Piotr Pęzik
Uniwersytet Łódzki

Exploring phraseological equivalence with Paralela

Zastosowanie korpusu Paralela w badaniach ekwiwalencji frazeologicznej

Streszczenie

Paralela to udostępniony niedawno korpus równoległy polsko-angielskich i angielsko-polskich tłumaczeń. Korpus zawiera aktualnie ponad 260 milionów segmentów słów (blisko 11 milionów segmentów tłumaczeniowych) w wersji polskiej. Dla korpusu opracowano również dostępną w postaci aplikacji WWW wyszukiwarkę (http://paralela.clarin-pl.eu), której składnia umożliwia formułowanie zapytań o pojedyncze wyrazy, frazy oraz wzorce leksykalno-gramatyczne. Możliwe jest także filtrowanie wyników według kryteriów typologicznych i bibliograficznych oraz ich eksportowanie w postaci arkuszy kalkulacyjnych. Szczegółowa zawartość korpusu, zarówno na poziomie metadanych jak też samych tekstów, może być przeglądana za pomocą specjalnego modułu wyszukiwarki.

Po przedstawieniu zawartości korpusu oraz funkcjonalności wyszukiwarki omówiono zastosowanie tych narzędzi w badaniu idiomatyczności tłumaczeń. W tym celu wprowadzone zostało pojęcie ekwiwalencji frazeologicznej, czyli tendencji do zachowania określonego poziomu idiomatyczności tekstu tłumaczenia. Zjawisko to polega na stosowaniu utrwalonych w języku tłumaczenia odpowiedników wielowyrazowych idiomów, kolokacji i innych jednostek frazeologicznych jako ekwiwalentów występujących w języku oryginału połączeń wyrazowych o podobnym statusie frazeologicznym. W tłumaczeniu nieidiomatycznym ekwiwalentami jednostek frazeologicznych są syntagmy, czyli doraźne połączenia wyrazów, których znaczenia są analizowane przez odbiorców tekstów poprzez dekompozycję, a nie częściowo lub całkowite przywoływane z pamięci poprzednich użyć, jak to się dzieje w przypadku jednostek frazeologicznych. Mimo iż tłumaczenie za pomocą kompozycyjnych odpowiedników jest czasami nieuniknione, to niska idiomatyczność całego tłumaczenia (w porównaniu z tekstem oryginalnym) może znacznie utrudniać jego przetworzenie w sensie psycholingwistycznym, a także zwiększa jego wieloznaczność. Zasada ta dotyczy szczególnie tekstów z gatunku użytkowych, naukowo-dydaktycznych i prasowych, w których pojawiają się frazemy

mające w podobnym stopniu utrwalone odpowiedniki frazeologiczne w języku tłumaczenia. Szczególnych trudności z zachowaniem porównywalnego stopnia utrwalenia frazeologicznego w oryginale i tłumaczeniu mogą nastręczać kolokacje, które w odróżnieniu od idiomów czystych i figuratywnych nie muszą się cechować całkowitą lub częściową niekompozycyjnością.

Na przykładzie korpusu Paralela staram się wykazać, że o ile lokalna ekwiwalencja frazeologiczna może być badana na poziomie pojedynczego tłumaczenia, o tyle występowanie frazeologicznej ekwiwalencji globalnej (czyli skonwencjonalizowanego stosowania ekwiwalentów frazeologicznych między parą języków) można badać jedynie, opierając się na odpowiednio dużych korpusach równoległych. Tezę tę ilustruję przykładami wybranych idiomów figuratywnych, które występują w korpusie Paralela, zaczerpniętymi z profesjonalnych i amatorskich tłumaczeń.

Keywords: parallel corpus, Polish, English, phraseology, equivalence
Słowa kluczowe: korpus równoległy, język polski, język angielski, frazeologia, ekwiwalencja

1. Introduction

A new parallel Polish-English corpus called *Paralela* has recently become available as part of the CLARIN-PL infrastructure of Polish language tools and resources. In this paper, I describe the current contents of this corpus and its dedicated search engine. I also attempt to show the usefulness of *Paralela* in the study of the idiomaticity of English-Polish translations. I conclude that large parallel corpora for which such specialized search tools are available are indispensible in investigating the phenomenon of global phraseological equivalence in translation.

2. The corpus

Paralela can be described as an open-ended, opportunistic parallel corpus of Polish-English and English-Polish translations. It currently contains 262 million words in 10,877,000 translation segments. When selecting the translations to be included in the corpus, we initially focused on large, publicly available multilingual text collections and open-source parallel corpora, in order to quickly build a sizeable collection, which could be used to develop and test a new parallel corpus search engine. The main sources of texts included in the corpus are listed in Table 1. The largest of these are the automatically aligned Polish-English subsets imported from the *OPUS* collection (Tiedemann, 2009), which include: the *JRC Acquis Communautaire*, *Open Subtitles*, *European Parliament Proceedings*, *EU Books* and *EMEA* corpora.

Subcorpus	Segments	Words	Alignment
JRC-Acquis	3 385 142	72 88 7270	Automatic
RAPID	3 952 181	66 304 435	Automatic
Open Subtitles	13 628 985	63 048 392	Automatic
CORDIS	761 057	17 162 287	Automatic
EP Proceedings	693 139	13 026 414	Automatic
EU Books	657 938	11 596 443	Automatic
EMEA	825 922	8 883 601	Automatic
114 Literary Classics	448 957	6 292 789	Manual
ESO	74 852	1 447 958	Automatic
OSW	60 363	1 335 858	Manual
Academia	17 750	317 426	Manual
Total	10 877 301	262 302 873	

Table 1. Current contents of the *Paralela* corpus. Word counts were calculated for the Polish segments only

We have also crawled a number of public domain websites including the European Commission Press Release database (RAPID)[1], the Community Research and Development Information Service (CORDIS)[2] and the European Space Observatory website (ESO)[3]. The Polish-English texts acquired from these websites were automatically aligned using the mALIGNa tool (Jassem, Lipski, 2008). In addition to these large, statistically aligned collections, *Paralela* contains a much smaller, but nevertheless significant number of manually aligned texts obtained from the publishers of *Academia* (a popular science journal) and the Center for Eastern Studies. Last but not least, 114 Polish-English and English-Polish translations of public domain literary classics were manually aligned and included in the corpus. The full list of these sources is provided in the 'Browse' section of the *Paralela* website (http://paralela.clarin-pl.eu). The ten largest books from this subset of the corpus are listed in Table 2 below:

Source	Segments	Words
Potop	28 301	430 143
David Copperfield	22 710	319 289
The Pickwick Papers	18 840	269 701
Ogniem i mieczem	16 515	247 887
Faraon	14904	200 035
Villette	11448	197 673
Great Expectations	10850	178 762
Quo Vadis	10252	170 850
Sons and Lovers	16297	164 567
Jane Eyre	10421	164 004

Table 2. Examples of the 114 manually aligned literary classics indexed in *Paralela*

1 http://europa.eu/rapid.

2 http://cordis.europa.eu.

3 http://www.eso.org.

Manual annotation of these texts was a time-consuming task. After developing a special web application called Mantel, we assigned them to trained annotators in order to have them aligned at the level of sentences. The annotators were instructed to use one of the following alignment markers of equivalence between source and target text sentences:

1. *Simple* – used to mark simple sentence to sentence equivalence
2. *Merge/Split* – used to mark many-to-one or one-to-many alignments wherever more than one sentence was translated into many sentences or vice versa
3. *Insertion/Deletion* – to mark 'extra' sentences in either the source or translation
4. *Crosslink* – used to mark equivalent sentences separated by one or more intervening segments
5. *Composite* – used to mark many-to-many segment blocks with overlapping sentence to sentence equivalence relations
6. *Compression* – used to mark complex mergers where several sentences are translated into significantly fewer sentences
7. *Paraphrase* – a last resort marker used to mark significant adaptations or paraphrases in the translation which could not be reasonably mapped at the level of individual sentences.

It is important to note that, in many cases, we had no way of knowing which edition of a particular classic novel was used by the translator. This may explain the high incidence of complex alignment types in texts which had several considerably different editions.

The complexity of the manual alignment procedure is illustrated in Figure 1, which shows the alignment of the first 9 sentences of *The Adventures of Tom Sawyer* in the user interface of Mantel. There are seven simple alignments in this section, one deletion and one split. In total, more than 500,000 segments were manually aligned with this tool and included in the *Paralela* corpus.

The aligned texts were stored in a relational database, part-of-speech tagged and then indexed by the *Paralela* search engine, which was implemented using the Apache Solr library. In addition to the alignment mark-up, a number of bibliographic and taxonomic metadata annotations are stored for most texts in the index. The current list of the metadata fields available for each parallel segment in the corpus database is shown in Table 3.

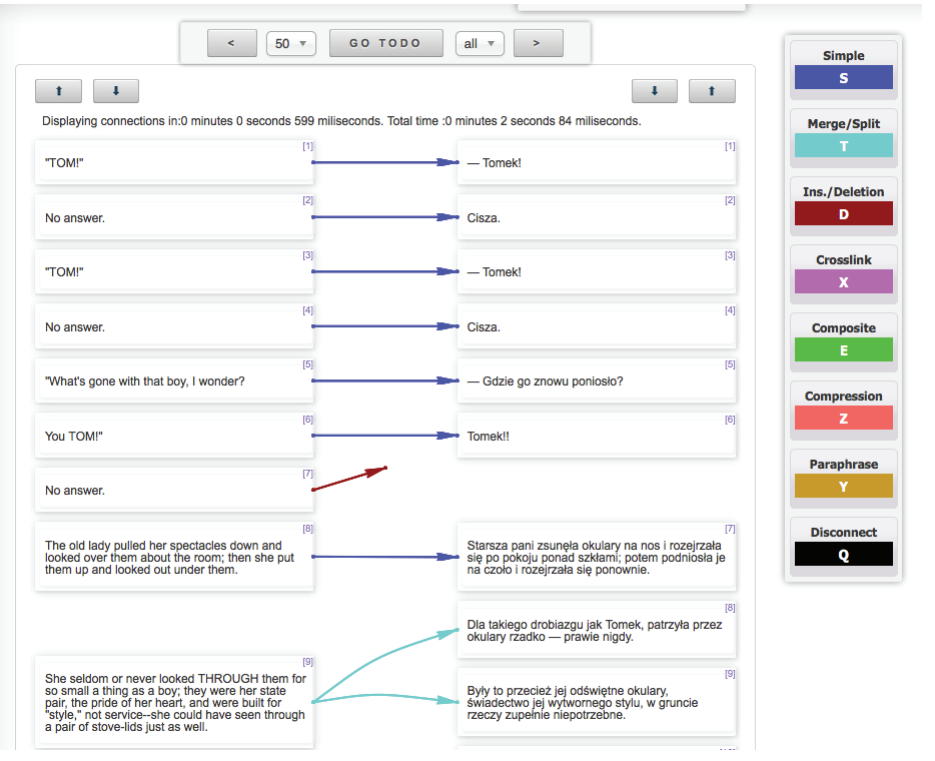

Figure 1. Manual alignment of *The Adventures of Tom Sawyer* in Mantel

Field name	Explanation	Example value
Id	Segment identifier	gkn9r
text_id_pl	Polish text identifier	ceae
text_id_en	English text identifier	Tja5
alignment_mode	Mode of alignment	MANUAL
lang_src	Source language	eng
lang_trg	Target language	pol
seq	Sequence in text	140
source	Source identifier	houndbaskervilles1
genre	NKJP genre tag	typ_lit_proza
medium	NKJP medium tag	kanal_ksiazka
word_total	Segment size	47
alignment_type	Type of alignment	SIMPLE
title_m_pl	Polish monograph title	Pies Baskerville'ów
title_a_pl	Polish section title	Przeklęty ród
title_m_en	English monograph title	The Hound of the Baskervilles
title_a_en	English section title	The Curse of the Baskervilles
authors_en	English authors	Arthur Conan Doyle

Table 3. Searchable metadata fields in the *Paralela* index

As further explained in the next section of this paper, all of the metadata fields listed in Table 3 can be used as additional metadata filters for corpus span queries. There are also some additional unexposed metadata fields, which are only used internally for corpus maintenance purposes.

3. Search engine and query syntax

Paralela supports the SlopeQ query syntax, which has been used in previous corpus projects, such as Spokes (Pęzik, 2014) and the Monco search engine (http://monitorcorpus.com). The scope of the syntax is illustrated in Table 4 below. Apart from basic surface form queries for single words, it is possible to search for loosely defined phrases with the so-called slop factor and lexico-grammatical patterns matching morphosyntactic codes.

#	English query	Matches translation segments containing
1	popular	The exact word form 'popular'
2	popular with	The exact phrase 'popular with'
3	popular with|among	Either of the two exact phrases: 'popular with' / 'popular among'
4	strike**	Different forms of the lemma 'strike' (both nouns or verbs)
5	strike** a balance	Phrases with different forms of 'strike' followed by the sequence 'a balance'
6	strike**|!striking a balance	Same as above, but not when strike** takes the form of 'striking'
7	(strike** a balance)=3	Same as above, but with up to 3 unspecified words between the query terms, e.g. 'struck a very delicate balance'
8	(strike** a balance)~3	Same as above, except that the query terms may occur in any order
9	(strike** balance**|deal**)=4	Co-occurrences of different forms of the lemmas 'strike' and 'balance' or 'deal'
10	word**|story** has it that	Different variants of the multiword expression 'word (or story) has it that…'
11	<lemma=strike tag=n.*>	Different forms of the lemma 'strike' as a noun
12	<tag=j.*> chance**	Co-occurrences of different forms of the lemma 'chance' with immediately preceding adjectives
13	(<tag=v.*> <tag=j.*> discovery**)=2	Sequences of a verb, followed by an adjective and followed by any form of the lemma 'discovery' with up to two word tokes in between

Table 4. *Paralela* supports the SlopeQ corpus query syntax

It is possible to specify bilingual SlopeQ queries for pairs of aligned segments as illustrated in Table 5 below. The first three of these queries are examples of how

one could search for fully specified formal lexical and phraseological equivalents of original words and phrases.

#	English query	Polish query	Matches translation segments containing
1	chance**	nadzieja**	Any inflectional form of 'nadzieja' as a possible equivalent of the lemma 'chance'
2	<tag=j.*> chance**	<tag=j.*> szansa**	Any form of 'nadzieja' (when it is preceded by an adjective) as a possible equivalent of 'chance' (similarly pre-modified by an adjective)
3	(give** to understand)=3	(dać** do zrozumienia)=3	A relaxed co-occurrence of the phrase 'dać do zrozumienia' when it is an equivalent of 'give someone to understand'.
4	(reach** agreement**)~3	(<tag=v.*> porozumienie**)~3	Verbs co-occurring with the noun 'porozumienie' when they are possible equivalents of the English collocation 'reach an agreement'
5	(give** no reason to <tag=v.*>)=3	powód**	An English lexico-grammatical pattern when it may be translated as a phrase containing the Polish noun 'powód'.

Table 5. Examples of bilingual span queries

The last two examples in Table 5 show how to specify a query which matches partly underspecified equivalents. For example, in query 4 any Polish verb is allowed in the equivalent of the English collocation *reach an agreement* and in query 5 we only specify one obligatory term to find potential equivalents of an English multiword expression. All corpus concordances generated with *Paralela* can be exported as Excel files for offline use.

4. Query-based word alignment

The *Paralela* search engine supports query-based word alignment. Once a monolingual query is entered, possible lexical equivalents of the original query terms are computed and ranked using the Dice coefficient (Dice, 1945). The highest scoring matches are then highlighted in the spans retrieved from the index. This solution eliminates the need for offline word alignment which would be very costly to compute and update on a regular basis. Word alignment of the results of bilingual queries is more straightforward: the search engine simply highlights the spans matching both parts of the queries in the retrieved concordances.

5. Metadata queries and search facets

It is possible to use a conjunction of a span query and a logical metadata query to filter the results retrieved from the index. Metadata queries can be formulated

using the Apache Solr DisMax syntax[4]. They are always appended as a logical conjunction to the obligatory span query. For example, the following metadata query:

> (genre:typ_lit_proza NOT source:wutheringheights AND (alignment:simple OR alignment:paraphrase) AND wc:[5 TO *])

would limit the results of the span query to segments found in literary prose (except for those from *The Wuthering Heights*), which are either marked as simple alignments or paraphrases and which contain at least 5 words. This kind of filters are particularly useful when a particular source or genre of texts contains a high number of matches of the query and it becomes necessary to explicitly eliminate such sources from the results.

Because such metadata queries can seem quite complicated to many users, we have introduced two features, namely query facets and predefined collections, which provide a similar functionality through the standard controls of the application user interface. Both of those features are shown in Figure 2 below.

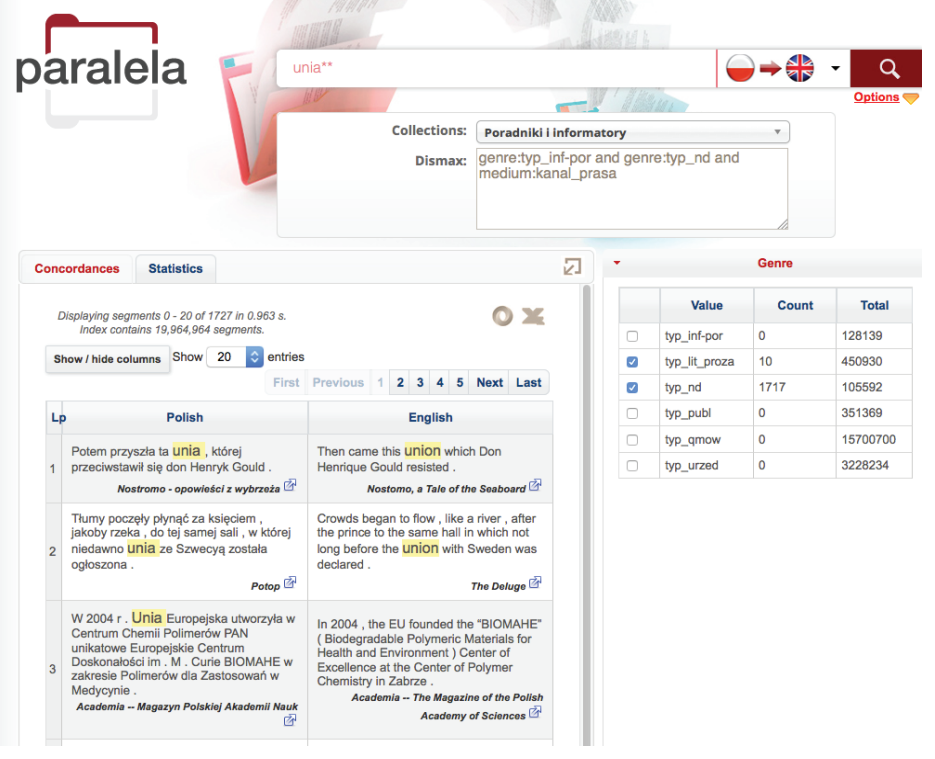

Figure 2. Query facets and predefined collections in *Paralela*

4 See https://cwiki.apache.org/confluence/display/solr/The+Extended+DisMax+Query+Parser. Accessed on 9th January 2016.

Predefined collections are simply a set of metadata queries which users can select from the drop-down list under the corpus query text box. In the example above, the user can select a predefined query which limits the results of the query to texts labeled as 'practical guides' ('Poradniki i informatory') in the corpus taxonomy.

For every query submitted by the user, the *Paralela* search engine also computes a summary of matches found in the different metadata categories in the entire corpus. These summaries are known as 'facets' and they are visualized as pie charts in the *Statistics* section of the results screen. They are also presented in the form of interactive tables as shown in Figure 2. Users can select or deselect some of the categories, thus narrowing down the results of the original span query. In the example above (Figure 2), having obtained a very large number of hits for the query 'unia**' from the *JRC Acquis* section of the corpus, the user decides to deselect all texts which are not marked as literary or scientific works ('typ_lit_proza', 'typ_nd'). This limits the set of matched occurrences of the lemma 'unia' to segments which occurred in such texts.

6. Phraseological equivalence

So far I have introduced the composition of the *Paralela* corpus and the search and exploration features of its search application. In the remaining sections of this paper, I will try to demonstrate that its query syntax is expressive enough and that its current size is sufficiently large to facilitate the investigation of subtle bilingual phenomena such as the idiomaticity of translation and the incidence of phraseological equivalence in English-Polish translations.

An idiomatic translation is sometimes defined as one "which has the same meaning as the source language, but is expressed in the natural form of the receptor language" and in which "the meaning not the form is retained" (Larson, 1984:10; cf. Beekman, Callow, 1974). What makes a translation 'natural' is often language-specific and only indirectly compositional. Given that idioms are prototypical examples of such specificity, it is understandable that the adjective 'idiomatic' is used in this definition to describe this quality of translation. This type of translation idiomaticity can also be viewed as an aspect of dynamic equivalence (Nida, 1964) and it is based on a very general understanding of 'idiomaticity' according to which almost any 'natural' translation could be described as 'idiomatic'.[5]

5 Idiomaticity and formulaicity are often viewed as fundamentally important aspects of 'native-like selection' (Pawley, Syder, 1983) and 'language naturalness' (Sinclair, 1984).

Although such generalizations are useful in that they succinctly express commonly shared intuitions, it is also possible to define an idiomatic translation as one which is characterized by a significant presence of idiomatic expressions which directly correspond to source text phraseological units (PUs). In this view, idiomaticity is understood in a much more restricted sense with PUs as its formal exponents. Normally, translators who encounter lexical or terminological units in the source text may try to translate them into equally conventional target language units to the extent that such simple word-for-word equivalence is justifiable in a give case. Such equivalence becomes more problematic when a non-compositional PU has to be translated. For example, when a figurative idiom found in the source text has no literal equivalent in the target language, it may require a more 'dynamic' translation. Such an equivalent may take the form of a functionally similar figurative idiom which is based on a different metaphor or metonymy, a single word lexical item, or a compositional paraphrase. What makes this rather well-known issue interesting is that some idiomatic equivalents are less 'dynamic' (i.e. more conventionalized and predictable) than they may seem to be in the context of just one translation. The availability of large parallel corpora makes it possible to observe how conventional pairings of source and target language idioms and other types of phraseological units are regularly found in independent translations.

To illustrate this point, let us consider the English idiom "to kill two birds with one stone", which may be translated into Polish as "upiec dwie pieczenie przy jednym ogniu" (lit. "to cook two roasts over one fire"). Looking at a single instance of such a translation, we might be tempted to consider it as a case of dynamic equivalence in that the original idiom has no literal equivalent in Polish, and so the nearest functional equivalent has to be used to ensure a desired level of target text 'idiomaticity'. The figurative meanings of the two expressions are very close and they can be used in similar registers. This translation may therefore work very well, unless the source text idiom is used in some humorous wordplay which takes advantage of its literal meaning.

Let us see how the predictability of this equivalent can be validated against a large parallel corpus. In order to get a sample of naturally occurring Polish translations of the English idiom in question, we could run the following query in *Paralela*:

(kill** bird** stone)=4

This query matches 50 contexts in which the words *kill*, *bird* and *stone* co-occur, with a maximum of four words in between in original English texts. The query may seem a little underspecified, but it is in fact optimized to match slight grammatical variants of the idiom without fetching too many false positives.

Although it is difficult to give an exact figure due to the 'borderline' cases, about 36 occurrences of the English expression "to kill two birds with one stone" were translated as "upiec dwie pieczenie na jednym ogniu". Some of them are shown in Table 6 below.

#	Example	Source
1	How do we kill two birds with one stone? Jak upiec dwie pieczenie na jednym ogniu?	Bottoms up
2	Owner knew he had bad tenants, wanted to kill two birds with one stone? Właściciel wiedział, że ma złych lokatorów i postanowił upiec dwie pieczenie na jednym ogniu.	Pretty Persuasion
3	I figured I could kill two birds with one stone. Zdałem sobie sprawę, że mogę upiec dwie pieczenie na jednym ogniu.	Dance with Somebody
4	Therefore we are in a very positive situation where we can kill two birds with one stone. Jesteśmy zatem w sytuacji, w której możemy upiec dwie pieczenie na jednym ogniu.	Proceedings of European Parliament

Table 6. A selection of predictable phraseological equivalents of the English idiom "to kill two birds with one stone"

Given the regularity with which we find this pairing of idioms in corpora of English-Polish translations, it could be argued that the choice of the Polish equivalent is largely predictable and similar to the way simple lexical and terminological equivalents are selected in other contexts. Should such translations be described as 'dynamic', or rather, as highly conventionalized and thus, in a sense, more formal than dynamic? This may sound like a terminological question, but the conventionality of seemingly dynamic translations is an observation with very practical implications for translators.

Needless to say, phraseological equivalents are not absolute or nearly as predictable as terminological equivalents in technical translation. For example, among the fifty translations of "kill two birds with one stone" there were three independent occurrences of the Polish idiomatic phrase "łapać dwie sroki za ogon" (lit. "to catch two magpies by the tail")[6], a partly formulaic paraphrase "zrobić dwie rzeczy za jednym zamachem" ("to do two things in one go"). There were also a few partly or entirely literal translations and some idiomatic mistranslations. It has to be noted, however, that most of these variants were found in amateur subtitle translations. Table 7 below shows some of these examples.

6 Incidently, this translation could be problematic. The Polish expression "łapać dwie|wiele sroki|srok za ogon" has a predominantly negative connotation of "trying to do too many things at once."

#	Example	Source
1	And kill two birds with one stone. Aha, zabić dwa ptaszki jednym kamieniem?	El Bola
2	Thought I'd kill two birds with one stone, you know. Dwa grzyby w barszcz.	Notting Hill
3	I guess I'll kill two birds with one stone. Chyba upiekę dwa ptaki na jednym ogniu.	Mr. Popper's Penguins

Table 7. Non-conventional phraseological equivalents

The first translation is literal and difficult to justify as such. The phraseological status of the original expression is lost and the Polish translation is certainly not idiomatic. In the second example, an erroneous idiomatic equivalent is used: the Polish idiom "dwa grzyby w barszcz" (lit. "[to put] two mushrooms in the borscht") is normally used to mean "an excess of something". The third example is particularly interesting in that it shows how translators may deal with idiom-based word puns. The line "I guess I'll kill two birds with one stone" comes from the script of *Mr. Popper's Penguins* and it is intentionally ditropic, i.e. its generally figurative meaning is literal in this case. The translation is based on the conventional Polish equivalent of the original idiom, but it also does some justice to the literal meaning of the English original. By replacing the noun *pieczenie* ('roasts') with *ptaki* ('birds'), the translator strikes a delicate balance between achieving phraseological equivalence and saving some of the original word play in the translation. Such a systematic parallel corpus-based analysis of the strategies applied by translators to deal with idiomatic expressions may help us generalize the notion of phraseological equivalence, which I try to define below.

Phraseological equivalence (PE) can be defined as the tendency for translators to use a target language phraseological unit, such as an idiom, a restricted or open collocation as an equivalent of the corresponding source language phraseological unit. Although this tendency is rarely absolute, a low level of phraseological equivalence may result in an insufficient level of idiomaticity of the translation. This in turn may have two negative implications. Firstly, the readers of a non-idiomatic translation may have to invest a larger amount of cognitive effort in understanding it than the readers of the original. Secondly, a non-idiomatic translation may be significantly more ambiguous than the original text, whose meaning is constructed, to the extent that it is idiomatic, from highly conventionalized phraseological units. Furthermore, we can distinguish between local phraseological equivalence between PUs in a particular text and global phraseological equivalence across many different texts of the kind illustrated above, which can only be studied through parallel and reference corpora.

Such corpora have to be sufficiently large to compensate for the fact, that many figurative and pure idioms are relatively rare (Moon, 2001).

Although PE can be regarded as a special type of lexical equivalence, it requires separate consideration, due to the partial compositionality of many phraseological units. The basic difference between lexical and phraseological equivalence boils down to the following observation: when translators encounter an orthographic word, they are quite likely to consider using its institutionalized lexical or terminological equivalent. The non-compositionality of words is a basic fact of derivational morphology (cf. Haspelmath, Sims, 2010: 62). By contrast, combinations of words are more likely to be compositional and translators are more likely to fail to recognize their phraseological prefabrication. In other words, phraseological units are not always as easy to recognize as lexical words. While most idioms, proverbs and speech formulas are relatively easy to spot as such, the conventionalization of restricted and open collocations can be much more subtle. The former types of phraseological units are therefore more difficult to translate idiomatically.

Compared with terminological equivalence, global PE is not usually a fixed one-to-one relation between lexical entities. It may be primarily a one-to-many, many-to-one or many-to-many relation between source and target PUs. For example, the abovementioned English idiom "to kill two birds with one stone" seems to have a Polish equivalent which is much more frequent than any of its alternatives. In many cases, local PE can be null, which means that source language phrasemes are translated as target language syntagmas (i.e. grammatically valid, spurious word combinations with no phraseological status) and vice versa. A high incidence of null PE in a text may result in a non-idiomatic translation. On the other hand, in some cases, null PE may be a conscious and well-justified choice. For example, a formally accurate idiomatic equivalent of a multiword unit may not yet exist in the target language: a regular Polish equivalent of the term "product placement"[7] was only recently established (as "lokowanie produktu") and the highly institutionalized English term "road rage" does not seem to have a stable equivalent in Polish. A quick *Paralela* query shows that it has a variety of similarly likely equivalents such as "agresja na drodze," "gniew na drodze" or "furia drogowa". Also, a context-dependent, humorous use of a ditropic idiom may require a hybrid equivalent of the kind illustrated above (cf. Table 7, example 3). Finally, the translator's attempt to achieve a state of PE may be unsuccessful (cf. Table 7, example 2). In other words, the exact choice of the target language PU is just as important as the recognition of the source language phraseological unit.

7 Although "product placement" could be described as a technical term, it is also a 'phraseological nomination' (Gläser, 1998).

The phenomenon of global PE is particularly subtle in the case of hundreds of thousands of restricted and open collocations which contribute to idiomaticity of the source text. Such word combinations are usually semantically compositional and they can easily be translated into compositional equivalents (cf. Pęzik, 2011, 2012). It takes a large parallel corpus to study the global PE of such items and to observe "the underlying rigidity of phraseology, despite a rich superficial variation" (Sinclair, 1991: 121).

7. Summary and future work

Although currently *Paralela* is not a balanced corpus, it can already be shown to contain a sufficiently large sample of different text varieties to be useful in the analysis of certain equivalence phenomena. The query syntax of its search engine is particularly useful in investigating phraseological equivalence, a notion which I have defined and briefly illustrated in this paper with examples extracted from the *Paralela* corpus. Having developed a scalable search and storage architecture, in the future we will focus on extending the coverage of the corpus. This is particularly important in view of the fact that despite the high incidence of phraseological prefabrication, individual PUs can be too rare to be spotted as particularly recurrent in small corpora.

Acknowledgments

The work described in this paper has been financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education. Apart from the author of this paper, the main developers of the *Paralela* core database and web application were Łukasz Dróżdż, Paweł Wilk and Paweł Kowalczyk.

References

Beekman, John, Callow, John (1974): *Translating the Word of God*. Grand Rapids, MI: Zondervan Publishing House.

Dice, Lee R. (1945): Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297-302. doi:10.2307/1932409.

Gläser, Rosemarie (1998): The Stylistic Potential of Phraseological Units in the Light of Genre Analysis. In: Anthony Paul Cowie (ed.): *Phraseology : Theory, Analysis, and Applications*. Oxford: Oxford University Press, 124–43.

Haspelmath, Martin, Sims, Andrea D. (2010): *Understanding Morphology*. 2nd Edition. Understanding Language Series. London: Hodder Education.

Jassem, Krzysztof, Lipski, Jarosław (2008): A New Tool for the Bilingual Text Aligning at the Sentence Level. In: *Proceedings of 16th International Conference on Intelligent Information Systems*, 279–86.

Larson, Mildred L. (1984): *Meaning-Based Translation: A Guide to Cross-Language Equivalence*. Lanham, MD: University Press of America.

Moon, Rosamund (2001): Frequencies and Forms of Phrasal Lexemes in English. In: Anthony Paul Cowie (ed.) *Phraseology : Theory, Analysis, and Applications*. Oxford: Oxford University Press, 79–100.

Nida, Eugene Albert (1964): *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: Brill Archive.

Pawley, Andrew, Syder, Frances Hodgetts (1983): Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In: Jack C. Richards, Richard W. Schmidt (eds.): *Language and Communication*. London: Longman, 191–225.

Pęzik, Piotr. (2011): Providing Corpus Feedback for Translators with the PELCRA Search Engine for NKJP. In: Stanisław Góźdź-Roszkowski (ed.): *Explorations across Languages and Corpora : PALC 2009*. Łódź Studies in Linguistics. Frankfurt am Main/ New York: Peter Lang, 135–44.

Pęzik, Piotr. (2012): NKJP w warsztacie tłumacza. In: Adam Przepiórkowski, Mirosław Bańko, Rafał Górski, Barbara Lewandowska-Tomaszczyk (eds.): *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN, 301–311.

Pęzik, Piotr. (2014): Spokes – a Search and Exploration Service for Conversational Corpus Data. Paper presented at the CLARIN Annual Conference 2014, Soesterberg, The Netherlands, October 25.

Sinclair, John (1984): Naturalness in Language. *Ilha Do Desterro. A Journal of English Language, Literatures in English and Cultural Studies* 5(11), 45–55.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tiedemann, Jörg. (2009): News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *Recent Advances in Natural Language Processing* 5, 237–48.