpus-based approaches*, eds Altenberg, B., & Granger, S. , 271-289. John Benjamins Publishing Company.

Kraif, Olivier. 2003. From Translational Data to Contrastive Knowledge: Using Bi-Text for Bilingual Lexicons Extraction. *International Journal of Corpus Linguistics*, 8(1): 1-29.

Krishnamurthy, Ramesh. 2008. Corpus-driven lexicography. *International Journal of Lexicography*, Vol. 21 No. 3, 231-242. DOI:10.1093/ijl/ecn028.

Kubicka, Emilia. 2019. So-called dictionary equivalents confronted with parallel corpora (and the consequences for bilingual lexicography). *Glottodidactica* XLVI/2. Adam Mickiewicz University Press, Poznań, 75-89. DOI: 10.14746/gl.2019.46.2.05

Marco, Josep. 2019. Living with parallel corpora: The potentials and limitations of their use in translation research. *Parallel Corpora: Creation and Applications,* eds Irene Doval and Maria Teresa Sánchez Nieto, 39–56. Amsterdam/Philadelphia: John Benjamins.

McEnery, Tony & Richard Xiao. 2007. Parallel and comparable corpora: What are they up to? In *Incorporating Corpora: the Linguist and the Translator*, eds Gunilla Anderman & Margaret Rogers, Clevedon: Channel View Publications.

McEnery, Tony, and Richard Xiao. 2010. *Corpus-Based Contrastive Studies of English and Chinese*. London and New York: Routledge.

Mikhailov, Mikhail. 2021 (forthcoming). God, Devil and Christ: A corpus-based study of Russian formulaic idioms and their English and Finnish translation equivalents In *Formulaic*

language: Theories and methods, eds Aleksandar Trklja and Łukasz Grabowski. Berlin: Language Science Press.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies: a guide for research.* London and New York: Routledge.

Mikhailov Mikhail, and Juho Härme. 2015. Parallelnyje korpusa hudožestvennyh tekstov v Tamperskom universitete. (=Parallel corpora of fiction texts at the University of Tampere). *Russkij jazyk za rubežom. Spetsvypusk*, 16-19.

Ooi, Vincent. 1998. *Computer corpus lexicography.* Edinburgh: Edinburgh Univ. Press.

Östling, Robert and Jörg Tiedemann. 2016. Efficient Word Alignment with Markov Chain Monte Carlo. *The Prague Bulletin of Mathematical Linguistics (PBML)*, Number 106: 125–146. http://ufal.mff.cuni.cz/pbml/106/art-ostling-tiedemann.pdf

Perdek, Magdalena. 2012. Lexicographic potential of corpus equivalents: The case of English phrasal verbs and their Polish equivalents . In *Proceedings of the 15th EURALEX International Congress.* 7–11 August, 2012, Oslo, 376-388.

Pöppel, Ludmila. 2018. The construction *возьми и + V IMP*: a corpus-based study. *Zeitschrift für Slawistik*, 63(1), 111–119.

RG 1980. = Švedova, Natalja Ju. (ed.). *Russkaja grammatika.* [Russian grammar]. Moscow: Nauka. http://www.rusgram.narod.ru/.

Ramón, Noelia and Belén Labrador. 2008. Translations of '-ly' adverbs of degree in an English-Spanish Parallel Corpus. *Target*, 20:2, 275–296. doi 10.1075/target.20.2.05ram

Salkie, Rafael. 2008. How can lexicographers use a translation corpus? In *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies. Xiao,* eds *Richard, Lianzhen He and Ming Yue.* Hangzhou: Zhejiang University. http://www.lancaster.ac.uk/fass/projects/corpus/UCCTS2008Proceedings/papers/Salkie.pdf

Salkie, Rafael. 2002. Two types of translation equivalence. In *Lexis in contrast,* 51#7. eds B. Altenberg & S. Granger. Amsterdam: John Benjamins.

Sinclair, John. 2001. Data-derived Multilingual Lexicons. *International Journal of Corpus Linguistics*, 6 (Special Issue): 79–94.

Štichauer, Pavel & Petr Čermák. 2016. Causative constructions of the hacer / fare + verb type in Spanish and Italian, and their Czech counterparts: a parallel corpus-based study. *Linguistica Pragensia* 2, 7-20.

Tarp, Sven. 2020. A dangerous cocktail: databases, information techniques and lack of visions. In *Studies on Multilingual Lexicography*, eds María José Domínguez Vázquez, Mónica Mirazo Balsa and Carlos Valcárcel Riveiro, 47-66. De Gruyter.

Terryn, Ayla Rigouts and Véronique Hoste and Els Lefever. 2020. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Lang Resources & Evaluation,* 54, 385–418. https://doi.org/10.1007/s10579-019-09453-9.

Teubert, Wolfgang 2001. Corpus Linguistics and Lexicography. *International Journal of Corpus Linguistics*, 6 (Special Issue), 125–153.

Tiedemann, Jörg, 1997. Automatical Lexicon Extraction from Aligned Bilingual Corpora. M.A. thesis. Department of Linguistics, University of Uppsala / Otto-von-Guericke Universität Magdeburg.

Tiedemann, Jörg. 1998. Extraction of Translation Equivalents from Parallel Corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Center for Sprogteknologi, Copenhagen.

Tiedemann, Jörg. 2004. Word to word alignment strategies. In *Proceedings of Coling 2004*, 212-218. http://stp.lingfil.uu.se/~joerg/published/coling04.pdf

Tiedemann, Jörg. 2012, Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Walter, Elizebeth, 2010. Using corpora to write dictionaries. In *The Routledge Handbook of Corpus Linguistics.* Eds. M. McCarthy & A. O'Keeffe. London ; New York, NY: Routledge, 428-443.

Zakharov, Viktor, and Svetlana Bogdanova. 2020. *Korpusnaâ lingvistik*a. Sankt-Peterburg: SPbGU.

Zalizniak Anna A., Denisova G. V., Mikaeljan I. L. 2018. Russkoe kak-nibud′ po dannym parallel′nyh korpusov. In *Komp'ûternaâ lingvistika i intellektual'nye tehnologii: po materialam meždunarodnoj konferencii «Dialog 2018»*, Moskva: RGGU.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. 2016. The United Nations Parallel Corpus. In *Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia, May 2016. https://conferences.unite.un.org/UNCORPUS/Content/Doc/un.pdf