RiCL Research in Corpus Linguistics

Phrasal Verbs in learner English: a semantic approach. A study based on a POS tagged spoken corpus of learner English

Joanna Wierszycka[1]
Adam Mickiewicz University / Poland

Abstract – Phrasal Verbs (PVs), understood as a verb and a particle, though very common in native speech, are reportedly difficult to learn by non-native speakers (NNSs) of English (see Celce-Murcia and Larsen-Freeman 1999). The hypothesis is therefore put forward that for Polish learners of English too the range of PVs is generally significantly smaller than for English native speakers (NSs) and that their degree of use of the semantic categories of PVs is inversely proportional to the PVs' level of idiomaticity. In other words, Polish learners have little trouble with transparent verbs, more with semi-transparent and most with opaque ones (see Dagut and Laufer 1985). In order to verify this hypothesis, we have used the evidence from the *PLINDSEI* corpus, that is, the Polish part of the *LINDSEI* (*Louvain International Database of Spoken English Interlanguage*), containing advanced English as spoken by NSs of Polish, and from the *LOCNEC* (*Louvain Corpus of Native English Conversation*), which we have used as a reference corpus. The comparison of PV usage by Poles as NNSs of English and by English NSs has been performed employing the scheme of contrastive interlanguage analysis (Granger 1996). We show learner over- and underuse of items and illustrate the searches conducted for identifying patterns of use. The methodology applied consists in a partially automatic extraction and a subsequent manual filtering of PVs from a POS-tagged NNS corpus and its reference NS corpus. A semantic analysis of the extracted PVs based on the notion of compositionality (see Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004) has been performed and the hypotheses verified.

Keywords – phrasal verbs, POS tagged corpus, semantic analysis, learner language, spoken learner corpus

1. INTRODUCTION

The aim of the paper is to verify the hypothesis that learners use idiomatically opaque Phrasal Verbs (PVs) less frequently than transparent items, assuming a linear scale of idiomaticity. The classification of PVs is performed by providing a semantic analysis of those items as used by Polish learners of English. The analysis has been done on a POS (part of speech)-tagged NNS corpus of oral English, *PLINDSEI* (that is, the Polish part of the *LINDSEI – Louvain*

[1] This project was funded by the National Science Centre, Poland (grant no. 3787/B/H03/2011/40). Special thanks are due to Paul Rayson for providing the tools and expertise needed to work with CLAWS4. I wish to thank Dr. Alejandro Alcaraz Sintes for his many suggestions and criticisms of an earlier draft of this paper.

International Database of Spoken English Interlanguage) and the *LOCNEC* (*Louvain Corpus of Native English Conversation*), which we have used as a reference corpus.

In Section 2 we provide an introduction to the grammatical annotation of the NNS or spoken learner corpus (Gilquin, De Cock and Granger 2010) that we have used for the purpose of the analysis. In the next section, we offer a more detailed analysis of the non-native use of PVs (Celce-Murcia and Larsen-Freeman 1999 and Armstrong 2004), as attested in the corpus data.

For the purpose of this paper, PVs are defined as a union of a lexical verb and a following particle[2] (see Quirk et al. 1972: 1150ff). PVs are to be distinguished from Prepositional Verbs (e.g., *call* [*on* NP] ('visit NP'), *cope* [*with* NP], *laugh* [*at* NP], *provide* NP [*with* NP]) and Phrasal-Prepositional Verbs (e.g., *face up* [*to* NP], *cut down* [*on* NP, e.g., *on expenses*], *fall back* [*on* NP, e.g., *on your wife's money*] (cf. Cappelle 2005). PVs as defined above are equivalent to, for example, Dehé's (2002) particle verbs or Pelli's (1976) verb-particle constructions. They "are sometimes [also] called two-word verbs because they usually consist of a verb plus a second word (...) a particle (...) to distinguish it from prepositions and other adverbs, although we acknowledge that (...) the same word can fit into more than one category" (Celce-Murcia and Larsen-Freeman 1999: 426). However, since the learner is the center of our research, the term most frequently used in pedagogical approaches and EFL course books, the label 'phrasal verbs', was decided upon.

The verb and the particle operate as a PV not only when they fulfill certain structural criteria, but also when they are found to function together semantically as a unit. The semantic unit features one of the following configurations (cf. Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004):

- Both the verb and the particle retain their literal meanings, the particle often indicating geographical direction, e.g., *come back.* These are directional PVs and are semantically transparent.
- The verb has a literal meaning and the particle provides an aspectual meaning, which is redundant, cf. Dehé (2002) and Hampe (2002) respectively, e.g., *read through.* These are aspectual PVs and are semantically semi-transparent.
- The verb and the particle have an idiomatic meaning, e.g., *come across.* These are idiomatic PVs and are semantically opaque.

Although very common in native speech (see Biber et al. 1999: 408), PVs are reportedly difficult to learn for NNSs of English (see Celce-Murcia and Larsen-Freeman 1999). However, Marks (2005: 12) points out that, against the common supposition of students and even teachers, the meaning of PVs is not illogical and random. What is more, the meaning can often be understood if learners recognize metaphorically extended meanings of particles and verbs. The underlying reason of the learners' difficulty with PVs might stem from the rule that regulates the native use of these items, namely, that the use of PVs by NSs is directly proportional to the PVs' level of idiomaticity.

Taking the above-mentioned factors into account, our hypotheses of PV use by Poles are the following:

- In the first place, the number of PVs used by learners tends to be significantly smaller than that of NSs.
- Secondly, PV use by learners is inversely proportional to the PVs' level of idiomaticity. In other words, Polish learners should have little trouble with semantically transparent verbs, more with aspectual verbs, and most with idiomatic ones. This would be contrary to how NSs use their language, as their use of idiomatic PVs grows together as their level of idiomaticity increases (see Celce-Murcia and Larsen-Freeman 1999).
- Finally, we suspect that the number of PVs used by Polish learners should vary according to their degree of exposure to the English language (number of years studying English in a classroom situation and number of months spent in an English-speaking country).[3]

All three hypotheses are summed up in Table 1.

---

[2] This particle is referred to as 'adverbial particle' in the CLAWS POS tagging system. For examples, see <http://ucrel.lancs.ac.uk/bnc2sampler/guide_c7.htm#m3prepadv-prep> (7 May 2012).

[3] The *LINDSEI* transcripts come along with learner metadata in the form of 'learner profiles' (see Gilquin et al. 2010). This has given us access to information on the time spent abroad and in the classroom by the Polish learners. However, it did not contain more specific information, such as, for example, the number of movies without voice-over watched by the learners.

HYPOTHESIS NO.	DEFINITION
H1	The number of PV tokens is significantly lower in NNSs than in NSs.
H2	The distribution of semantic categories of PVs in NNS is inversely proportional to the PV's level of idiomaticity as observed in NS use. (Polish EFL speakers underuse the idiomatic semantic category of PVs most).
H3	Differences in L2 exposure among NNS do matter. It is assumed that a longer exposure to the English language will produce more PVs in the learners' speech.

Table 1. Hypotheses pertaining to learners' PVs use

2. POS TAGGING OF THE *PLINDSEI* CORPUS: RESOURCES AND REASONS

In the case of this study, a part-of-speech (POS)-tagged corpus constituted an important, yet independent, part of the methodology. The two corpora that have been POS-tagged for the purpose of this study are *PLINDSEI* and *LOCNEC*. Both comprise subcorpora of the mother project, the *LINDSEI*. This is a corpus of advanced English learner speech produced by young adults with different mother languages (Gilquin 2012). *PLINDSEI* is the Polish IL[4] component of *LINDSEI*. Alongside the non-native varieties of English, a native English comparative corpus, the *LOCNEC* was compiled, in order to carry out contrastive interlanguage analyses both across various NNS categories, and between any of the NNS speech and the NS speech, since the *LOCNEC* was compiled following the same criteria as all *LINDSEI* subcorpora (see Figure 1). The *LINDSEI* language data came from informal interviews conducted in a question-answer format and constitutes continuous discourse. The target size of each subcorpus was aimed at approximately 100,000 words, but the subcorpora usually consist of a larger amount of data.[5] For the exact numbers, see Gilquin et al. (2010).

Although data annotation is the natural next step after data collection and transcription,[6] there were no available POS taggers trained on learner data when the *LINDSEI* data were annotated (academic year 2008–2009). It was therefore decided to test CLAWS (Garside 1995), a well-known tagger which had already been successfully trained on native spoken language on the *British National Corpus*.

For reasons of space there is no room to discuss at length the whole process of POS-tagging *PLINDSEI*. Still, it is important to stress that the main aim of tagging *PLINDSEI* was to find and pinpoint learner dysfluencies which caused erroneous POS tagging, in order to train the tagger on this difficult kind of input, and therefore improve its accuracy. We managed to achieve an overall tagger accuracy of 98.5%. Since it was the first attempt at POS-tagging a spoken learner corpus (see Mukherjee 2007; Aijmer 2009), it is hoped that other scholars engaged in this field of study will profit from our experience. What must be remembered is that the *PLINDSEI* corpus was POS-tagged with no particular grammatical or vocabulary item, such as, for example, PVs, in focus. This approach, that is, giving the same importance to all parts of speech, will greatly enhance further research in ways that the people responsible for tagging the corpus had not projected. In fact, as Leech (2004)[7] wrote, "no one in their right mind would offer to predict the future uses of a corpus".

At the same time it is vital to notice that there are pre-assigned grammatical categories, such as infinitives, adverbs or prepositions, which are essential to define the tags, as it is impossible to approach corpus tagging in a theory-free manner.

Once an effective POS tagging is done, it enhances grammatical searches enormously. General searches for PVs (i.e., any form of verb plus adverbial particle) would have been impossible without it. However, in searches for concrete examples of particles, or whole PVs, no POS tagging is needed. In the latter approach the researcher has a raw corpus at his or her disposal. Research on learner PV use so far has been based on raw corpora and, as a consequence, has been conducted on selected PVs only (e.g., Gilquin 2012). Therefore, the searches for PVs for the purpose of this study were done with the use of POS tags and with the help of *WordSmith Tools* software (Scott 2008).

3. METHODOLOGY

The investigation presented in this paper is based on the evidence of advanced spoken English by NSs of Polish, as described in section 1, and was performed using the scheme of contrastive interlanguage analysis (Granger 1996), the

---

[4] Granger's (1996) NL (native language) corresponds to our NS (native speaker), and her IL (interlanguage) corresponds to our NNS (non-native speaker).

[5] For example, the Polish part of *LINDSEI* has 114,862 words, while the French part has 143,044 words.

[6] However, see Sinclair (1991).

[7] <http://users.ox.ac.uk/~martinw/dlc/chapter2.htm> (7 December 2011).

Native Language (NL) vs. Interlanguage (IL) branch in particular (see Figure 1). Ellis (1994) stresses the importance of collecting comparable samples of learner language in interlanguage comparisons. However, the issue of comparability has not always been so obvious to scholars and, in fact, one of the reasons why many important aspects in applied linguistics have remained inconclusive is that researchers have not been "comparing like with like" (Granger 1996: 44). Granger also suggests deciding on genre sensitivity, rather than ease of access, when choosing the right reference corpus for our study. For this reason, the *PLINDSEI* and *LOCNEC* corpora were considered most appropriate for this study. Both were compiled according to the same criteria and can, therefore, be considered as comparable.

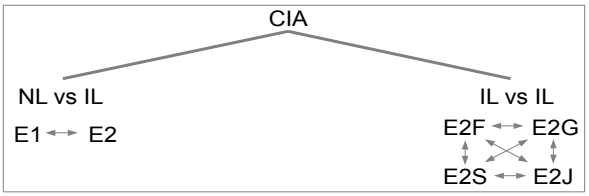

Figure 1. Contrastive interlanguage analysis (after Granger 1996: 44)

The methodology applied consisted in a partially automatic extraction and then manual filtering of PVs from the POS-tagged Polish spoken learner corpus of English and its reference NS corpus. The corpus data were used in their normalized frequencies of B-turns only. The rejected A-turns comprised the sections of the interviewers and they were of no interest for the research. Their only purpose was to keep the conversation going. B-turns, however, consisted of the interviewee part of the recordings, that is the students, selected along criteria of language and age level (for details see Gilquin et al. 2010). The overall number of words were 114,816 and 161,724 for the learner and the reference corpora, respectively. In order to arrive at a fully comparable corpus, a decision was taken to normalize the native corpus, which was done by eliminating 11 files randomly. Only the B turns analyzed which also limited the number of the number of words. The corpora produced 95,906 words in the learner corpus and 118,554 in the NS corpus.
words. Table 3 below provides the final word numbers after normalization in both corpora.

4. PHRASAL VERBS: WHAT, WHY, HOW?

It would be impossible to count all the PVs in the English language. One of the main reasons is that new ones are being constantly added, in some cases nouns or adjectives being the witnesses of the newest inventions, e.g., *I'm Christmased out* ('I'm sick of Christmas') in the 1996 movie *Elmo saves Christmas*, now widely used by the British and the American alike in the pre-Christmas period. It is an isolated example of a PV not to be found in dictionaries, but used by NSs. The number of such PVs to be "discovered" will remain unknown, and the number of coinages that will see the light of day will most probably remain a mystery. The other side of the coin is that some PVs are naturally getting out of use, and therefore dying out. There is also the factor of polysemous PVs (see, e.g., Table 5 below and the example of *pick up* in section 5.1). For all the above-mentioned reasons counting all PVs in the English language is not a straightforward and easy task. Still, there are researchers who have attempted to guess the exact number of PVs, which varies from as few as 700 (Bywater 1969: 97) up to 12,000 (Courtney 1983). It therefore seems that the attempt to count PVs is similar to that of calculating all vocabulary items in a given language.

For the purpose of this paper, PVs are selected in a two-stage procedure, after which their classification comprises another two stages. First, the basic structural definition is adapted from Quirk et al. (1972), where PVs are understood as constructions formed of lexical verbs followed by adverbial particles, e.g., *drink up.* They are to be distinguished from prepositional verbs, e.g., *dispose of*, phrasal-prepositional verbs, e.g., *get away with*, and other multi-word verbs. At this step PV candidates are automatically selected. What follows is a manual filtering stage of lexical verbs with particles from prepositional verbs and phrasal-prepositional verbs to arrive at a list of only true PVs ready for further semantic classification. The second step of the PV definition determines semantic classification and consists of two stages: classification of PVs according to the particle and further division of PVs into semantic categories along the lines of compositionality (adapted from Darwin and Gray 1999 and Armstrong 2004).

The rationale behind our research is to verify the existing research on the actual difficulty experienced by Polish speakers when learning PVs. The problem is that PVs are very common in native speech and therefore the difference between native and learner use strikes particularly hard (e.g., Celce-Murcia and Larsen-Freeman 1999; Darwin and Gray 1999; Armstrong 2004).

In terms of the procedure, PVs were extracted using the *WordSmith Tools* software, employing the search on tags. These were six different forms of a lexical verb (VV*), each followed by a particle as its context word (RP). All the C7 tags together with their explanations and examples are set together in Table 2 for clarity.

The outcome of this procedure was a list of potential PVs with all possible verb forms present in the corpus, e.g., while the particle *up* was searched for, together with the verb forms of *get*, the native corpus brought the following results: *gets up*, *getting up*, *get up* and any other string of signs that the data would show.

C7 TAG	DEFINITION + EXAMPLE
VV0	base form of lexical verb (e.g., *give*, *work*)
VVD	past tense of lexical verb (e.g., *gave*, *worked*)
VVG	*–ing* participle of lexical verb (e.g., *giving*, *working*)
VVI	infinitive (e.g., *to give... It will work...*)
VVN	past participle of lexical verb (e.g., *given*, *worked*)
VVZ	*–s* form of lexical verb (e.g., *gives*, *works*)
RP[8]	prep. adverb, particle (e.g., *about*, *in*)

Table 2. C7 tags used for searching PVs[9]

It should be noticed at this point that not only particles immediately following verbs were taken into consideration in the analysis, but also words with further search horizons, that is one or more words separating the verb from the particle, e.g., *<VVG> showing* ***<PPIO2>us*** *<RP> round <VVI>,* and *put* ***<AT> the <NN1> paper <.>.*** *<RP> back.* In this way, it may be assumed that all possible "verb plus particle" cases were retrieved and analyzed. Table 6 below presents a full list of the particles found in both the native and the non-native corpora.

Apart from the fully automatic extraction based only on the criteria of CLAWS-implemented POS grammatical categories, manual filtering was necessary, as the data were not free from noise. The first filter consisted of the application of the transitivity criterion (Armstrong 2004) to both the POS tagged Polish spoken learner corpus of English and its reference corpus.[10] Whether a PV is intransitive or not is a topic for separate discussion (see, e.g., Quirk et al. 1972).

The transitivity criterion serves to decide if a candidate for a PV (based on all the criteria described above) is actually a PV (Armstrong 2004). This, in turn, is based on the following premise: particles must be intransitive, i.e., they must form a unit with the verb, not with their object. Based on this principle, the PV candidates *set in* and *get up,* although seemingly perfect PVs, because they are made of a verb and a particle, were not considered to be true PVs, because the context of the PV candidate was taken into consideration. To be precise, the particle, which belongs to the object in *set in an aerobics class*, *getting up these mountains*, gives away lack of affinity with the PV category at the same time.

5. DISCUSSION OF FINDINGS

Having complied with all the above-mentioned criteria, the following results were reached when it concerns the overall number of PVs.

	NNS	NS
PVs	227	875
Corpus size B turns, (NS normalized)	95,906	95,862
Chi-square		384.28

Table 3. PV general token counts

[8] Bolinger (1971: 26, 28) uses the term 'prepositional adverb' to refer to particles which can be used both as adverbs and as prepositions, e.g., *in*, *up*. 'Prepositional adverb/particle' is CLAWS terminology. However, not only are all PV candidates extracted from the corpora with the use of POS tags at step 1, but they also go through manual filtering where prepositional adverbs are excluded and only true particles are left.

[9] For other POS tags, see CLAWS manual <http://ucrel.lancs.ac.uk/bnc2sampler/ guide_c7.htm#m3prepadv-prep> (7 May 2012).

[10] However, in order to have data tagged in a systematic way, none of them was manually annotated. The 1.5% error is small enough to leave the resource and not mingle with data manually, since introducing a human factor means initiating uncertainty as to the quality of the data which a computer corpus is not normally associated with.

	NNS	NS
PVs	85	274
Corpus size in types	4,917	5,606
Chi-square value		79.86
TTR[11]	37.44 %	31.31 %

Table 4. PV general type counts + TTR

In compliance with the first hypothesis, the learners' token number of PVs is indeed significantly smaller than that of NSs, as shown by the chi-square value (see Table 3). NNs appear to use almost four times as many tokens of PVs as learners, hence the high chi-square value. In order to be able to count the TTR, the types for both speaker groups are also presented in Table 4. From the tables above one may conclude that the learner's speeches seem to be lexically more varied than those of NSs. However, this may be due to the presence of hapaxes, which may not necessarily reflect learners' real competence of the vocabulary. The TTR values will be compared against values of semantic PV groups later in the paper.

5.1. The semantic approach to PV grouping: stage one (the particle)

PVs are semantic units. There is a bond between the verb and the particle but the degree of attachment varies. "In one case, the main factor determining the unity between the verb and the particle is semantic, mainly lexical, in the other, formal syntactic" (Sroka 1972: 180). Semantic categories comprise three groups: transparent, semi-transparent and opaque (cf. Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004). The first aim of the research was to check if the distribution of learners' semantic categories of PVs was inversely proportional to the level of PV idiomaticity as observed in NSs' use (e.g., Howarth 1998: 178). As the first stage of the semantic approach, PVs were grouped according to the particle (e.g., Rudzka-Ostyn 2003), bearing in mind that the particle carries more semantic load than the verb. Therefore, it was more logical to list PVs not according to the verb, but under the particle. Such a classification was carried out for both speaker groups separately and, as a result of that, interesting observations came into light (see further down).

Naturally, it frequently happened that within a given PV form, different meanings occurred. Although structurally identical, e.g., *pick up*, in *pick up a girl*, *pick up a job* and *pick up a tent* were classified as different types, because they come from three different semantic domains (see Bentivogli et al. 2004) and as such they could hardly be classified under one semantic term. All the aforementioned examples come from the *LINDSEI* corpus, but similar examples also occurred in the *LOCNEC* corpus, showing how the two speaker groups diverged not only in the number of verbs, but also in their sense distribution. Some of the differences in meaning expressed by the two speaker groups in a group of several PVs from the *LINDSEI* and *LOCNEC* corpora are shown in Table 5.

PV	NS SENSE	NNS SENSE	NS & NNS SENSE
take off	to achieve wide use or popularity	start	–
work out	to accomplish by work or effort		to prove successful, effective, or satisfactory
come over	approach		pass as somebody
go through		travel	experience, examine
come up			approach

Table 5. Polysemy across speaker groups: Common PVs with different senses

Subsequent analysis dividing the PVs into three semantic categories revealed that the polysemous PVs also crossed the idiomaticity line sometimes, not only across the speaker groups, as presented in Table 5, but also within one speaker group. Such a situation happened, for example, in the case of *come down*, which may be a literal, compositionally transparent PV, as in *prepare for the curtain to come down*, meaning 'to move downward', but which may also be a totally idiomatic, opaque PV, as in *she broke her hip and came down with cancer*, meaning 'to become sick with (an

[11] Type/Token Ratio (TTR): the number of types divided by the number of tokens. This indicates how rich or lexically varied the vocabulary in the text is. In the example of NNS, the TTR is 85 (types) ÷ 227 (tokens) x 100 = 37.44 %.

illness)'. All the form-sense differences do not mean that there are no common PVs between those two speaker groups, as may be seen in the fourth column of Table 5.

Another interesting observation pertaining to particle use by the two speaker groups is that, out of all of the particles present in the corpora, quite a few were absent from the learners' repertoire and one particle was used by the students only. The particle division into groups is shown in Table 6 below.

	PARTICLES
COMMON	*across*, *along*, *around*, *back*, *down*, *in*, *off*, *on*, *out*, *over*, *through*, *up*
SOLELY NS	*(a)round*, *after*, *away*, *by*, *for*, *to*
SOLELY NNS	*about*

Table 6. Particles in NS and NNS speakers

The division of PVs according to the particle was the first stage of the semantic analysis of PVs, because the basic meanings of the particles were to help out in deciphering the transparent and semi-transparent PVs. This, in turn, paved the way for the third category of PVs: opaque PVs, since the meaning of some of the opaque PVs is not fully opaque, but has its semantic roots in the meaning of the semi-transparent and transparent categories (see Rundell 2005).

Rudzka-Ostyn (2003), in her pedagogically-oriented book, employed the cognitive approach as a means for effective acquisition of PVs, and proposed 17 particle meanings, for which the leading meanings are presented below. Some of the particle senses are listed here in order to clarify further PV division.

- *on* stands for continuation of an action or situation, e.g., *walked on*, *rambled on.*
- *around* stands for location or motion (in different directions) often viewed from a central point, paths in all kinds of directions, e.g., *travel around*, *come (a)round*, *look around*, *bossing around*.
- *through* stands for motion inside an entity from end to end, activities viewed as complete(d) motions, e.g., *drive through*, *slept through*, *soaked through.*
- *over* is being or moving higher than and close to something or from one side to the other, examining thoroughly from all sides, e.g., *turning over*, *lingered over*.

In a similar fashion, other researchers have tried to group particles according to their meaning. By way of comparison to the particles above, the following definitions are worth quoting here:

- *on* means some more (Jackendoff 1997), expresses continuative action if used with activity verbs (Celce-Murcia and Larsen-Freeman 1999), i.e., verbs that express action not state, e.g., *carry on*, *keep on.*
- *around* stands for in a circle or with a circular motion, expresses absence of purpose (continuative) if used with activity verbs (Celce-Murcia and Larsen-Freeman 1999), e.g., *mess around*, *play around.*
- *through* means from beginning to end, e.g., *read through*, *think through.*
- *over* is again or re-, iterative if with activity verbs, e.g., *think over*, *do over (and over again)*, (Celce-Murcia and Larsen-Freeman 1999).

As can be noticed from the comparison of the two approaches, Jackendoff's (1997) and Celce-Murcia and Larsen-Freeman's (1999) definitions are simpler, which enables quicker grasping of the idea of particle, and thereby of PV transparency. Rudzka-Ostyn's (2003) definitions, on the other hand, are more elaborate, provide a thorough analysis within each particle and are accompanied by elaborate exercises, which is more suitable for self-conscious students.

5.2. The semantic approach to PV grouping: stage two (compositionality)

What follows from the grouping of PVs according to the particles is the semantic analysis of PVs, based on the idea of compositionality (Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004). Compositionality means that the inherent parts of a PV (verb and particle) either mean the same as they do when they are used on their own (as in *pulled up the anchor*), i.e. they are semantically transparent, or they are partially transparent (e.g., *locked up the office*, where *up* does not indicate upward position, but a redundancy aspect), or they cannot be semantically broken down at all (e.g., in *came across sth*, neither *came* nor *across* mean what they do when they are used on their own).

There are various labels for the same terminology in the literature. Table 7 sums up the major ones. Before going on to the analysis of the corpora, each of the semantic PV categories needs to be defined: opaque, semi-transparent and transparent.

Opaque PVs (Armstrong 2004) are also referred to as idiomatic (Jackendoff 1997; Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004) or noncompositional (Jackendoff 1997). These PV combinations consist of a verb and

a particle which are both opaque. Like other types of idiom, they are probably stored as whole units in the lexicon and as such they have to be 'learnt' as units, e.g. *come across*.

Semi-transparent PVs (Celce-Murcia and Larsen-Freeman 1999; Armstrong 2004), also referred to as aspectual (Jackendoff 1997; Dehé 2002) or semi-transparent (Celce-Murcia and Larsen-Freeman 1999), are formed by a verb which retains its lexical meaning and a particle which does not (Armstrong 2004). Other researchers claim that what the particles express is not aspect, but *aktionsart* (e.g., Brinton 1988). Jackendoff (1997) and Celce-Murcia and Larsen-Freeman (1999) talk of aspectual PVs when the particle conveys its (aspectual) meaning. According to Jackendoff (1997: 541), "[i]n these cases, the particle does not satisfy an argument position of the verb; rather it contributes an aspectual sense, often paraphrased by some sort of adjunct PP. *Run/sing on*, for instance, means roughly 'run/sing some more'".

Fully transparent PVs (Armstrong 2004), otherwise called directional, literal (Celce-Murcia and Larsen-Freeman 1999) or fully compositional (Jackendoff 1997), are PVs in which the particle has a directional meaning and the verb is a verb of motion. The subject therefore moves in the direction specified by the particle in the manner specified. Here we can expect relatively little difficulty with the use of those PVs on the part of the learner (see, e.g., Celce-Murcia and Larsen-Freeman 1999). Some particles lend themselves easily to the transparent and fully compositional category, e.g., *back*, which is almost always directional, e.g., *come back*, *give back.* Another feature of compositional PVs is that they are self explanatory in terms of their semantics, hence there is no need for them to be listed as separate items in the lexicon. Jackendoff (1997: 541) stresses in this respect that "there is no need to list the verb-particle combinations in the lexicon, since the particle satisfies one of the verb's argument positions, and the meaning is fully compositional".

AUTHORS(S)	DEGREE OF IDIOMATICITY FROM HIGH TO LOW		
Quirk et al. 1972	highly idiomatic	semi-idiomatic	nonidiomatic
Jackendoff 1997	idiomatic = noncompositional meaning	aspectual	directional = compositional meaning
Celce-Murcia and Larsen-Freeman 1999	idiomatic	semi-transparent	literal
Armstrong 2004	idiomatic = opaque	aspectual = semi-transparent	directional = transparent

Table 7. PV compositionality

Both in Table 7 and in the description of the categories the PVs have been presented by introducing the degree of idiomaticity from absolute to none for clarity. However, it needs to be stressed that there is no direct endpoint to either of the categories, as is pointed out by Downing and Locke (2006: 343), who claim that "[i]t is by no means easy to establish boundaries between what is idiomatic and what is not", and by Biber et al. (1999), who maintain that verbs should rather be graded according to relative fixedness rather than to binary categories, given the categorization difficulties. Figure 2 below demonstrates the linear scale of idiomatic compositionality. It needs to be remembered that the categories are not points on the scale of compositionality. It is relative boundaries between them that need to be borne in mind, not so much the categories themselves. All of the PVs within each of the definitions of the compositional categories have been classified along the definitional criteria, and the outcome is summed up in Tables 8-10 below.

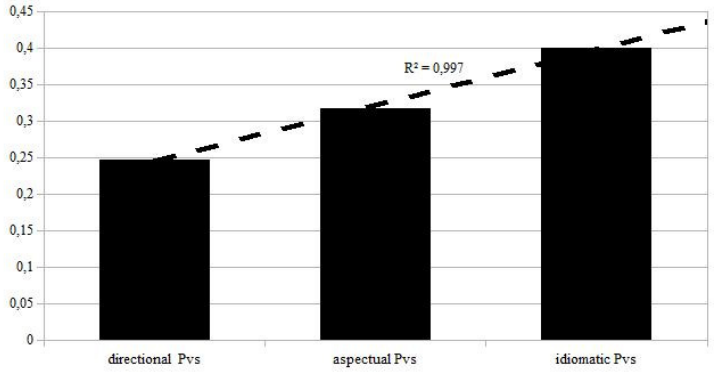

Figure 2. Incidence of PV use in the *LOCNEC* corpus expressed in PV TTR according to the categories of compositionality

The three tables below present different ways of looking at the data. Table 8 presents the types of PVs compared against the number of word types for the calculations to be valid. Chi-square values are presented in the last column to demonstrate statistical significance. Even assuming the strictest significance values, for $p<0.0001$, where the critical

value is 15.13, the chi-square values are still higher and therefore show significant differences between the two corpora in all three cases, with the opaque PVs showing the greatest discrepancy. One can obviously spot the gross differences between the corpora, but what strikes us immediately the most is that learners have a rather equal distribution of PVs, which might imply their having relatively few problems with idiomaticity.

Types account for only part of the picture, though. Table 9, in turn, concentrates on PV tokens derived from each corpus, and the chi-square values are also calculated here, this time against the number of word tokens, demonstrating again significant differences. What is striking in this comparison is that as idiomaticity grows, the significance diminishes. A quick look at the raw numbers reflects this tendency in both corpora. However, only a TTR comparison offers a deeper insight into the actual tendency of use when those two groups are compared.

COMPOSITIONALITY	NNS	NS	CHI-SQUARE
transparent PVs	27	85	30.06
semi-transparent PVs	31	89	28.06
opaque PVs	27	100	41.99

Table 8. Categories of PV compositionality in NS and NNS corpora, expressed in PV types

COMPOSITIONALITY	NNS	NS	CHI-SQUARE
transparent PVs	84	344	158.06
semi-transparent PVs	70	281	126.94
opaque PVs	73	250	97.08

Table 9. Categories of PV compositionality in NS and NNS corpora, expressed in PV tokens

COMPOSITIONALITY	NNS	NS
transparent PVs	32.1%	24.7%
semi-transparent PVs	44.2%	31.67%
opaque PVs	36.9%	40%

Table 10. Categories of PV compositionality in NS and NNS corpora, expressed in PV TTR

Table 10 sums up the TTR for each of the groups within the NNS and NS corpora, respectively. What can be concluded from this juxtaposition is that Polish speakers overuse the transparent category (32.1% vs. 24.7%), but their tendency of use is not linear as it is in the case of NSs. Natives exhibit more types than tokens of PVs as the level of idiomaticity grows (24.7% > 31.67% > 40%), while Polish speakers break the linearity at the level of the semi-transparent category. If the tendency was linear, there should be more than 44.2% of semi-transparent PVs used. At this point it seems valuable to compare the TTR values to the overall TTR for the general PV use, calculated at 37.44% and 31.31% for the NNS and the NS corpora, respectively (see Table 4). Such a comparison would enable us to conclude that learners' lexical variability probably stemmed from the use of transparent and semi-transparent PVs rather than the use of idiomatic PVs.

It is important to notice that the learner distribution of compositional categories turns out to be unequal only after comparison to the reference corpus. When looked at in isolation, learners employ all compositional categories of PVs with comparable 'ease' (see Tables 8 and 9). However, when compared to the reference corpus, where the distribution of the categories is not equal, the picture is different. In this respect, Celce-Murcia and Larsen-Freeman (1999) appear to be correct in their prediction that learners will be afraid or reluctant to use aspectual PVs, and even more so with idiomatic ones. It had been expected, however, that the tendency would be inversely proportional to the growth of idiomaticity, an expectation which did not come true.

Hypothesis number two can be verified at this stage. Table 10 above demonstrates that the tendency in PV use in the native group is directly proportional to idiomaticity. Just as idiomaticity grows, so does the use of PVs increase. However, in the non-native corpus, the tendency is not preserved. It is, however, also not inversely proportional, as had been predicted, so that the linearity of PV compositionality in the case of the non-native corpus is broken.

In trying to explain this unexpected distribution, one important observation must be made. Although PVs in general have been shown to be underused by learners, not many unusual, understood as non-dictionary PVs, can be noticed within the native corpus. This might be attributed to the conditions in which the data were collected. It was interview type of data rather than surreptitious recordings, so that the interviewees were fully aware that they were being recorded. Despite the advantages of such data, speakers could potentially have refrained from using vocabulary items they judged colloquial or inappropriate. After all, the material was being recorded in an academic environment and was intended to be made available for scientific use. Marks (2005: 12) stresses the fact that although the reality is more complicated, there is some basis for at least the first four of the beliefs that are still shared by teachers and students alike: that PVs are "colloquial, casual, informal, characteristic of speech rather than writing (...) and perhaps even a bit

sloppy or slovenly, uneducated, not quite proper". He calls them widespread popular wisdom about PVs among learners and teachers. The learners' proportionally higher semi-transparent PV use, on the other hand, may, interestingly, also stem from the same conditions. In the case of the Polish corpus, though, the recordings, although resembling an exam situation, were taking place in front of the interviewees' colleagues (which is reflected in the transcripts). On the one hand, this situation could have relaxed the subjects, by making them feel at ease. On the other, learners could have felt the pressure of being recorded, and therefore felt as if in an exam situation, which, in turn, could have made them recall and employ all grammar and vocabulary normally reserved for such occasions.

6. INDIVIDUAL LEARNER TENDENCIES

Apart from the verification of the major hypotheses, there are other observations to be made from the data analyzed. It was assumed that there would be differences in PV use resulting from the variation in foreign language experience of the learners. Despite the language level criterion being externally estimated in all *LINDSEI* language learners as advanced,[12] their exposure to the English language naturally varied. By "exposure to language" two factors are meant, namely, time spent in an English-speaking country and years of English language studied in a classroom situation prior to university.

The first look at the distribution of PVs among learners (see Figure 3) in relation to their length of stay in an English-speaking country does not clear up the picture. The lowest number of PVs used (counted in tokens) is 0 (learners nos. 25 and 48), the highest 27 (learner no. 24). When looking at the time spent in English-speaking countries, it is 2 and 4 months for the people who did not use PVs in their speech at all, and 0 months for the person who used 27 PVs. The list of the 8 highest values of PVs used, set against the length of stay in an English-speaking country, is presented in Table 11.

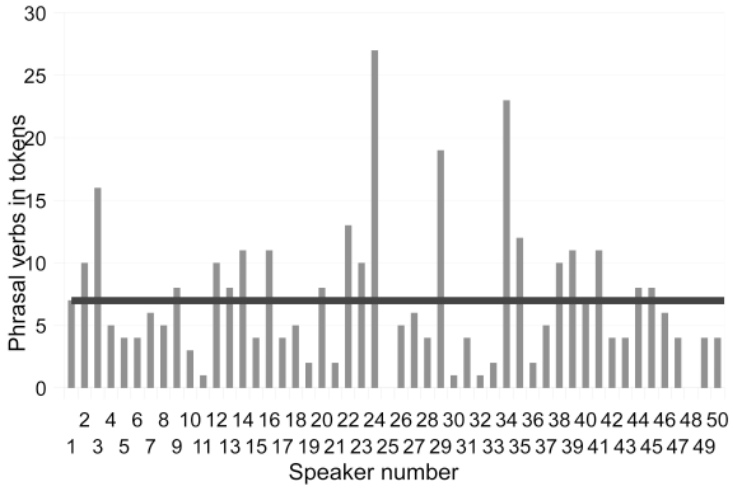

Figure 3. General PV distribution in individual users – *PLINDSEI*

Another factor worth taking into consideration is the number of years spent learning English in a classroom situation. One would expect that the students who had the longest exposure to English at school would have a good command of English PVs. One striking observation, however, is that the majority of students who had 8-11 years of English at school (longest periods) used 0-1 PVs in their conversations and only 1 student used 23 PVs. When we have a look from the perspective of learners with the highest token number of PVs in the corpus, the relationship between the token number and years of English at school is not clear either: these learners (tokens of PVs being 18-27) learnt English in a classroom situation from merely 4 years to as many as 10. In order to make sure if there is no correlation between the years of English at school and the number of PVs used, the Pearson correlation coefficient for the two variables was counted. The result, taking all 50 learners into account, is –0.02. We can thus safely confirm that it is difficult to say that it is the exposure to classroom English that predisposed any learners towards using PVs.

Most of the students also spent some time in an English speaking country so the correlation of this factor and the number of PVs used was calculated using Pearson product-moment correlation coefficient. The length of stay varies as much as 0-7 months, and the PV use within this group varies from 1 to 27. One speaker who did not travel to an

[12] All *LINDSEI* participants were third- or fourth-year 'English Language and Linguistics' University students.

English-speaking country happened to use 27 PVs; another used only 1 PV. Similarly, a 3-month stay brought about a result of 18 PVs in the interview of speaker no. 29, and only 1 PV of speaker no. 30. The Pearson correlation coefficient for the length of stay in an English-speaking country is 0.1. The results therefore suggest that no correlation between the number of phrasal verbs used and either of the ratio variables was found.

SPEAKER NO.	PVS USED	LENGTH OF STAY IN UK (IN MONTHS)	YEARS OF ENGLISH AT SCHOOL
24	27	0	6
29	18	3	4
34	23	1	10
25	0	2	11
48	0	4	5
11	1	0	8
30	1	3	8
32	1	7	8

Table 11. Number of PVs in relation to length of stay in UK and number of years of English at school

Thus, the third hypothesis, namely, assuming that language learning experience, defined as "years spent learning English in natural environment and in a classroom",[13] bears a noticeable influence on the quantity of learner PV use, cannot be confirmed. It turned out that the length of stay in an English-speaking country does not unravel the causes of PV underuse by learners, and neither does the length of English learning in school conditions, expressed in years. My supposition is that it is the quality of learning and teaching, rather than the number of years, that influences the use of PVs. As far as exposure to the English language in natural conditions is concerned, it involves more than months of passive living in a country for a foreign language learner to progress to a higher level.

As mentioned before, NSs also display different levels of command of vocabulary, even if we look at their PV distribution (Figure 4). Their PV use is, however, much more balanced than the learners', with the lowest PV token being 7, the highest 44. What is also characteristic of the NS use is high token and low type values in the most of them, which means that speakers have the tendency to repeat what is already in their repertoire, and so the PVs *go down* (9), *pick up* (a skill) (9) and *go over* (8) belong to the most commonly used PVs. On the one hand, *spread around*, *turn around*, *get away*, *look after*, *pay back*, *get by*, *put down*, *feel for*, *put on*, *leave out*, *take over*, *fall through* and *split up* are hapaxes. Had the corpus been larger, such hapaxes might have had greater occurrence and would have therefore exerted a stronger influence on the overall interpretation of the data.

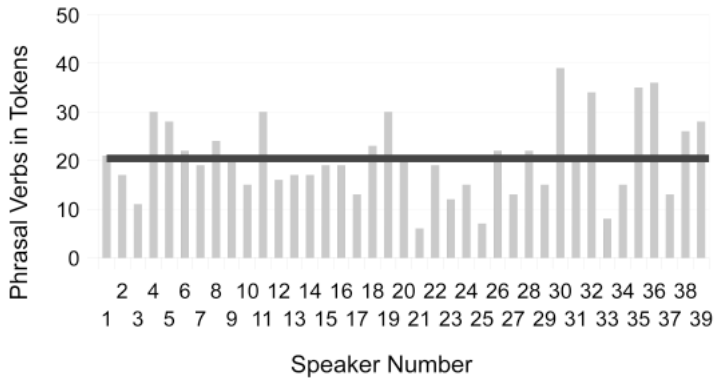

Figure 4. General PV distribution in individual users – *LOCNEC*

Finally, there is also a group of PVs common to both groups (native vs. learner) in terms of their tokens used. These are: *come back* (41 vs. 28), *go back* (56 vs. 13), *get back* (14 vs. 5), *come in* (13 vs. 2), *go out* 'leave' (26 vs. 15), *sit down* (15 vs. 3), *go on* 'continue' (32 vs. 10), *come on* (2 vs. 8), *wake up* (10 vs. 1), *go out* (socially) (51 vs. 7), *find out* (7 vs. 10), *make up* 'invent' (5 vs. 13) and *show off* (5 vs. 5). Apparently, out of the group of thirteen PVs in common, five belong to the transparent category of PVs, and four PVs belong to semi-transparent and idiomatically opaque categories. The sample is, however, too tiny to attempt any comparison, and the distributions of single PVs are not

[13] The choice of the given variables was motivated by the availability of metadata for *LINDSEI* sound files. Each of them is linked to a profile which contains information about the learner, the interviewer and the interview itself. This information makes it possible to study the potential influence of certain factors on learner language.

equal (e.g., single occurrence vs. 56 occurrences at times). Thus, the group of PVs common to both learners and natives cannot be compared so easily along the compositional category lines.

7. Conclusions and directions for further research

In this paper the problem of learner underuse of PVs has been presented using the example of Polish advanced speakers of English. General substantial underuse has been verified thanks to the employment of a POS tagged corpus, without which this research would not have been possible. PVs were divided along the lines of the semantic compositionality criterion, preceded by their classification according to the particle. What was found out in the analysis of PV compositionality is that while the native use of PVs is linear, learners do not appear to follow this tendency. They underuse PVs within all of the compositional categories, but the idiomatically opaque PVs are neglected the most.

In an attempt to find the key to this underuse, proficiency in English was called up. Neither of the two variables checked (length of stay in an English-speaking country and years of English at school) brought meaningful results, however. From the angle of language proficiency, it remains an open question where the observed differences stem from, as participants with similar education and language experience displayed varying degrees of PV use.

Another possibility is that learners simply avoided using PVs and tried using one-word equivalents instead. It is however debatable if the one-word equivalents truly reflect the meaning of the PVs. *Dress up* and *disguise* are approximate synonyms, where *disguise* suggests an intention to deceive while *dress up* does not. The PV *sail through something* means 'to succeed' and is roughly the equivalent of 'to pass' when referring to an exam. However, only in the case of PV use is there the connotation of effortlessness (see Marks 2005). Checking whether learners do consciously avoid PVs would naturally require a systematic study in order to find out what vocabulary items were used instead of PVs and if they were effective replacements.

Finally, as regards further research, it would be necessary to investigate into the reasons why certain learners underuse PVs more than others. As the available learner metadata did not provide an answer to this question, it might perhaps be more worthwhile to examine the way in which PVs are taught and learned in language course books. Further research into learner underuse of PVs might be to design an observation exercise or a questionnaire for teachers, and to observe which and how many PVs are used by teachers in their communication with students.

References

Aijmer, Karin (ed.). 2009. *Corpora and language teaching*. Amsterdam: John Benjamins.

Armstrong, Kevin. 2004. Sexing up the dossier: a semantic analysis of phrasal verbs for language teachers. *Language Awareness* 13: 213–224.

Bentivogli, Luisa Pamela Forner, Bernardo Magnini and Emanuele Pianta. 2004. Revising WordNet Domains hierarchy: semantics, coverage, and balancing. In *COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, Switzerland, August 28, 2004*, 101–108. <http://www.aclweb.org/ anthology-new/W/W04/W04-2200.pdf> and <http://wndomains.fbk.eu/publications/Coling-04-ws-WDH.pdf> (12 July 2013).

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

Bolinger, Dwight. 1971. *The phrasal verb in English*. Cambridge, Mass.: Harvard University Press.

Brinton, Laurel J. 1988. *The development of English aspectual systems: aspectualizers and post-verbal particles.* Cambridge: Cambridge University Press.

Bywater, F.V. 1969. *A proficiency course in English*. London: University of London Press.

Cappelle, Bert. 2005. *Particle patterns in English: a comprehensive coverage*. PhD dissertation. Leuven: Katholieke Universiteit Leuven.

Celce-Murcia, Marianne and Diane Larsen-Freeman. 1999. *The grammar book*: *an ESL/EFL teacher's course*. Second edition. Boston, MA: Heinle and Heinle.

Courtney, Rosemary. 1983. *Longman dictionary of phrasal verbs*. Harlow: Longman.

Dagut, Menachem and Batia Laufer. 1985. Avoidance of phrasal verbs – a case for contrastive analysis. *Studies in Second Language Acquisition* 7/1: 73–79.

Darwin, Clayton M. and Loretta S. Gray. 1999. Going after the phrasal verb: an alternative approach to classification. *TESOL Quarterly* 33/1: 65–83.

Dehé, Nicole. 2002. *Particle verbs in English. Syntax, information structure and intonation*. Amsterdam: John Benjamins.

Downing, Angela and Phillip Locke. 2006. *English grammar. A university course*. Second edition. London: Routledge.

Ellis, Rod. 1994. *The study of second language acquisition*. Oxford: Oxford University Press.

Garside, Roger. 1995. Grammatical tagging of the spoken part of the British National Corpus: a progress report. In Geoffrey Leech, Greg Myers and Jenny Thomas (eds.), *Spoken English on computer: transcription, mark-up and application*. Harlow: Longman, 161–167.

Gilquin, Gaëtanelle. 2012. The ups and downs of phrasal verbs in spoken and written learner Englishes. Paper presented at ICAME 33. University of Leuven, 30 May–3 June 2012.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger (comp.). 2010. *Louvain International Database of Spoken English Interlanguage* (*LINDSEI*). Louvain-la-Neuve: UCL Presses universitaires de Louvain.

Granger, Sylviane. 1996. From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in contrast. Textbased cross-linguistic studies*. Lund: Lund University Press, 37–51.

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

Howarth, Peter Andrew. 1998. The phraseology of learners' academic writing. In Anthony Paul Cowie (ed.), *Phraseology*. Oxford: Clarendon Press, 161–186.

Jackendoff, Ray. 1997. Twistin' the night away. *Language* 73/3: 534–559.

Leech, Geoffrey. 2004. Adding linguistic annotation. In Martin Wynne (ed.), *Developing linguistic corpora*: *a guide to good practice*. <http://users.ox.ac.uk/~martinw/dlc/chapter2.htm> (7 December 2011).

Marks, Jonathan. 2005. Phrasal verbs and other 'phrasal' vocabulary. In Michael Rundell (ed.), *Macmillan phrasal verbs plus.* Oxford: Macmillan.

Mukherjee, Joybrato. 2007. Exploring and annotating a spoken English learner corpus: a work-in-progress report. In Sabine Volk-Birke and Julia Lippert (eds.), *Proceedings of Anglistentag 2006*. Trier: Wissenschaftlicher Verlag Trier, 365–375.

Pelli, Mario G. 1976. *Verb-particle constructions in American English*. Bern: Francke.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1972. *A grammar of contemporary English.* Harlow: Longman.

Rudzka-Ostyn, Brygida. 2003. *Word power: phrasal verbs and compounds. A cognitive approach*. Berlin: Mouton de Gruyter.

Rundell, Michael (ed.). 2005. *Macmillan phrasal verbs plus*. Oxford: Macmillan.

Scott, Mike. 2008. *WordSmith Tools* version 5. Liverpool: Lexical Analysis Software.

Sinclair, John. 1991. *Corpus, concordance, collocation: describing English language.* Oxford: Oxford University Press.

Sroka, Kazimierz A. 1972. *The syntax of English phrasal verbs*. The Hague: Mouton de Gruyter.