Lexicographic potential of corpus equivalents: The case of English phrasal verbs and their Polish equivalents

Magdalena Perdek

Keywords: *phrasal verbs*, *equivalence*, *parallel corpora*.

Abstract

The aim of this paper is to investigate Polish equivalents of English phrasal verbs as found in an English-Polish (E-P) parallel corpus *PHRAVERB*. Given the semantic idiosyncrasy exhibited by phrasal verbs, it is assumed that the equivalents generated by *PHRAVERB* will often differ from those found in E-P dictionaries. The qualitative corpus analysis aims to show that arriving at the desirable Polish counterpart involves a detailed semantic breakdown of the English structure, a careful analysis of the context in which it is used, as well as linguistic and translation skills, necessary to detect the nuances and subtleties of meaning in both languages. *PHRAVERB* is used to analyze the lexicographic potential (LP) of corpus equivalents. Four levels of LP have been established – high, average, low and zero – to evaluate which corpus-derived equivalents are eligible for inclusion in E-P dictionaries. To this end, 2,514 occurrences of PVs in the parallel corpus, with their equivalents, have been identified and analyzed.

1. Introduction

The English phrasal verb is a peculiar union of a verb and a particle (prepositional or adverbial) that often produces a unique meaning, uninferable from the meanings of its constituents. This semantic unpredictability of phrasal verbs (PVs) along with their specific syntactic configurations, poses major problems for the non-native speakers who often consciously choose to avoid using the structures and instead fall back on the synonymous, "safer", Latinate verbs. Adding to the comprehension difficulties is the often stressed informal and colloquial character of phrasal verbs. The widespread conviction that PVs are typical of unofficial discourse contributes to their "pedagogical notoriety" but, at the same time, convinces learners that mastery of phrasal verbs (along with idioms) is the key to achieving the much-desired, native-like command of English. The features described above add up to a vivid picture of a lexical item so concise in form, yet complex in content. It is, therefore, not surprising that PVs might be quite a challenge for both translators and bilingual lexicographers.

2. Parallel corpora in bilingual lexicography and translation

The use of monolingual corpora in compiling monolingual dictionaries has become a standard in the lexicographic practice pioneered by the Cobuild project in the 1980s (Sinclair 1987). Landau (2001: 305) claims that "the corpus is a tool that has breathed new life into the art of lexicography". The main purpose of the corpora was "to ensure authenticity and empirical adequacy in lexicography" (Altenberg and Granger 2002: 33). Thanks to corpora, "lexicographers have become aware of new regularities and systematic behaviours in language use, such as chunking and semantic prosody" (Varantola 2006: 218). Obviously, monolingual corpora can be just as useful in preparing a bilingual dictionary, or at least some part of it. This includes selecting frequent collocational, grammar and usage patterns of the lemma as well as example phrases or sentences, either modified or left unchanged. However, as pointed out by Teubert (2002: 204), "even where bilingual dictionaries record the evidence encountered in monolingual corpora, they still have to rely on the lexicographers' bilingual competence to determine the translation equivalent of any semantic conglomerate". Lexicographers can search for equivalents in other bilingual dictionaries or draw from their own experience and create their own equivalents not yet recorded. Such equivalents will, "under normal circumstances not be wrong. But [they] will not necessarily reflect the translation practice", which is exactly what parallel corpora record (Teubert 2002: 204).

Translators, by consulting monolingual language corpora, can find similar contexts to the one they are dealing with in their translation, especially if they need confirmation that they understand the original or if, in the absence of a lexicographic equivalent, they must create their own, based on the numerous contexts the word naturally occurs in. Clearly, in hunting for equivalents, a bilingual corpus seems like the best solution. Examining an original with its translation(s) does not just offer ready-made equivalents (which can, of course, be questioned and rejected), but simultaneously gives essential information on the lexical item to be translated. Therefore, as Teubert (2002: 193) puts it, "parallel corpora are repositories of translation units and their equivalents in the target language" to be re-used in subsequent translations.

Atkins and Rundell (2008) use *parallel corpora* as the umbrella term for *translation* and *comparable corpora*. In their approach, a translation corpus is one in which a set of texts is translated into another language.[1] A comparable corpus consists of "two individual language corpora, selected on the basis of at least one shared parameter, usually the subject matter, together with possibly other properties shared by the texts, such as date and/or the medium (books, newspapers, conversations etc.)" (Atkins and Rundell 2008: 479).[2] Figure 1 illustrates the difference between the two types of bilingual corpora.

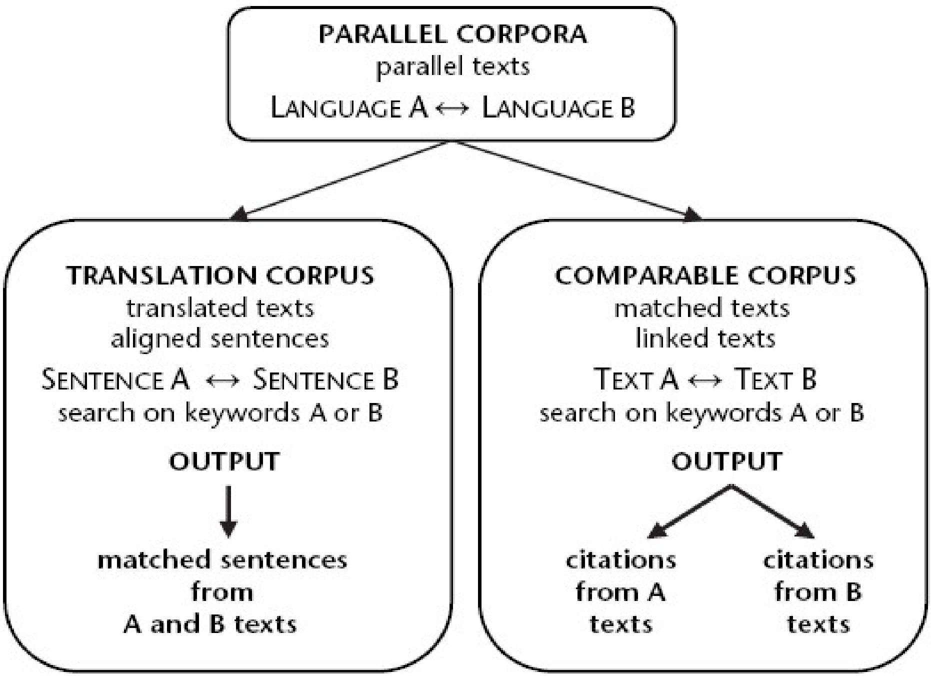

Figure 1. Two types of bilingual corpora (after Atkins and Rundell 2008: 477).

Parallel corpora can be used to analyze different levels of language – lexis, syntax as well as discourse. "By facilitating the mapping of correspondences between languages, parallel corpora can not only shed light on the commonalities and differences between language pairs, but also improve the accuracy of descriptions of individual languages"

(Kenning 2010: 493). Since both bilingual lexicography and translation start from meaning, parallel corpora may also be used "to establish contrastive lexical-semantic fields. This is done by going back and forth between translations. Words which share a number of translations are semantically close (…). In this way we can establish systems and subsystems based on the strength of the equivalents" (Aijmer 2008: 278).

As mentioned before, the advantage of using parallel corpora in bilingual lexicography and translation practice is that they provide syntagmatic data about lexical items in both languages but most importantly, they generate ready-made equivalents, or rather, candidates for equivalents to appear in a bilingual dictionary or a translated text. Atkins and Rundell (2008: 478) list pros and cons of using translation corpora in compiling a bilingual dictionary. On the upside, there is no need for "hunting for equivalence candidates", there is a "wealth a context-sensitive translations" and all equivalence candidates actually have specific contexts. On the downside, there are "too many equivalence candidates [and] every one of them seems essential to lexicographers at that point", which results in oversized dictionaries unfit for printing and entries overloaded with details unimportant for most users. Consequently, the production process of a dictionary is considerably slowed down (Atkins and Rundell 2008: 478). So, even with a ready-made collection of corresponding texts, a lexicographer-cum-translator must interpret "what the corpus returns and select what is relevant" (Béjoint 2010: 369) and this, of course, includes choosing relevant equivalents. The beneficial potential of translation corpora is duly noted. What a parallel corpus gives is

> direct access to equivalents present in the texts comprising the corpus. In cases when counterparts identified in the aligned sections constitute actual lexicographic equivalents, such an access may not only facilitate the process of compiling bi- and multilingual dictionaries which would be based on real equivalents used in translation practice, but also make the process more reliable (Waliński 2005: 43).[3]

Reliability is definitely a desirable quality, but as Rundell and Atkins (2008) pointed out earlier, large amount of much information offered by parallel corpora is not necessarily a good solution for a bilingual dictionary. Since, to date, there are no "bilingual dictionaries of general language based on parallel corpora, we still do not know to what extent they can complement, improve and validate existing dictionaries" (Teubert 2002: 204).[4]

Waliński (2005) further lists some of the implications of applying parallel corpora in translation practice. For example, through the comparison of equivalent linguistic items in the original and translated texts, parallel corpora allow us to verify various hypotheses concerning natural aspects of the translation process and its results (Waliński 2005: 43, cf. Hunston 2002: 128, Kenning 2010: 492). On the other hand, "it is well-known that textual choices often differ depending on the individual translator, and there might be outright errors in translation" (Johansson 2007: 9).[5] The upside of the specific make-up of parallel corpora is the stable structure of the corresponding texts, which, in theory, should facilitate assigning equivalent structures – starting from the text level, through paragraphs and sentences, to phrases and words in the original and the translation (Waliński 2005: 44). However, it often happens that in the process of translation the original structural arrangement is lost (e.g. some paragraphs or sentences get shortened or omitted while others are lumped together into one translation sequence), which automatically rules out alignment of the texts based on structural analogies (Waliński 2005: 44). The linguistic equivalents generated by parallel corpora are also criticized for their unnaturalness resulting from the influence of SL structures on the TL phrases which lead to loss of the naturalness of syntax, phraseology and lexicon in the translated text. (Waliński 2005: 44, cf. Aijmer 2008: 284, Kenning 2010: 492).[6] Another aspect of the translation process revealed by parallel corpora is how translators deal with the lack of straightforward equivalence at word or phrase level (Zanettin 2002: 11, cf. Aijmer 2008: 285-286). Therefore, apart from being repositories of re-usable equivalents (Teubert 2002: 193), parallel corpora are also "repertories of strategies deployed by past translators" (Zanettin 2002: 11, cf. Sinclair 1996: 174, Teubert and Čermáková 2005: 155) in dealing with both equivalence and lack thereof. The extent to which corpus equivalents might prove to be "re-usable" depends on the size and content of the corpora. While large general corpora offer large numbers of equivalents, smaller and more specialized corpora can provide a more detailed picture of lexical units. For example, a compilation of domain-specific bilingual texts can offer more insight into the behavior of the unit in context and its collocational patterns, which inevitably affects the translation strategies and consequently the choice of equivalents.

Parallel corpora are definitely a useful tool in searching for equivalents, one that lexicographers should turn to, given that bilingual dictionaries are not very "instructive" (Teubert and Čermáková 2005: 124), and also because they

> lack the richness of context that occurs in parallel texts; and furthermore they lack the flexibility afforded by using parallel texts, where any number of patterns can be searched for. The student/investigator is not limited to the words and phrases that happen to have been chosen by the dictionary maker. In addition, dictionaries vary greatly in how well they deal with collocational information (Barlow 2000: 114).

However, as Malmkjær (1998: 6) rightly observes, a translation corpus "still only provides, for each instance, the result of one individual's introspection, albeit contextually and contextually informed". The human factor is, then, what both reference resources have in common, which is why

> there is no reason to put dictionaries and parallel corpora in competition since they have different strengths and weaknesses. Whereas it is true that parallel corpora can show more contexts than are possible in dictionaries, they are also full of noise, including incorrect and imprecise translations, and they do not provide the detailed description possible through the introspection of a highly-trained lexicographer. Thus it is toward convergence rather than dominance of one genre or the other that the field should seek to move (Lubensky and McShane 2007: 920).

The obvious solution would be to connect electronic bilingual dictionaries to parallel corpora (of good quality, semantically tagged, syntactically parsed) so that the equivalents (and contexts, however limited) offered by lexicographers can be checked against what is found in the corpora and vice versa. Until it becomes a standard to integrate both resources in such a way, translators and lexicographers alike must rely on both tools separately because. Once other features are also incorporated, for example interactive functions, user-profiles, user-specified filters and display modes (e.g. browser modes, look-up modes) what we will get will be an intelligent bilingual reference work (Varantola 2006: 223).

3. English phrasal verbs and equivalence

While the research on PVs in the monolingual setting is quite copious, including studies on the syntactic (e.g. Sroka 1972), semantic (e.g. Campoy-Cubillo 1997), pragmatic (e.g. Hampe 2002) and cognitive (e.g. Lindner 1983) features of PVs, the bilingual aspect is somewhat neglected as only few studies exist on interlingual equivalence and PVs. Those based on corpus data were conducted by Claridge (2002) with German equivalents and Dezortová

(2010) with Czech equivalents.[7] Other studies, much smaller in scope, which examined the equivalents of PVs were done for Polish (Masiulanis 1974, Kretowicz 2005), Russian (Yatskovich 1999) and Arabic (Aldahesh 2009). Given the semantic complexity of PVs and difficulties with their understanding by non-native speakers of English, it is worth taking a deeper look into the interlingual differences in the rendition of PVs into different languages and discover some of the translation strategies used by various translators.

4. *PHRAVERB* – a unidirectional, English-Polish parallel corpus with an index to PVs

The study of corpora can deepen the understanding of a language by examining usage in a variety of contextual relationships, registers and topics. A standard in monolingual lexicography, corpora are still waiting for their heyday in the realm of bilingual lexicography. Separate monolingual corpora have already been used in the compilation of large bilingual lexicographic works (e.g. *Oxford Hachette French dictionary* or *PWN-Oxford wielki słownik angielsko-polski* to name just a few) but it is parallel corpora that are likely to provide more valuable data not only on interlingual differences and similarities but, most importantly, on equivalence relations and translation techniques. For the purpose of this study an English-Polish unidirectional parallel corpus – *PHRAVERB* – has been compiled to analyze Polish equivalents of English phrasal verbs, and, more specifically to evaluate the lexicographic potential of those equivalents for the purpose of inclusion in future English-Polish dictionaries, both general-purpose and specialized ones. A total of 2,514 occurrences of PVs have been identified in the corpus.

4.1. *Size and content of the corpus*

The parallel corpus consists of 408 English press articles and their Polish translations.[8] They were collected between July 2006 and March 2011. The majority of the English articles (95.08%) have been taken from American websites, with only a small proportion (4.92%) derived from British internet sources. The Polish translations have been taken from three main online sources – www.gazeta.pl, www.interia.pl and www.onet.pl. Table 1 presents the distribution of the sources of the articles.

Table 1. Sources of articles used in the creation of *PHRAVERB*.

<table>
<tr><th colspan="5">American websites</th></tr>
<tr><td>New York Times</td><td>Forbes</td><td>Boston Globe</td><td>Washington Post</td><td>Other</td></tr>
<tr><td>365 articles</td><td>15 articles</td><td>3 articles</td><td>2 articles</td><td>3 articles</td></tr>
<tr><td>89.46 %</td><td>3.67%</td><td>0.73%</td><td>0.49 %</td><td>0.73%</td></tr>
<tr><th colspan="5">British websites</th></tr>
<tr><td colspan="2">The Guardian</td><td colspan="3">Daily Telegraph</td></tr>
<tr><td>15 articles</td><td>3.67%</td><td colspan="2">5 articles</td><td>1.22%</td></tr>
</table>

PHRAVERB contains 926,725 words. Table 2 shows the size and structure of the corpus on both sides.

Table 2. Structure and size of *PHRAVERB*.

	Types	Lemmas	Tokens	Punctuation
English	35,805	33,958	488,941	24,124
Polish	67,225	37,782	437,784	82,831

Types are unique tokens, no lemmatization is performed. Lemmas are unique base forms derived in lemmatization process and distinguished for morphosyntactic category (e.g. open_JJ (adjective) and open_RB (adverb), poza_prep and poza_subst are counted separately). Tokens are understood as non-unique word forms (number of strings between spaces or punctuation). Punctuation tokens include English SENT, GENERAL JOINER and [: , '' ``]; e.g. *I am who I was* (types: 4, lemmas: 3, tokens: 5, punctuation: 1).[9] The corpus can be viewed in any internet browser. Figure 6 shows the corpus format

4.2. *Processing of the texts and alignment tools used*[10]

The processing of textual data used to create *PHRAVERB* involved performing semi-automatic normalization of the characters used in the original (source language – SL) and translated (target language – TL) texts. As for sentence splitting of the SL and TL texts, different procedures were used for English and Polish, relying on abbreviation sets appropriate for each language. The Polish sentence splitter was a modified version of the Lingua::EN::Sentence from search.cpan.org. Titles of the articles were semi-automatically converted into sentences by placing a full stop and inserting a hard line break. *Hunalign* was used to align sentences in corresponding sentence-split text files and default settings were used for alignment.[11] Beads of 1-to-many and many-to-1 were allowed. The output format was explicit – both source and target sentences were specified explicitly rather than by their indexes. The bilingual dictionary used to optimize alignment was a 60k word pair list based on *PWN-Oxford wielki słownik polsko-angielski.* Morphological tagging for Polish was performed using TaKIPI18 tagger (http://nlp.pwr.wroc.pl/takipi/), while for English the morphological tagging was performed using TreeTagger (http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html).

Dawid Weiss' frequency list (http://www.cs.put.poznan.pl/dweiss/research/pv/) was used to automatically extract phrasal verbs. The list contained 991 phrasal base forms. These were used by Weiss to create the list of 12925 verbs by providing all inflectional variants and each inflectional variant with regular expressions allowing for up to 3 words between the verb and the preposition/particle. Additionally, phrasal verbs not included in Weiss' list had been detected while reading the articles and were subsequently added to the index of phrasal verbs.

4.3. *Limitations of the corpus*

Even though the corpus generated over 2,500 instances of PVs, a bigger corpus could undoubtedly yield even larger sample, which would most likely influence the distribution of the lexicographic potential. Another limitation of the corpus is the origin of the texts, i.e. primarily American press, which leaves the British variation underrepresented, especially in view of the fact that there is no definitive research confirming that PVs are more prevalent in American English as some scholars ventured to claim (Mencken [1919] 1969: 199, Vallins 1956: 130, Lipka 1972: 161, 229, cf. Meyer 1975). At the time of the compilation of *PHRAVERB*, most of the translations found on the Polish sites published mainly articles from the American press. Additionally, the fact the only journalistic texts were used in the corpus might limit the representativeness of the corpus as genre variation clearly enhances the quality of data obtained from the corpus. Finally, individual styles and strategies used by particular translators might also bias the results. However, the translator's name has not been provided in 68% of the texts so it was difficult to notice any idiosyncrasies and recurring patterns in the translations.

5. Lexicographic potential (LP) of corpus equivalents

The main purpose of the corpus analysis is to evaluate the *lexicographic potential* (LP) of the non-lexicographic equivalents found in *PHRAVERB*. Lexicographic potential is understood here as the eligibility of the equivalent for inclusion in an E-P dictionary, based on its accuracy and applicability in various contexts and with different arguments. Four levels of LP have been established – high, average, low and zero LP.

Before assigning the LP, all of the occurrences have been analyzed to see which ones were rendered using equivalents found in 17 E-P dictionaries published between 1997 and 2009. The analysis revealed that lexicographic equivalents have been used in 1,420 sentences (56.48%). Phrasal verbs have been omitted in translation in 330 cases (13.13%).[12] That leaves 764 (30.39%) phrasal verbs translated with non-lexicographic equivalents. Table 3 presents the types of translations found in *PHRAVERB*.

Table 3. Types of translations in *PHRAVERB*.

Type of translation	Number of occurrences	Percentage
Lexicographic equivalents	1,420	56.48%
Non-lexicographic equivalents	764	30.39%
Omissions (type 1)	281	11.18%
Omissions (type 2)	49	1.95%
Total number of occurrences	2,514	100%

It is not surprising that the majority of the translated PVs have been rendered with lexicographic equivalents. Excluding the omissions, in 65% of the cases, a dictionary equivalent has been implemented. That is not to say, of course, that the translators actually looked it up in a dictionary since they might simply know the right rendition based on their translation competence but the fact that, if needed, 65% of the equivalents is already recorded in E-P dictionaries gives reasons to be optimistic about the quality of E-P lexicographic works. From a lexicographic perspective, however, the most interesting cases are those where translators used non-lexicographic equivalents for which lexicographic potential may be applied in order to evaluate their eligibility for inclusion in future reference works.

5.1. *High LP criteria*

At this level, the corpus equivalent (either a verb or a verb phrase) must be synonymous to the lexicographic equivalent(s) and can be used in a considerable number of contexts or with the most common arguments. Table 4 shows selected examples of high LP.

Table 4. Examples of corpus equivalents with high LP.

	English	Polish	Lexicographic potential
1.	Though the Nazis drew up invasion plans (…) they never **acted on** them. (Bunker)[13]	Choć hitlerowcy planowali inwazję na Szwajcarię, (…) ostatecznie nigdy tych planów nie **zrealizowali.**	High LP. *Zrealizować* (‘to realize’) conveys the meaning of the phrasal verb with objects like *plan/ pomysł/ zamiar* (‘plan/idea/intention’)

2.	But before anyone could **act on** this impulse, the rules of jihadi etiquette kicked in.(Jihad)	Jednak zanim ktokolwiek **zrealizował** ten pomysł, zadziałała etykieta dżihadu	High LP. Change of objects from *impuls* ('impulse') to *pomysł* ('idea') in the translation, which are not interchangeable. However, with *pomysł* the equivalent works.
3.	The contest last year, asking consumers to finish incomplete vignettes in the long-running "Priceless" campaign, "had 100,000 entries, which **blew us away**," Mr. Jogis said… (Oscars)	Ubiegłoroczny konkurs, w którym konsumenci mieli dokończyć zdania wykorzystywane w kampanii pod tytułem "Priceless", miał 100 tys. Uczestników. Taka ilość **powaliła nas na kolana** - mówi Jogis	High LP. This idiomatic expression is similar to *zwalić z nóg* (literally 'to knock down') when talking about something that surprises us a great deal not necessarily in a good way.
4.	He says the education system is conservative, and **bogged down** with ideology. (Afraid)	Cały system nauczania jest konserwatywny i **obciążony** ideologią - dodaje	High LP. The participle *obciążony* ('burdened') can collocate with some abstract nouns like *praca/obowiązki* ('work/duities') and can also be used in the active voice.
5.	Its cost, (…) is widely regarded as a central impediment to **bringing down** the French structural deficit and a significant drag on French competitiveness. (Sarkozy)	Jej koszt, (…) jest powszechnie uważany za główną przeszkodę na drodze do **zredukowania** deficytu strukturalnego Francji i obciążenie dla jej konkurencyjności na gospodarczej arenie.	High LP. *Zredukować* ('to reduce'), as a synonym to *zmniejszać* ('to reduce'), could be used in many contexts, especially with abstract nouns like *wydatki/ inflacja/ bezrobocie* ('expenses/inflation/unemployment')
6.	…when they learned that she and her siblings were **growing up** without religion. (Atheist 33) ...who **grew up** in the well-to-do Sherman Oaks section of Los Angeles…(Diary 4) …as far away as possible from the village in Bavaria where she had **grown up.** (Nazi 54)	…gdy dowiadywali się, że ona i jej rodzeństwo **wychowują się** bez religii. …która **wychowywała się** w zamożnej dzielnicy Los Angeles, Sherman Oaks …jak najdalej od wioski w Bawarii, w której **się wychowała.**	High LP. This use of *wychowywać się* ('to be raised') is well established in Polish but, surprisingly, none of the E-P dictionaries records it. Adding the phrase **to grow up in (some place)** to the entry seems justified.

5.2. *Average LP criteria*

The corpus equivalent is semantically similar to the lexicographic equivalent(s) but its scope is limited due to structural differences or selection of arguments. At this level, changes in the structure of the equivalent, i.e. everything other than a verb are included. This includes sentences where the Polish translation contains a part of speech (usually a noun but it can also be an adjective) which is morphologically linked with a verb of the same meaning, e.g. *realizować – realizacja* ('to realize – realization'), *odchodzić - odejście* ('to depart – departure'). Table 5 shows selected examples of average LP.

Table 5. Examples of corpus equivalents with average LP.

	English	Polish	Lexicographic potential
1.	This, of course, is not exactly how things **turned out.** (Blavatnik 22)	Oczywiście, sprawy **przybrały** nieco inny **obrót.**	Average LP. *Przybrać obrót* ('to take a turn') would always need a subject like *sprawy* ('affairs'), *wydarzenia* ('events') or the name of some specific event.

2.	In fact, the subprime-mortgage crisis was the first severe market downturn since online trading **took off** here…(Home 19)	W rzeczywistości ostatni kryzys rynku kredytów hipotecznych by pierwszym poważnym kryzysem odkąd w Japonii **zaczął się boom** na inwestowanie online	Average LP. *Boom* is a foreign word in Polish but already well established, especially in journalistic discourse where it would be suitable. It would have to be used with a verb like *zacząć się* ('to begin') or *nastać* ('to increase') to convey the sense of beginning. There is also a shift of the noun from subject to object position.
3.	I ask him how his life has changed since Facebook **took off**. (Facebook 66)	…więc pytam go, co się zmieniło od momentu, kiedy Facebook **stał się** naprawdę **popularny.**	Average LP. *Stać się popularnym* ('to become popular') or *zyskać/zdobyć popularność* ('to gain popularity') convey the meaning of being successful. However, it could only be used with names of products or *pomysł* but not with firma.
4.	For travelers who (...) carry around sensitive data, it is worth **looking into** programs like (...) LoJack…(Surf)	Podróżni, którzy przewożą ze sobą (…) istotne dane, powinni **zainteresować się** takimi programami jak LoJack…	Average LP. *Zainteresować się czymś* ('to become interested in sth') conveys the meaning of the phrasal verb, especially if it is used in the progressive. An example phrase of **to be looking into sth** would have to be included in the dictionary entry.

5.3. *Low LP criteria*

The corpus equivalent is a translation of the definition, which can be used in a limited number of contexts with arguments determined by the manner of translation, usually resulting in some degree of under- or overspecification of the original meaning. Table 6 shows selected examples of low LP.

Table 6. Examples of corpus equivalents with low LP.

1.	Relations with Rome grew colder as the **calls for** him to **stand for** the presidency mounted. (Lugo)	Stosunki z Rzymem uległy dalszemu ochłodzeniu po tym, jak odezwały się głosy, by **spróbował sił** w wyborach prezydenckich	Low LP. *Spróbować sił w* ('to try one's hand at sth') conveys the sense of trying something out rather than serious consideration but in the political context it might be synonymous with *kandydować* ('to run for').
2.	He should **spell out** all the ways America will guarantee Israel's security. (Jerusalem)	Powinien **wyraźnie wymienić** wszystkie środki, którymi Ameryka będzie gwarantować izraelskie bezpieczeństwo…	Low LP. The adjective *wyraźnie* ('clearly') pertains to the meaning of the phrasal verb but *wyraźnie wymienić* ('to clearly list sth') is not a natural collocate in Polish. A better verb to go with the adjective would be *nakreślić* ('to outline') or *zaznaczyć* ('to indicate')
3.	It's true that firms **scaled back** the corporate excesses, like fancy retreats and private jets, for which they were vilified as a brutal recession gripped the country.	Prawdą jest, że firmy **powściągnęły** korporacyjne ekscesy, takie jak wymyślne wycieczki i prywatne odrzutowce, za które to zachcianki spadła na nie zmasowana krytyka, gdy brutalna recesja zamknęła Amerykę w swoich kleszczach	Low LP. *Powściągać* ('to stop') collocates with *wybryki* ('pranks', 'frolics') which is a synonym for *ekscesy* ('excess') but cannot be used with many other objects like *import, wydatki, płace* ('import, expenses, pay') mentioned in E-P dictionaries.

4.	…the heads of girls continually **pop up** from narrowly constructed 10-foot shafts. (Silicon)	…dziewczęce głowy **wynurzające się** raz po raz z wąskich, trzymetrowych biedaszybów.	Low LP. The participle *wynurzające się* (‘looming’) does refer to appearing but it has considerable collocational restrictions as not everything can be the subject of *wynurzyć się* (‘to loom’)

5.4. *Zero LP*.

The corpus equivalent is limited to a singular context without any possibility of extending its scope to a wider range of contexts or arguments. All equivalents resulting from mistranslations are treated as having zero LP.

6. Results and discussion

The non-lexicographic equivalents constitute 30.39% (764) of the total occurrences of PVs. The analysis based on the LP criteria has shown that over half (54.45%) of the instances can be classified as having zero LP, therefore exhibiting no lexicographic value. Table 7 presents detailed distribution of the LP levels.

Table 7. LP levels of non-lexicographic translations of PVs.

	High LP	Average LP	Low LP	Zero LP	Total
Occurrences	116	100	132	416	764
Percentage	15.18%	13.09%	17.28%	54.45%	100%

When it comes to the number of different equivalents used in translations, as many as 717 have been identified and the distribution of LP is similar, with the zero LP equivalents constituting nearly 60% of all the equivalents. Table 8 shows the distribution of LP levels based on the number of equivalents.

Table 8. Number of different equivalents and their LP.

	High LP	Average LP	Low LP	Zero LP	Total
Number of equivalents	90	82	129	416	717
Percentage	12.55%	11.43%	17.99%	58%	100%

Even though most of the investigated equivalents of PVs turned out to be strictly context-based and unique enough not to be inserted in different collocations, still, they are a testimony to the translators’ creativity or failure, for that matter (e.g. mistranslations). This group needs to be further examined in terms of approaches to particular types of phrasal verbs (idiomatic vs. literal) and structural patterns (morphological, syntactical etc.) that might appear in the Polish equivalents.

As for the other three levels of LP, the difference in distribution is not that sharp to indicate just one type of LP as prevailing. The assessment of LP must be considered, to some extent, arbitrary and some cases might be treated as borderline cases. A larger sample of PVs would definitely shift the balance and an evaluation of LP levels from more professional translators would enhance the precision of LP assignment thus giving a more objective view of the quality of corpus equivalents .

When considering lexicographic implications of the corpus analysis, much depends on what type of dictionary the equivalents were to be included in. If it was a large bilingual dictionary, average LP could be factored in. On the other hand, in a specialized dictionary of phrasal verbs where (ideally) the structures receive an in-depth coverage, more equivalents with low LP could be included and specific phrases added to a list of collocations. The best-case scenario is to develop an English-Polish electronic (online) dictionary of phrasal verbs where a parallel corpus, like *PHRAVERB*, would be readily available for consultation. If this was the case, the group of equivalents with no LP could be further explored and used in other translations (excluding the cases of mistranslations which have a didactic value of their own, e.g. in translation training).

7. Concluding remarks

The parallel corpus with 2,514 identified phrasal verbs generated useful equivalents of phrasal verbs, some of which are quite good candidates for inclusion in future English-Polish dictionaries. The criteria of lexicographic potential against which the equivalents have been compared showed that it is the context and sentence-structure that mostly affect the way translators render phrasal verbs. The equivalents granted the zero-LP status while rejected in this study as non-eligible for inclusion in dictionaries (but not incorrect unless they are mistranslations) might prove to be a good material for translation pedagogy.

Notes

[1] For most applications, a parallel corpus is aligned and the standard unit of alignment is the sentence. This can be done automatically or manually. There are some difficulties connected with automatic alignment, e.g. one sentence might correspond to two or three in the translated text.

[2] For terminological discrepancies concerning the terms *parallel corpus*, *translation corpus* and *comparable corpus,* see e.g. Baker (1995), Laviosa (1997), Johansson (1998), Olohan (2004), Hartmann (2007), Aijmer (2008). In the present paper, *parallel corpus* will be used synonymously with *translation corpus*, similarly to Sinclair (1996).

[3] The translations of quotes from Waliński (2005) are mine.

[4] *Oxford Hachette English-French/French-English Dictionary* (1994) was one of the first bilingual dictionaries based on corpora. However, it makes use of two separate monolingual corpora, one in English and one in French, each containing over ten million words.

[5] For example, errors stemming from wrong comprehension of the original, literal translation of idioms, or applying linguistic calques.

[6] The level of naturalness undoubtedly depends on the linguistic competence of the translator and his inventiveness. The issue of naturalness is also addressed in Lewandowska-Tomaszczyk (2001).

[7] Dezortová also analyzed lexicographic equivalents but the scope of her investigation included only five phrasal verbs: *carry out*, *go back*, *go on*, *pick up* and *set up*.

[8] Both the originals and their translations were harvested manually by copying the online content into text editor. All rich content was then removed and the texts were coded in order to create bitexts for alignment.

[9] The complete tag set for Tree Tagger can be found at http://courses.washington.edu/hypertxt/csar-v02/penntable.html.

[10] Text processing, alignment, morphological tagging and the html site of *PHRAVERB* were done by dr Grzegorz Krynicki, a specialist in parallel corpora.

[11] The sentence-to-sentence alignment in *hunalign* is about 88% (F-score combining precision and recall equals 0.8756 – Krynicki 2006: 152). Both precision and recall measures were not used for evaluation of the automatic indexing of PVs. The semi-automatic method was used because the goal was to create a practical tool with an easy access to the selected PVs and not to find the best method of automatic PV search.

[12] Two types of omissions occur in the corpus. The first type involves skipping the phrasal verb in a sentence that was otherwise translated into Polish. The other type of omission is a result of skipping the whole English sentence (or larger chunk of the text) that features the phrasal verb.

[13] Brackets at the end of each sentence indicate the title of the article from which the sentence is taken. If the same phrasal verb occurs more than once in a given article, the number of the line is also given.

References

Aijmer, K. 2008. ‘Parallel and comparable corpora.’ In A. Lüdeling and M. Kytö (eds.), *Corpus linguistics. An international handbook,* vol. 1: 275–292. Berlin: Walter de Gruyter.

Aldahesh, A.Y. 2009. *Translating idiomatic English phrasal verbs into Arabic. A contrastive linguistic study*. Saarbrücken: VDM Verlag Dr. Müller.

Altenberg, B. and S. Granger 2002. *Lexis in Contrast*. Amsterdam: John Benjamins.

Atkins, B. T. S. and M. Rundell 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baker, M. 1995. ‘Corpora in Translation Studies. An Overview and Suggestions for Future Research.’ *Target* 7.2: 223–43.

Barlow, M. 2000. ‘Parallel texts in English teaching.’ In S. P. Botley, T. McEnery and A. Wilson (eds.), *Multilingual corpora in teaching and research*. Amsterdam: Rodopi.

Bejont, H. 2010. *The lexicography of English. From origins to present*. Oxford: Oxford University Press.

Campoy Cubillo, M. C. 1997. *Semantic analysis of adverbial, prepositional and adverbial-prepositional verbs*. Castellón: Universitat Jaume I.

Claridge, C. 2002. ‘Translating phrasal verbs.’ In B. Kettemann and G. Marko (eds.), *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 361–373.

Dezortová, J. 2010. *Phrasal verbs and their translations into Czech. A corpus-based study*. M.A. Dissertation, Masaryk University.

Hampe, B. 2002. *Superlative verbs. A corpus-based study of semantic redundancy in English verb-particle constructions*. Tübingen: Gunter Narr Verlag.

Hartmann, R.R.K. [1992] 2007. ‘Contrastive linguistics: (How) Is it relevant to bilingual lexicography?’ In R.R.K Hartmann, *Interlingual lexicography. Selected essays on translation equivalence, contrastive linguistics and the bilingual dictionary*. Tubingen: Max Niemeyer, 89–94.

Hunston, S. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Johansson, S. 1998. ‘On the role of corpora in cross-linguistic research.’ In S. Johansson and S. Oksefjell (eds.), *Corpora and cross-linguistic research: Theory, method, and case studies*, 3–24. Amsterdam: GA: Rodopi.

Johansson, S. 2007. *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam: John Benjamins.

Kenning, M. 2010. ‘What are parallel and comparable corpora and how can we use them?’ In: A. O’Keffee and M. McCarthy (eds.), *The Routledge handbook of corpus linguistics*, 487–501.

Kenny, D. 1998. ‘Equivalence.’ In M. Baker (ed.), *The Routledge encyclopedia of Translation Studies*. London: Routledge, 77–80.

Kretowicz, J. 2005. *The two Polish translations of The Martian Chronicles by Ray Bradbury. Phraseological equivalence in translation*. M.A. Dissertation, Jagiellonian University.

Krynicki, G. 2006. *Compilation, Annotation and Alignment of a Polish-English Parallel Corpus*. PhD Thesis, Adam Mickiewicz University.

Landau, S. 2001. *Dictionaries: The art and craft of lexicography*. (Second edition). Cambridge: Cambridge University Press.

Laviosa, S. 1997. ‘How comparable can 'comparable corpora' be?’ *Target* 9.2: 289–319.

Lewandowska-Tomaszczyk, B. 2001. ‘Dictionaries, language corpora and naturalness in translation.’ In Marcel Thelen and Barbara Lewandowska-Tomaszczyk (eds.), *Translation and Meaning, Part 7. Proceedings of the Maastricht Session of the 3rd*

International Maastricht-Łódź Duo Colloquium on "Translation and Meaning". Maastricht: Universitaire Pers Maastricht, 177–185.

Lindner, S. J. 1983. *A lexico-semantic analysis of English verb-particle constructions with UP and OUT*. Trier: LAUT.

Lipka, L. 1972. *Semantic structures and word-formation: verb-particle constructions in contemporary English*. Munich: Wilhelm-fink-Verlag.

Lubensky, S. and M. McShane. 2007. 'Bilingual phraseological dictionaries.' In H.d Burger (ed.), *Phraseology: An international handbook of contemporary research*. Vol 2, 919–929.

Malmkjær, K. 1998. 'Love thy neighbour: Will parallel corpora endear linguists to translators?' *Meta* XLIII, 4: 1–8.

Masiulanis, J. 1974. *English phrasal verbs and their Polish equivalents: A contrastive study with pedagogical implications.* M.A. Disseration, Adam Mickiewicz University, Poznań.

Mencken, H. L. [1919] 1969. *The American language*. New York: A.A. Knoff.

Meyer, G. A. 1975. *The two-word verb: a dictionary of the verb-preposition phrases in American English*. 1975. The Hague : Mouton.

Olohan, M. 2004. *Introducing corpora in translation studies*. London: Routledge.

Sinclair, John (ed.). 1987. *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English Language Dictionary.* London: HarperCollins.

Sinclair, John. 1996. 'Corpus to corpus: A study of translation equivalence.', *International Journal of Lexicography* 9.3: 171–178.

Sroka, K. A. 1972. *The syntax of English phrasal verbs*. The Hague: Mouton

Teubert, W. 2002. 'The role of parallel corpora in translation and multilingual lexicography.'In B. Altenberg and S. Granger (eds.), *Lexis in contrast. Corpus-based approaches,* 189–214.

Teubert, W. and A. Čermáková. 2005. 'Directions in corpus linguistics.' In M.A.K. Halliday, W. Teubert, C. Yallop and A. Čermáková, *Lexicology and Corpus Lingustics*. London: Continuum, 113–165.

Vallins, G. H. 1956. *The pattern of English*. London: Adre Deutsch.

Varantola, K. 2006. 'The contextual turn in learning to translate.' In L. Bowker (ed.), *Lexicography, terminology and translation. Text-based studies in honour of Ingrid Meyer*. Ottawa: University of Ottawa Press, 215–226.

Waliński, J. 2005. 'Korpusy wielojęzyczne – równoległe i porównywalne.' In B. Lewandowska-Tomaszczyk (ed.), *Podstawy językoznawstwa korpusowego*. Łódź. Wydawnictwo Uniwersytetu Łódzkiego, 42–60.

Yatskovich, I. 1999. "Some ways of translating English phrasal verbs into Russian". *Translation Journal* 3.3. (http://translationjournal.net/journal/09russ.htm).

Zanettin, F. 2002. 'Corpora in Translation Practice.' In E. Yuste-Rodrigo (ed.). *Language Resources for Translation Work and Research, LREC 2002 Workshop Proceedings*. Las Palams de Gran Canaria, 10–14.