International Journal of English Language & Translation Studies
ISSN: 2308-5460

Extraction of Phrasal Verbs from the Comparable English Corpus of Legal Texts

[PP: 184-194]

Marija Bilić
Faculty of Humanities and Social Sciences, University of Split
Croatia

Angelina Gaspar
Faculty of Humanities and Social Sciences
Catholic Faculty of Theology,
University of Split, **Croatia**

ABSTRACT

This paper presents a corpus-based approach to semi-automatic extraction of English phrasal verbs, very productive, but complex and often non-transparent lexical units, via particles (prepositions, adverbs) they consist of and which are among the top-ranking functional words in the list of running words of the British National Corpus (BNC). The research is carried out on a comparable English corpus of publicly available legal texts consisting of 392 255 words and using *WordSmith Tools 6.0.* The evaluation of the system efficiency is conducted via the statistical measures of *Precision, Recall* and *F-measure*, whereas the list of phrasal verbs is checked against the reference source *Cambridge Phrasal Verbs Dictionary* (2015). The results show that the process of semi-automatic extraction of phrasal verbs requires a considerable human intervention as well as control via their verbal segments since it revealed instances of wrong phrasal verb usage. Furthermore, the results point to the low frequency of phrasal verbs in legal texts since they account for only 2% in the total number of words, and their unequal distribution since 5 most frequent phrasal verbs account for nearly half, and 25 for more than 90% of all such items. Finally, tendency towards nominalisation of phrasal verbs, which is in line with the nature of legal language, is evident, especially in the texts originally written in English.

Keywords: *Automatic Extraction, Particles, Phrasal Verbs, Comparable Corpus, Legal Language*

ARTICLE INFO	The paper received on	Reviewed on	Accepted after revisions on
	15/04/2018	**07/05/2018**	**10/07/2018**

Suggested citation:
Bilić, M. & Gaspar, A. (2018). Extraction of Phrasal Verbs from the Comparable English Corpus of Legal Texts. *International Journal of English Language & Translation Studies*. *6(2). 184-194.*

1. Introduction

The paper focuses on the possibility of the automatic extraction of phrasal verbs - a structure consisting of a verb and one or two morphologically invariable particles and acting as a unique lexical and semantic unit - from the comparable English corpus of legal texts, and the analysis of their presence and frequency in the legal texts originally written in English and translations in English.

English phrasal verbs are chosen for the analysis since they are one of the most characteristic and productive features of the English language, but also complex, and difficult to acquire due to their structural, syntactic and semantic features. Moreover, since they are multi-word units, they are also believed to pose a problem for the automatic extraction, and natural language processing, e.g. machine translation and computer-assisted translation.

Legal language is chosen for the analysis both for linguistic reasons since it is a genre characterised by unambiguousness, precision, repetition, concision, i.e. a genre in complete opposition with phrasal verbs which are very often polysemic, non-transparent and redundant (since they are multi-word units), as well as for purely practical reasons since the legislation of both the EU and the Republic of Croatia is publicly available. The following hypotheses are tested in this paper:

a. phrasal verbs can be semi-automatically extracted via particles (adverbs, prepositions) they consist of by using a key-word extraction program that gives a list of the most frequent words where functional words (adverbs, prepositions, articles, pronouns, etc.) are top-ranking words;

b. since phrasal verbs are a typical feature of the English language, their presence in domain-specific texts is statistically significant, regardless of their redundancy, polysemy and the principle of language economy;

c. distribution and frequency of phrasal verbs in English source texts differs from English translations.

2. Literature Review

Due to their diverse syntactic and semantic features, phrasal verbs have been attracting linguistic attention for the last 300 years or so (Thim, 2012). Firstly, scholars have been proposing many detailed descriptions and classifications, but eventually acknowledged the difficulty in making clear-cut distinctions between multi-word verbs, as many of them may belong to more than one category depending on the context. For example, *come back* may be interpreted either as a phrasal verb meaning ´to resume an activity´ or as a free combination meaning ´to return´ (Biber et al., 1999).

This research is based on Darwin and Gray`s (1999) alternative and inclusive approach according to which '[...] linguists should consider all verb + particle combinations to be potential phrasal verbs until they can be proven otherwise´, and extended version of their definition whereby all structures consisting of a verb proper and one or two morphologically invariable particle/s that function as a single lexical and semantic unit are considered as a phrasal verb.

Secondly, a lot of debate has been revolving around the use of phrasal verbs in different genres. While, for example, Dempsey et al. (2007) consider phrasal verbs as text genre identifiers since they are believed to be more common in spoken and informal registers, Fletcher (2005) believes that phrasal verbs are not just an informal version of 'purer' English since in many cases they fill important lexical gaps: that is, they express concepts for which there is no obvious single-word equivalent or single-word equivalents sound stilted or pompous (e.g. *put on/ don*). Thim (2012), however, simply believes that the traditional view of phrasal verbs as a typically English and particularly colloquial construction has its roots in the 18th century as the indirect result of a number of metalinguistic and stylistic factors which he describes as a *colloquialization conspiracy*, e.g. the normative verdicts against preposition stranding, monosyllables and pleonasm.

The aim of this research is to analyse the presence and frequency of phrasal verbs in the corpora of legislative texts of the EU and English translations of legislative texts of the Republic of Croatia due to the fact that, with the aim of promoting simplicity, unambiguity, precision, and economy in the drafting of EU legislative documents (Novak et al., 2003), the EU has prepared different legal acts on the quality of the drafting of EU legislation (e.g. *Birmingham Declaration* (1992), *Interinstitutional Agreement* (1999/C 73/01), *White Paper on European Governance* (2001), etc.) and guidelines (e.g. *Joint Practical Guide* (2003), *Interinstitutional style guide* (2011), etc.).

Likewise, the Croatian Ministry of Foreign Affairs and European Integration has, for the purposes of the translation of the Croatian legislation in English, a prerequisite for the accession of the Republic of Croatia to the EU which occurred on July 1, 2013, also prepared manuals and guidelines (e.g. *Priručnik za prevođenje pravnih akata Europske unije* (2003), *Priručnik za prevođenje pravnih propisa Republike Hrvatske na engleski jezik* (2006)) which actually incorporate parts of the abovementioned EU guidelines and whose aim is to achieve consistent and high-quality translations.

Therefore, this research will analyse the frequency of phrasal verbs, which are undoubtedly very often polysemic, not precise enough and not economic (since they are multi-word units) in the legal language which is, as mentioned above, characterised by concision, precision and impersonal style, i.e. features opposite to those of phrasal verbs, and in which nouns and noun groups prevail, qualitatively and quantitatively, over verbs and adjectives (Cabré, 1999).

Thirdly, there have been many attempts of automatic identification and extraction of phrasal verbs from electronic text corpora. Rehbein and Ruppenhofer (2017) presented a method for the automatic identification and extraction of causal relations from text, based on a large English-German parallel corpus. They succeeded in identifying and extracting 100 different types for causal verbal triggers, with only a small amount of human supervision. Dealing with compositionality in verb-particle construction, Bhatia et. al. (2017) identified the core senses of particles that have broad application across verb classes; the information was used while building computational lexicons. They demonstrated through an example how grammatical/semantic/ontological information that enables compositional parsing is used to obtain full semantic representation of sentences. Vincze (2017) investigated the behaviour of verb-particle

constructions in English questions, in three English corpora. The results showed that there are significant differences in the distribution of WH-words, verbs and prepositions/particles in sentences that contain VPCs and sentences that contain only verb+prepositional phrase combinations.

Exploring the role of prepositions in context, Gong et. al. (2017) revealed that sense-specific preposition representations not only encode semantic relations but aid paraphrasing of phrasal verbs when used in a simplistic compositional manner. Also, they explored the task of inferring the meaning of the phrasal verb from its components, i.e., the verb and preposition sense representation, casting that as a lexical paraphrasing task of finding one word that captures the meaning of the verb-particle construction (e.g. *climb down = descend; carry on = conducted*).

However, due to their flexible multi-word character and semantic richness that results in translation asymmetry, i.e. n:1 and n:n relationship, phrasal verbs will, undoubtedly, continue to pose a great challenge.

This aim of this research is to evaluate the efficiency of phrasal verb identification and extraction via particles and with the use of *WordSmith Tools* 6.0 developed by Mike Scott in 1996 at the University of Liverpool.

3. Methodology

This research is a part of a larger research conducted for the completion of the doctoral dissertation (Bilić, 2018) with the first phase entailing the creation of a bi-directional English-Croatian parallel corpus and comparable English and Croatian corpora of diversified legal texts consisting of 743 936 words in total.

3.1. Data on corpora

The EU parallel sub-corpus consists of 16 legal texts and its translations created in 2013 and 2014 and publicly available at the EUR-Lex portal.

Table 1: Statistical data on EU parallel sub-corpus

EU parallel sub-corpus			
N	**document**	**en_SL**	**hr-TL**
1.	Report 2013/C 365/35	4 036	3 576
2.	Report 2013/C 365/22	2 754	2 406
3.	Report 2013/C 365/14	3 154	2 744
4.	Commission Staff Working Document – Communication from the Commission	2 250	2 071
5.	Commission Staff Working Document – Proposal for a Regulation	5 698	5 315
6.	Supplementary rules	758	551
7.	Regulation (EU) No 1285/2013	15 285	12 715
8.	Regulation (EU) No 1286/2013	4 622	3 979
9.	Regulation (EU) No 1287/2013	9 367	8 254
10.	Regulation (EU) No 1288/2013	13 401	11 453
11.	Directive 2014/107/EU	17 342	16 233
12.	Directive 2014/104/EU	12 847	11 829
13.	Directive 2014/92/EU	18 568	16 774
14.	Directive 2014/90/EU	18 166	15 368
15.	Decision (2009/371/JHA)	20 809	17 150
16.	Decision No 1386/2013/EU	22 061	21 147
total		**171 118**	**167 596**

The Hr parallel sub-corpus consists of 8 legal texts and its translations created between 2005 and 2010 and downloaded from the CIDRA portal (today: Digital Information Documentation Office of the Government of the Republic of Croatia).

Table 2: Statistical data on Hr parallel sub-corpus

Hr parallel sub-corpus			
N	**document**	**cro-SL**	**en_TL**
1.	Law on Croatian Radio and Television	7 282	7 998
2.	Law on Police	11 512	13 392
3.	Media Act	7 664	10 102
4.	Regulation on Internal Organization of the Ministry of the Economy, Labour and Entrepreneurship	21 255	22 804
5.	Drug-abuse Prevention Act	9 087	9 723
6.	Maritime Code	96 781	120 094
7.	Ordinance on the Dossier Requirements for the Evaluation of Active Substances Contained in Plant Protection Products	26 602	31 354
8.	Ordinance on the Management of End-of-life Vehicles	3 902	5 607
total		**184 085**	**221 137**

Tables 1 and 2 show that, although texts originally written in English (en_SL) and Croatian (hr_SL) contain almost the same number of words, translations in English (en_TL) contain 29% more words than texts originally written in English (en_SL), due to the fact that, on the one hand translations in Croatian (hr_TL) contain 2% less words than texts originally written in English (en_SL), and, on the other hand, translations in Croatian (hr_TL) contain 20% more words than texts originally written in Croatian (hr_SL).

Whether the reason for such a difference in the number of words is related not only to the different nature of the two languages (English an analytic, and Croatian a syntactic language), but also to the different usage of phrasal verbs in English as a source and target language may be the focus of a further research.

3.2. Data on tools

For the purposes of this research, *WordSmith Tools* 6.0 (2015) and its programmes *WS KeyWord List*, *KWL*, *WS WordList*, *WL* and *WS Concord* are used.

a) *WS WordList*, *WL* is a programme that generates lists of words and word-clusters set out in alphabetical or frequency order, detailed statistics on the number and ratio of types and tokens, mean word length, number of sentences, paragraphs and sections.
b) *WS KeyWord List*, *KWL* is a programme that generates lists of words with the highest frequency in comparison with a reference set of words usually taken from a large corpus of text, e.g. British National Corpus (BNC)
c) *WS Concord tool* is a programme that gives a chance to see any word or phrase in context.

The evaluation of the system efficiency is conducted via the statistical measures of *Precision*, *Recall* and *F-measure*.

According to Lopes at al. (2010:251), *Precision* (P) indicates the capacity that the method has to identify the correct terms, considering the reference list, and it is calculated with the formula (1), which is the ratio between the number of terms found in the reference list (RL) and the total number of extracted terms (EL), i.e., the cardinality of the intersection of the sets RL and EL by the cardinality of set EL.

$$P = \frac{|RL \cap EL|}{|EL|} \quad (1)$$

Recall (R) indicates the quantity of correct terms extracted by the method and it is calculated through the formula (2).

$$R = \frac{|RL \cap EL|}{|RL|} \quad (2)$$

F-measure (F) is considered a harmonic measure between precision and recall, and it is given by the formula (3).

$$F = \frac{2 \times P \times R}{P + R} \quad (3)$$

3.3 Research phases

For the purposes of this research only the comparable English corpus (en_SL and en_TL) consisting of 392 255 words is analysed in terms of the presence and frequency of phrasal verbs.

The building of the corpora is followed by a research conducted on a sample corpus of 10 en_SL documents which includes the manual extraction of phrasal verbs as well as the testing of the possibility of the automatic extraction of phrasal verbs via particles they consist of using *WS KeyWord List*, *KWL* and *WS WordList*, *WL* programmes of the *WordSmith Tools* 6.0 (2015).

The list of phrasal verbs is checked against the reference dictionary *Cambridge Phrasal Verbs Dictionary* (2015) and the evaluation of the system efficiency is conducted via the statistical measures of *Precision*, *Recall* and *F-measure*.

The third phase of the research includes the repetition of the same steps on the whole comparable English corpus.

The fourth phase of the research includes the verification of the obtained list of phrasal verbs in the comparable English corpus via their verbal segment using *WordList* and *Concord* programmes of *WordSmith Tools* 6.0.

It is followed by a discussion on the similarities and differences between the two English subcorpora in terms of the presence and frequency of particles, as well as phrasal verbs.

4. Analysis and Discussion

4.1. Testing of automatic phrasal verb extraction via particles on sample en_SL corpus

The testing of the automatic extraction of phrasal verbs, i.e. structures consisting of a verb and one or two morphologically invariable particles and acting as a unique lexical and semantic unit, via particles they consist of is conducted on a sample corpus of 10 en_SL documents using *WordSmith Tools* 6.0.

Since the list of top 500 key words in the 10 en_SL corpus, obtained using the programme *WS KeyWord List*, contains 8 particles (*of*, *to*, *in*, *for*, *with*, *by*, *under* and *within)*, which is expected given the size of the sample corpus, as opposed to 28 particles in top 500 key words of BNC, other particles forming a phrasal verb are extracted using the programmes *WS WordList*, *WL* and *WS Concord tool*.

Table 3: List of particles which form phrasal verbs in the sample corpus

particles	total	in PV
of	3 024	16
to	1 949	306
in	1 586	10
for	775	37
with	572	34
on	533	65
by	434	0
at	194	21
from	168	12
under	157	0
out	106	97
through	83	0
up	73	42
down	49	49
over	35	13
across	23	0
about	15	1
along	6	0
around	4	0
back	3	2
off	3	3
total	**9 792**	**715**

Table 3 shows that only 20% tokens of the particle *to* form 42% of all phrasal verbs and that, given the total number of their tokens, *out*, *up*, *down*, *back* and *off*, which make only 2% of all particles and are not on the list of key words, are actually the particles that most often form a phrasal verb (rather than standing on their own), thus forming 32% of all phrasal verbs 27% of which goes on particles *out*, *up* and *down*.

On the basis of the data from Table 3 the extraction of phrasal verbs is conducted and the list of phrasal verbs, presented in Table 4, is created.

Table 4: List of phrasal verbs in the sample corpus

	PV	N		PV	N		PV	N		PV	N
1.	*refer to*	167	15.	*carry over*	10	29.	*start up*	3	43.	*go down*	1
2.	*relate to*	80	16.	*derive from*	9	30.	*write off*	3	44.	*lay out*	1
3.	*set out*	57	17.	*follow up*	9	31.	*engage in*	2	45.	*pass on*	1
4.	*lay down*	45	18.	*lead to*	9	32.	*open up*	2	46.	*rely on*	1
5.	*carry out*	38	19.	*allow for*	8	33.	*pay back*	2	47.	*roll out*	1
6.	*base on*	35	20.	*draw up*	8	34.	*stem from*	2	48.	*take on*	1
7.	*contribute to*	34	21.	*result in*	8	35.	*step up*	2	49.	*tide over*	1
8.	*associate with*	28	22.	*depend on*	8	36.	*pay off*	2	50.		
9.	*provide for*	24	23.	*deal with*	6	37.	*take over*	2	51.		
10.	*aim at*	21	24.	*account for*	5	38.	*take up*	2	52.		
11.	*set up*	20	25.	*pertain to*	5	39.	*bring about*	1	53.		
12.	*focus on*	14	26.	*build on*	4	40.	*build up*	1	54.		
13.	*consist of*	13	27.	*break down*	3	41.	*depart from*	1	55.		
14.	*amount to*	11	28.	*dispose of*	3	42.	*draw on*	1	56.		

total: 715

Table 4 shows that phrasal verbs have low frequency in the sample corpus of legal texts since they make only 2% of the total number of words[1], as well as uneven distribution since top 5 phrasal verbs make 54%, and top 25 phrasal verbs 93% of all phrasal verbs.

4.1.1. Evaluation of the WS WordSmith Tools 6.0 system efficiency

On the basis of the data from Table 4, the efficiency of the *WS WordSmith Tools* 6.0 system is evaluated. Since only 715 out of the total of 9 792 particles in the sample corpus form a phrasal verb, *Precision* (P) is very low and amounts to only 7.3%. *Recall* (R), the ratio between the automatically extracted phrasal verbs and the reference list of phrasal verbs created manually and containing 485 phrasal verbs, is high and amounts to 67.8%. *F-measure*, the harmonic measure between precision and recall, is expectedly low and amounts to 13.1%.

Thus, the results show that the automatic extraction of phrasal verbs via particles they consist of is possible but, since *Precision* is low, a considerable human intervention is needed in order to refine the results initially offered by the system. The measure of *Recall* shows that, regardless of the low level of *Precision*, the automatic extraction is more efficient than a purely manual method of extraction.

The semi-automatic extraction is, undoubtedly, a much faster, simpler and more organized method of research which offers many different possibilities of analysis, in this case particles which make such a small percentage in the total number of words of the sample corpus.

4.2. Creation of the list of phrasal verbs in the comparable English corpus

The creation of the list of phrasal verbs in the comparable English corpus (en_SL and en_TL) is preceded by the creation of the list of particles that constitute phrasal verbs. The list of phrasal verbs presents the level of their presence and frequency.

4.2.1. List of particles which constitute phrasal verbs

Since the list of top 500 key words, obtained using the programme *WS KeyWord List*, in the case of en_SL corpus contains 8 particles (*of, to, in, for, with, by, under* and *within)*, and in the case of en_TL corpus 6 particles (*of, for, by, on, under* and *from*), other particles forming a phrasal verb are extracted, as in the case of the sample corpus, using the programmes *WS WordList*, *WL* and *WS Concord tool*.

Table 5: List of particles which constitute phrasal verbs in the comparable English corpus

[1] PVs x 2 since they are multi-word units

	en_SL					en_TL			
	particle	total	in PV	%	% / PVs	total	in PV	%	% / PVs
1.	*of*	7939	20	0.3%	1.1%	13871	60	0.4%	2.6%
2.	*to*	5446	693	12.8%	38%	5616	1037	18.5%	44.6%
3.	*in*	4249	40	1%	2.3%	4810	77	1.7%	3.5%
4.	*for*	2101	89	4.3%	4.9%	2990	85	2.8%	3,6%
5.	*with*	1569	71	4.5%	3.9%	1380	27	2%	1.2%
6.	*by*	1443	0	0	0	1782	0	0	0
7.	*on*	1356	194	14.3%	10.6%	2187	191	8.8%	8.3%
8.	*at*	542	34	6.3%	1.9%	665	23	3.5%	1%
9.	*from*	526	25	5%	1.4%	1278	9	0.7%	0.4%
10.	*under*	404	0	0	0	288	0	0	0
11.	*out*	318	289	90.1%	15.7%	475	431	91.4%	18.7%
12.	*within*	218	0	0	0	480	0	0	0
12.	*up*	176	134	79%	7.6%	315	268	85.7%	11.6%
13.	*through*	174	0	0	0	100	0	0	0
14.	*down*	170	162	95.3%	8.8%	14	14	100%	0.6%
15.	*over*	66	14	21.2%	0.8%	96	36	36%	1.5%
16.	*about*	40	5	12.5%	0.3%	80	0	0	0
17.	*across*	37	0	0	0	4	2	50%	0.1%
18.	*forward*	14	3	21.4%	0.2%	22	0	0	0
19.	*forth*	12	12	100%	0.7%	6	6	100%	0.3%
20.	*along*	11	0	0	0	15	0	0	0
21.	*around*	11	0	0	0	5	0	0	0
22.	*back*	4	3	75%	0.2%	10	0	0	0
23.	*off*	3	3	100%	0.2%	10	3	30%	0.1%
24.	*aside*	1	1	100%	0.1%	14	14	100%	0.6%
25.	*away*	0	0	0	0	10	3	30%	0.3%
total		**26612**	**1792**			**36043**	**2286**		

Table 5 shows that the two English subcorpora considerably differ in terms of the overall frequency of particles since particles *of, on, from, out, up* and *about* are much more frequent in en_TL, and particles *under* and *down* in en_SL. Since the statistical measure of *Precision* is almost the same for the two English subcorpora (en_SL - 6.7%, en_TL - 6.3%) and very close to that for the en_SL sample corpus (7.3%), it can be concluded that any increase in the corpus size would probably generate similar results.

Furthermore, Table 5 shows that phrasal verbs in the two English subcorpora are formed by a similar number of different particles, i.e. 18 in en_SL, and 17 in en_TL, with 15 being the same. However, particles *for, with, on, at, from, forward, back* and *off* are more productive in en_SL although particles *for, on* and *from* are among the key words in en_TL, and particles *of, to, in, out, up, down* and *over* are more productive in en_TL, although particles *to* and *in* are among the key words of en_SL.

The particles listed among the top 500 key words constitute 50% phrasal verbs in en_SL, with 38% going on particle *to* and, only 15% phrasal verbs in en_TL, with 8% going on particle *on*, while particles *by* and *under* do not constitute any phrasal verb in the comparable English corpus.

Although they make only 3% (en_SL) and. 2% (en_TL) of all particles and are not on the list of key words, *out, up, down, forth* and *aside,* are the particles that more often enter the combination of a phrasal verb than they stand on their own, and constitute 34% (en_SL) and 32% (en_TL) of all phrasal verbs. These results are in line with those of the research conducted on the sample corpus. Furthermore, the two English subcorpora considerably differ in terms of the particles *back* and *off* which in en_SL mostly constitute phrasal verbs while in en_TL stand on their own. Taking into the consideration the frequency of particles in the total number of phrasal verbs, the two English subcorpora differ in terms of the particles *to* (en_SL - 38%; en_TL – 45%) and *down* (en_SL – 9%, en_TL - 1%).

The potential relationship between the differences in the overall frequency of the above mentioned particles and differences in the use of phrasal verbs in the two English subcorpora as well as the comparison with the results of Gardner et. al. (2007) in terms of the function of particles (adverbs or prepositions) forming a phrasal verb may be an interesting topic of a further research.

4.2.2. List of phrasal verbs in the comparable English corpus

Table 6: List of phrasal verbs in the comparable English corpus

	en_SL					en_TL				
	PV	N		PV	N		PV	N		PV	N
1.	*refer to*	406	39.	*dispose of*	4	1.	*refer to*	624	39.	*derive from*	2
2.	*relate to*	170	40.	*phase out*	4	2.	*carry out*	392	40.	*fill out*	2
3.	*lay down*	158	41.	*break down*	3	3.	*follow up*	219	41.	*interfere with*	2
4.	*set out*	143	42.	*lay out*	3	4.	*relate to*	214	42.	*rely on*	2
5.	*carry out*	134	43.	*open up*	3	5.	*pertain to*	150	43.	*tie up*	2
6.	*base on*	81	44.	*pay back*	3	6.	*base on*	122	44.	*turn over*	2
7.	*contribute to*	69	45.	*start up*	3	7.	*provide for*	76	45.	*break down*	1
8.	*provide for*	62	46.	*stem from*	3	8.	*engage in*	46	46.	*buy off*	1
9.	*associate with*	51	47.	*take over*	3	9.	*set out*	37	47.	*call for*	1
10.	*set up*	45	48.	*write off*	3	10.	*decide on*	37	48.	*call off*	1
11.	*draw up*	44	49.	*call on*	2	11.	*dispose of*	32	49.	*fill up*	1
12.	*aim at*	34	50.	*depart from*	2	12.	*draw up*	30	50.	*narrow down*	1
13.	*pass on*	32	51.	*detract from*	2	13.	*aim at*	23	51.	*pay off*	1
14.	*lead to*	29	52.	*join up*	2	14.	*associate with*	23	52.	*start up*	1
15.	*result in*	26	53.	*reach out*	2	15.	*consist of*	23			
16.	*depend on*	25	54.	*roll out*	2	16.	*depend on*	23			
17.	*rely on*	19	55.	*take on*	2	17.	*take over*	22			
18.	*derive from*	18	56.	*attend to*	1	18.	*take in*	19			
19.	*allow for*	17	57.	*back up*	1	19.	*conform to*	17			
20.	*consist of*	16	58.	*bring forward*	1	20.	*set aside*	14			
21.	*engage in*	14	59.	*build up*	1	21.	*amount to*	13			
22.	*amount to*	13	60.	*go down*	1	22.	*contribute to*	12			
23.	*follow up*	13	61.	*link up*	1	23.	*lay down*	12			
24.	*set forth*	12	62.	*put forward*	1	24.	*result in*	12			
25.	*deal with*	12	63.	*set aside*	1	25.	*hand over*	12			
26.	*focus on*	11	64.	*sign up*	1	26.	*impose on*	7			
27.	*step up*	11	65.	*take forward*	1	27.	*lead to*	7			
28.	*carry over*	10	66.	*take out*	1	28.	*stem from*	7			
29.	*build on*	10	67.	*tide over*	1	29.	*set forth*	6			
30.	*take up*	9				30.	*set up*	6			
31.	*call for*	6				31.	*account for*	5			
32.	*decide on*	6				32.	*deprive of*	5			
33.	*draw on*	6				33.	*take up*	5			
34.	*bring about*	5				34.	*lay up*	4			
35.	*pertain to*	5				35.	*allow for*	3			
36.	*combine with*	4				36.	*take away*	3			
37.	*account for*	4				37.	*come across*	2			
38.	*interfere with*	4				38.	*deal with*	2			
total:					**1 792**						**2 286**

Table 6 shows that phrasal verbs have low frequency in the comparable English corpus of legal texts since they make only 2% in the total number of words[2], which confirms the results of the research conducted on the sample corpus.

Top 5 phrasal verbs make 55% (en_SL), i.e. 69% (en_TL) of all phrasal verbs, and top 25 phrasal verbs 91% (en_SL), i.e. 96% (en_TL) of all phrasal verbs. In en_SL 48% of phrasal verbs appear less than 5 times, and 16% only once, while in en_TL 35% of phrasal verbs appear less than 5 times, and 13% only once. Therefore, it can be concluded that phrasal verbs are unevenly distributed in both English subcorpora which also confirms the results obtained for the sample corpus.

[2] PVs x 2 since they are multi-word units

En_TL contains greater number of phrasal verbs than en_SL which, on the other hand, contains more different phrasal verbs than en_TL (en_SL – 67; en_TL – 52).

Since 36 phrasal verbs are present in both English subcorpora, and represent more than 90% (en_SL - 93%; en_TL - 97%) of all phrasal verbs , it results that the comparison between the two English subcorpora is possible, regardless of the fact that en_TL subcorpus contains 29% more words than en_SL subcorpus.

However, there are considerable differences between the two English subcorpora in terms of the frequency of 36 phrasal verbs, especially top 5 phrasal verbs, as it is presented in Table 7.

Table 7: Phrasal verbs contained in both English subcorpora

	PV	en_SL	en_TL		PV	en_SL	en_TL
1.	*refer to*	406	624	19.	*consist of*	16	23
2.	*relate to*	170	214	20.	*engage in*	14	46
3.	*lay down*	158	12	21.	*amount to*	13	13
4.	*set out*	143	37	22.	*follow up*	13	219
5.	*carry out*	134	392	23.	*set forth*	12	6
6.	*base on*	81	122	24.	*deal with*	12	2
7.	*contribute to*	69	12	25.	*take up*	9	5
8.	*provide for*	62	76	26.	*call for*	6	1
9.	*associate with*	51	23	27.	*decide on*	6	37
10.	*set up*	45	6	28.	*pertain to*	5	150
11.	*draw up*	44	30	29.	*account for*	4	5
12.	*aim at*	34	23	30.	*interfere with*	4	2
13.	*lead to*	29	7	31.	*dispose of*	4	32
14.	*result in*	26	12	32.	*break down*	3	1
15.	*depend on*	25	23	33.	*start up*	3	1
16.	*rely on*	19	2	34.	*stem from*	3	7
17.	*derive from*	18	2	35.	*take over*	3	22
18.	*allow for*	17	3	36.	*set aside*	1	14

With the aim of explaining the differences between the two English subcorpora, further research should include a detailed analysis of the use of phrasal verbs in the comparable English corpus, both in terms of the context in which they appear, and their translation equivalents. In order to identify the phrasal verbs which are typical of the legislative texts, the list of phrasal verbs in the comparable English corpus should be compared to the list of 25 top phrasal verbs in general English, i.e. BNC (Gardner and Davies, 2007) and EU English, i.e. CEUE (Trebits, 2009).

As far as productivity is concerned, Table 6 shows that the most productive particles, in terms of the number of different verbs they collocate with in the phrasal verb combination, in en_SL are *up* (12), *on* (10), *to* (8), *out* (7), *from*, *for* and *with* (4), *down*, *forward* and *over* (3), *in* and *of* (2) while particles *at*, *forth*, *off*, *back* and *about* collocate with one verb only.

The most productive particles in en_TL are *up* (8), *to* (7), *on* (5), *for* (4), *down*, *in*, *off*, *out* and *over* (3), *of*, *with* and *from* (2) while particles *at*, *forth*, *aside*, *across* and *away* collocate with one verb only.

The most productive verbs in en_SL are *take* (5), *set* (4), *bring* (3), *build*, *bring*, *lay*, *carry*, *draw*, *result* and *call* (2), and in en_TL *set* and *take* (4) and *lay*, *call* and *fill* (2).

Therefore, it can be concluded that the most productive particles are *up*, *on* and *to*, and the most productive verbs are *take* and *set*.

The verb *take* is the most productive verb in Trebits (2009) as well, since it collocates with 8 different particles forming phrasal verbs *take away*, *take back*, *take forward*, *take off*, *take on*, *take out*, *take over* and *take up*.

However, it should be stated that the particle *to* and the verb *set* form a considerably greater number of phrasal verbs than other particles and verbs.

4.2.2.1. Derivatives from phrasal verbs

The Table 8 shows the list of nouns and adjectives derived from the phrasal verbs listed in the Table 6, which proves the fact that nominalisation is a feature of the legal language.

Table 8: Productivity of phrasal verbs

en_SL		en_TL	
derivatives	**N[3]**	**derivatives**	**N**
carry-over, n. (10)	10	*filled out*, adj. (2)	2
the drawing-up, n. (1)	44	*follow up*, n. (56); *follow-up*, n. (3); *follow up*, adj. (2)	219
the follow-up, n.(11); *the follow up*, n. (1); *follow-up*, adj. (1)	13	*laid-up*, adj. (1)	4
joined-up, adj. (2)	2	*paid-off*, adj. (1)	1
the passing-on, n. (12); *a pass-on*, n. (1); *passing- on*, adj. (1)	32	*the setting up*, n. (3)	6
the phasing-out, n. (1)	1	*start-up*, n. (1)	1
the setting up, n. (3); *the setting-up*, n. (1); *a set-up*, n. (2)	45		
start-up, n. (3)	3		
the take-up, n. (5); *the taking up*, n. (1)	9		
tide-over, adj. (1)	1		
Written-off, adj. (1)			

Table 8 shows that, on the one hand, en_TL contains more derivatives of phrasal verbs than en_SL (en_SL - 52 nouns and 8 adjectives; en_TL - 61 nouns and 9 adjectives) which are, on the other hand, more diversified in en_SL (8 nouns and 6 adjectives) than in en_TL (3 nouns and 3 adjectives).

Furthermore, Table 8 shows that in en_SL the derivatives of phrasal verbs are

distributed among *pass on* (14), *follow up* (12), *carry over* (10), *set up* (6) and *take up* (6), while in en_TL they are mostly related to one phrasal verb only, i.e. *follow up*.

Phrasal verbs which appear only in the form of nouns of adjectives are *carry over*, *follow up*, *join up*, *phase out*, *start up*, *tide over* and *write off* in en_SL, and *fill out*, *pay off* and *start up* in en_TL.

The most productive particle forming derivatives of phrasal verbs is the particle *up* (en_SL- 32, en_TL- 66).

Table 8 also points to the problematic use of a hyphen (-) in derivatives of phrasal verbs. En_SL contains 5 cases of nouns without a hyphen (*the follow up* (1), *the setting up* (3); *the taking up* (1)), while en_TL contains 59 cases of nouns without a hyphen (*the follow up* (56); *the setting up* (3)) and 7 cases of adjectives without a hyphen (*fill out* (2); *follow up* (5)). The results show that particularly problematic are nouns consisting of present participle and a particle as well as the derivatives of the phrasal verb *follow up*.

Since the rules of writing derivatives of phrasal verbs are specified in the Point 3.23-4 of the handbook for authors and translators in the European Commission, *English Style Guide* (2011), it may be concluded that the authors of the EU legislation and translators of the Croatian legislation have not been sufficiently using the resource prepared especially for them.

4.3. Verification of the list of phrasal verbs in the comparable English corpus via their verbal segment – REORGANIZED

Verification of the list of phrasal verbs in the comparable English corpus via their verbal segment resulted in the following findings:

- en_TL contains examples of the use of a wrong particle due to the probable interference with the Croatian as the source language

1) *… the party authorized* ***to dispose with**** *the cargo…* (12x)
2) *...the obligation* ***to contribute in*** *general average shall exist even when …* (2x)
3) *…carries out the following … activities* ***related with*** *the trade policy of …*

- tokens of certain phrasal verbs are not included in the initial list of phrasal verbs in the comparable English corpus obtained via particles due to:

a) particles being left out

4) *...the environment-related elements set out in the Commission's reform proposals, ...**backed** by the proposals for greening the Union budget under the Multi-Annual Financial Framework 2014–2020 are designed to...* (en_SL)
5) *…if* ***disposes*** *or leaves behind unattended the used needles or syringes…* (en_TL)
6) *...preventive measures are measures taken with a view to reducing the quantity of end-of-life vehicles,* ***pertaining*** *materials and substances comprised in motor vehicles...* (en_TL-2x)
7) *… which of these amounts the maritime liens or mortgage... are* ***related****...* (en_TL)
8) *...**carries** other activities within its scope...* (en_TL-3x)

b) misspelling

9) *...The Financing Section* ***carries our*** *prior review of texts of financial agreements and project summaries...* (en_TL)

c) the insertion of a great number of words between the verbal segment and the particle constituting a phrasal verb

en_SL:

10) *...**contribute**, in the context of the deployment and exploitation phases of the Galileo programme and the exploitation phase of the EGNOS programme,* ***to*** *the promotion and marketing of the services...* (2x)
11) *…that is,* ***referring*** *of such requests, issues, information or applications for cooperation* ***to*** *other competent authorities…* (5x)
12) *...an infringement of competition law* ***to*** *which the action for damages* ***relates****...*
13) *...**combining** a modernisation of the provisions on the machinery on the clearance of vacancies and applications for employment* ***with*** *the reinforcement of the delivery of the EURES service offer...*
14) *...performance criteria* ***on*** *which the allocation of budget funds between Member States for the actions managed by the national agencies should be* ***based****....*

en_TL:

15) *...and shall specify* ***to*** *which pledge creditors individual claims* ***pertain*** *and...* (9x)
16) *...only the ship* ***to*** *which the lien, mortgage or the claim* ***refers*** *can be intercepted...* (2x)
17) *...**referring** of such ... applications for cooperation* ***to*** *other competent authorities...*
18) *...a document is found* ***on*** *which the transferor's ownership right is* ***based****...*

The examples show that, in some cases, even more than 20 words can be inserted between the verb and the particle constituting a phrasal verb which results in the verb being too distant from the particle

to be shown in the *window of the WS Concord programme*.

- one-word derivatives of phrasal verbs are not included in the initial list of phrasal verbs obtained via particles

19) *...should be gradual and conditional on the successful completion of an appropriate* ***handover*** *review...* (2x)

20) *...would increase the annual* ***turnover*** *of the Union waste management and recycling sector by...* (2x)

21) *...the* ***rollout*** *of the updated Common Integrated Risk Analysis Model (CIRAM v 2.0) was initiated with translation into selected EU languages...*

It can be concluded that one-word nouns derived from phrasal verbs are not included in the initial list obtained via particles since *WordSmith Tools* 6.0 does not recognize them as multi-word nouns made of a verb and a particle. Therefore, for a more successful semi-automatic extraction of phrasal verbs via particles it would be better if they were written with a hyphen.

-nominal structures made of verbs that constitute a phrasal verb

En_SL contains 48 examples where the verb *relate* is used in the *noun-related +noun* structure, i.e. more often than in the phrasal verb *relate to*.

22) *...addressing international environmental and* ***climate-related*** *challenges...* (48x)

En_TL contains only 18 such examples but in 14 examples the hyphen is wrongly left out.

23) *...documentation for approval of* ***energy related investment plans****...*

En_SL contains 45 examples where the verb *base* is used in the *noun-based +noun* structure, while no such examples are found in en_TL.

24) *...the development of* ***market-based instruments*** *and indicators beyond GDP...*

These examples explain the differences in the frequency of phrasal verbs *relate to* and *base on* in the two English subcorpora, show the tendency towards creating fixed nominal structures derivated from phrasal verbs, which is in line with the overall tendency of the legal language towards nominalisation, as well as the aim of achieving greater language economy. Also, the use of the passive voice and the possibility of a great number of words being inserted between the verb and the particle forming the phrasal verb are avoided whereby their natural language processing, e.g. machine and computer-assisted translation, is facilitated.

Due to a considerable number of these nominal structures made of verbs that constitute a phrasal verb, further research in the *WordList WS* programme focused on the nouns *relation, basis* and *conformity* which act as synonyms of the phrasal verbs *relate to, base on* and *conform to* when used in the sentences. The research resulted in the following findings:

En_SL contains more examples of the structure *in relation to* than en_TL (en_SL-25; en_TL-14).

25) *...allows consumers to execute an unlimited number of operations* ***in relation to*** *the services referred to in paragraph 1.*

The structure *on the basis of* is more often used in the English comparable corpus than the phrasal verb *base on*, and more often in en_TL (172) than in en_SL (92)

26) *...****On the basis of*** *the results of that evaluation, the Commission shall decide...*

Instead of the phrasal verb *conform to*, en_SL contains structures such as *in conformity with* (16), and *conformity assessment procedures/ activities/ bodies/ results/ tasks/ certificate* etc. (75).

En_TL contains only 7 examples of *in conformity with* structure.

These examples further explain the differences in the frequency of phrasal verbs *relate to, base on* and *conform to* in the two English subcorpora and underline the tendency of the legal language towards nominalisation which is especially evident in the en_SL subcorpus.

5. Conclusion

This paper presents the process of the semi-automatic extraction of phrasal verbs via particles they consist of, and highlights the need for the verification of the obtained list of phrasal verbs via their verbal segment since it revealed examples of certain phrasal verbs being excluded from the initial list due to various reasons (e.g. particles being wrongly chosen, left out or misspelled, insertion of a great number of words between the verbal segment and the particle, and the problematic nature of one-word derivatives of phrasal verbs) as well as examples of certain nominal structures and phrases related to phrasal verbs the usage of which not only is in line with the tendency of the legal language towards nominalisation, but contributes to language economy and facilitates natural language processing.

The fact that the results of the research conducted on the whole comparable English corpus confirm the results of the research conducted on a sample of 10 en_SL legal

texts in terms of the low frequency and unequal distribution of phrasal verbs suggests that any further increase in the size of the comparable English corpora would probably generate similar results.

Although the two English subcorpora differ considerably in size, the research showed that the comparison is possible due to the fact that they share 36 phrasal verbs which represent more than 90% (en_SL - 93%; en_TL - 97%) of all phrasal verbs. Whether the difference in size is related not only to the different nature of the two languages (English an analytic, and Croatian a syntactic language), but also to the different usage of phrasal verbs in English as a source and target language, and application of different techniques in the translation process may be the focus of a further research. Furthermore, the reasons of a significantly different frequency of certain phrasal verbs, especially top 5 phrasal verbs, in the two English subcorpora may be found through a detailed analysis of the use of phrasal verbs in the comparable English corpus, both in terms of the context in which they appear, and their translation equivalents.

Mistakes identified in the process of the verification of the initial list of phrasal verbs via their verbal segment as well as the problematic usage of the hyphen in the case of the nouns and adjectives derived from phrasal verbs, underline the problematic structural and syntactic features of phrasal verbs, and predict the potential instances of their problematic semantic features.

References

Bhatia, A., Tehng, C.M., Allen, J. F. (2017). Compositionality in Verb-Particle Constructions. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017).* Valencia, Spain, April 4, 2017. Association for Computational Linguistics. 139–148.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow: Longman.

Bilić, M. (2018). Korpusna analiza engleskih fraznih glagola u jeziku prava. (Unpublished doctoral dissertation). Faculty of Humanities and Social Sciences, University of Zagreb.

Birmingham Declaration http://europa.eu/rapid/press-release_DOC-92-6_en.htm

Cabré, M. T. C. (1999). *Terminology: Theory, methods, and applications*. Amsterdam, Netherlands: John Benjamins.

Cambridge Phrasal Verbs Dictionary. (2006, 2015). Cambridge University Press.

CIDRA portal (today: Digital Information Documentation Office of the Government of the Republic of Croatia, http://www.digured.hr/)

Darwin, C. M., Gray, L. S. (1999). Going After the Phrasal Verb: An Alternative Approach to Classification. *TESOL Quarterly* 33 (1). 65-83.

Davies, M. (2004-). BYU-BNC. (Based on the British National Corpus from Oxford University Press). Dostupno na: http://corpus.byu.edu/bnc/

Dempsey K. B., McCarthy P. M., McNamara D. S. (2007). Using phrasal verbs as an index to distinguish text genres. In: Wilson, D., Sutcliffe (ed.). *Proceedings of the Twentieth International Florida Artificial Inteligence Research Society Conference.* Menlo Park, CA: The AAAI Press. 217-222.

EUR-Lex, http://eur-lex.europa.eu.

European Commission, Directorate General for Translation. (2011). *English Style Guide*. (7.ed.). Available at: http://ec.europa.eu/translation/english/guidelines/documents/styleguide_english_dgt_en.pdf

European Communities. (2003). *Joint Practical Guide of the European Parliament, the Council and the Commission*. Luxembourg: Office for Official Publications of the European Communities. Available at: http://bookshop.europa.eu/en/joint-practical-guide-of-the-european-parliament-the-council- and- the-commission-for-persons-involved-in-the-drafting-of-legislation-within-the-community- institutions-pbKA4502094/

European Union. (2011). *Interinstitutional style guide*. Brussels, Luxembourg. Available at: https://nellip.pixel-online.org/files/publications_PLL/11_Interinstitutional%20style%20guide%202011.pdf

Fletcher, B. (2005). Register and phrasal verbs. In: Rndell, M. (ed.) Macmillian Phrasal Verbs Plus. Oxford: Macmillian: LS 13-15. Available at: http://www.macmillandictionaries.com/MED- Magazine/September2005/33-Phrasal-Verbs-Register.htm [visited: 20.04.2014.]

Gardner, D., Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly* 41 (2). 339-359.

Gong H., Mu, J., Bhat, S., Viswanath, P. (2017). Prepositions in context. arXiv preprint arXiv:1702.01466

Interinstitutional agreement on better law-making http://eur-lex.europa.eu/legal-

content/EN/TXT/HTML/?uri=CELEX:32003Q1231%2801%29&from=HR

Lopes, L., Vieira, R., Finatto, M. J., Martins, D. (2010). Extracting compound terms from domain corpora. *Journal of the Brasilian Computer Society* 16. 247-259. Available at: http://www.inf.pucrs.br/~linatural/Docs/publica.pdf

Ministarstvo vanjskih poslova i europskih integracija. (2006). *Priručnik za prevođenje pravnih propisa Republike Hrvatske na engleski jezik*. Zagreb. Available at: http://www.mvep.hr/files/file/prirucnici/prirucnik_za_prevodenje_pravnih_propisa_RH.pdf

Novak, J. i sur. (2003). *Priručnik za prevođenje pravnih akata Europske unije*. Zagreb: Ministarstvo za europske integracije Republike Hrvatske. Available at: . http://www.mvep.hr/files/file/prirucnici/MEI_PRIRUCNIK.pdf

Rehbein, I., Ruppenhofer, J. (2017). Catching the Common Cause: Extraction and Annotation of Causal Relations and their Participants. *Proceedings of the 11th Linguistic Annotation Workshop*. Valencia, Spain, April 3, 2017. Association for Computational Linguistics. 105–114.

Scott, M. (1996). *WordSmith Lexical Analysis Tools Software*. Oxford: OUP. Available at: http://lexically.net/WordSmith Tools /purchase/

Scott, M. (2015). *WordSmith Tools 6.0.* Oxford: OUP

Thim, S. (2012). *Phrasal Verbs: The English Verb-Particle Construction and its History* (Topics in English Linguistics 78). Berlin and New York: De Gruyter Mouton.

Trebits, A. (2009). The most frequent phrasal verbs in English language EU documents−A corpus-based analysis and its implications. *System* 37 (3). 470-481. Available at: http://www.euenglish.hu/wp-content/uploads/2010/11/A.Trebits_Phrasal-verbs-in-EU-English_System.pdf

Vincze, V. (2017). Verb-Particle Constructions in Questions. *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain, April 4, 2017. Association for Computational Linguistics. 155–160.

White Paper on European Governance. Available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2001:0428:FIN:EN:PDF