Improving English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment

Iliana Simova
Dept. of Computational Linguistics
Saarland University, Germany
ilianas@coli.uni-saarland.de

Valia Kordoni
Dept. of English and American Studies
Humboldt-Universität zu Berlin, Germany
kordonie@anglistik.hu-berlin.de

Abstract

This work describes an experimental evaluation of the significance of phrasal verb treatment for obtaining better quality statistical machine translation (SMT) results. Phrasal verbs are multiword expressions used frequently in English, independent of the domain and degree of formality of language. They are challenging for natural language processing due to their idiosyncratic semantic and syntactic properties. The meaning of phrasal verbs is often not directly derivable from the semantics of their constituent tokens. In addition, they are hard to identify in text because of their flexible structure and due to ambiguous prepositional phrase attachments. The importance of the detection and special treatment of phrasal verbs is measured in the context of SMT, where the word-for-word translation of these units often produces incoherent results. Two ways of integrating phrasal verb information in a phrase-based SMT system are presented. Automatic and manual evaluations of the results reveal improvements in the translation quality in both experiments.

1 Introduction

Multiword expressions (MWEs) are units which consist of two or more lexemes and whose meaning is not derivable, or is only partially derivable, from the semantics of their constituents. Some examples are idiomatic expressions such as *take advantage of*, or *break a leg*, nominal compounds such as *traffic light*, and phrasal verbs, such as *hold up* and *take away*, which also can exhibit different degrees of semantic compositionality.

MWEs play an important role in natural language communication. They are used with high frequency and appear in various contexts in everyday and literary language, independent of genre and degree of formality. Jackendoff (1997) estimates that the amount of MWEs in a speaker's lexicon is nearly the same as the amount of single words.

The high frequency of usage, and the idiosyncratic semantic and syntactic properties of these constructions, indicate the need for their special handling in Natural Language Processing (NLP). From the perspective of semantics, MWEs need to be treated as units, because their meaning spans over word boundaries. From the perspective of syntax, however, these expressions are often hard to identify, because of their resemblance to ordinary verb or noun phrases. The MWE *kick the bucket*, for instance, which on the syntax level is just an ordinary verb phrase, can receive a very different semantic interpretation than the intended one, if not treated as a unit (Sag et al., 2002). In addition, while most MWEs have a relatively fixed structure, some allow for certain syntactic variations. Separable verb-particle constructions, for example, can appear in two forms: with their direct object separating the verb and particle or following them.

Several experiments to date have suggested that the special handling of MWEs is a necessary preliminary step to robust syntactic and semantic NLP and, as such, can lead to significant improvements in the performance of NLP applications.

Statistical Machine Translation (SMT) is a pro-

Proceedings of the Workshop on Multi-word Units in Machine Translation and Translation Technology, Nice, September 3, 2013.
Monti, J., Mitkov, R., Corpas Pastor, G. Seretan, V., eds.

totypical, we may say, task for which the appropriate treatment of MWEs can be beneficial, as suggested by a number of experiments which we thoroughly present in the following sections. As far as the alignment between the source and target language is concerned, MWEs constitute a major challenge, since it is very often the case that they do not receive exact translation equivalents. One example of an asymmetry caused by MWEs are phrasal verbs (PVs) in English to Bulgarian translation. In Bulgarian, phrasal verbs do not occur as multiword units, but are usually translated as single verbs. The word-for-word translation of PVs leads to incoherent translations or loss of information, in cases when the semantics of the PV can partly be derived from that of its verb and particle. The appropriate treatment of PVs could therefore improve translation quality in many of these cases.

The work we present in this paper concentrates on phrasal verbs in the context of English to Bulgarian phrase-based SMT, and is a pilot study for this language pair. The presented experiment aims at revealing the importance of the correct identification of phrasal verbs for improving the performance of an SMT system. We use two methods in order to integrate phrasal verb knowledge into the translation process. The significance of the choice of integration strategy is measured in an automatic and a manual evaluation. The manual evaluation furthermore aims at determining how the different integration mechanisms' performances are influenced by the levels of idiomaticity of the translated phrasal verbs.

The paper is structured as follows: Section 2 provides background information on the basic characteristics of phrasal verbs. Section 3 discusses some related works. Section 4 presents our experiments, as well as the language resources and tools which are used in them. Section 5 focuses on the evaluation results, which include manual evaluations of phrasal verb identification module, as well as manual and automatic evaluations of translation quality. We conclude with an outline of possible future research developments.

2 Phrasal Verbs as Multiword Units

Phrasal verbs are multiword expressions which can be divided into two classes with respect to their syntactic structure: verb particle constructions (VPCs), and prepositional verbs.

VPCs consist of a main verb and a particle, which can be an intransitive preposition (take *off*), an adjective (cut *short*), or a verb (let *go*) (Baldwin and Kim, 2010). These constructions are either intransitive (*come back*), or take a direct object argument (*call off* a meeting). However, this argument is subcategorized for by the VPC as a unit, and not by its particle or verb ([look up [^{obj} a word]], *[look [up [^{obj} a word]]]).

Variations in the structure of transitive VPCs are possible - some of them allow for the direct object of the verb to appear between the verb and particle, while others are strictly inseparable.

- Separable VPCs - the verb and particle may or may not be separated by the object (a); if the object is of pronominal type, it must appear between the verb and the particle (b).
  - (a) She *turned* the light *on*. She *turned on* the light.
  - (b) She *turned* it *on*. *She *turned on* it.
- Inseparable VPCs - the verb and particle must be adjacent.
  - (c) She *fell off* a tree. *She *fell* a tree *off*.
  - (d) She *fell off* it. *She *fell* it *off*.

In addition to the direct object of the VPC, only some non-manner adverbs (e.g., *right*, *back*, *straight*) may appear between the verb and the particle (Sag et al., 2002):

- (e) She *turned* the light back *on*.
- (f) *She *turned* the light quickly *on*.

Prepositional verbs consist of a verb and a transitive preposition (refer *to*, look *for*). Their structure is not as flexible as that of VPCs, and they never take the form of a separable construction since the direct object is an argument of the preposition.

Due to the surface similarity in the structure of VPCs, prepositional phrases, and ordinary verb-preposition combinations, the correct identification of the different classes is a major challenge ("The boy [looked] up at the sky.", "The boy [looked up [to [^{obj} his brother]]]"). In the current work, the focus is placed on trying to identify phrasal verbs, and avoid marking ambiguous constructions where verbs are simply modified by

prepositional phrases. No effort is made to distinguish VPCs from prepositional verbs.

Phrasal verbs exhibit different levels of semantic compositionality. In some cases their meaning cannot be directly derived from the semantics of their constituent tokens. For instance, the meaning of the verb *do in* in the sense of *tire, exhaust* cannot be inferred from *do* or *in*. In other cases the meaning of the phrasal verb is closer to the semantics of its components, and can be partially derived from it. These compositional/semi-compositional constructions usually have a verb which preserves its original meaning, and a particle which indicates direction (*carry in*), or a manner in which the action is performed (e.g., continuously: *go on*). Another example is the particle *up*, which, when combined with some verbs, denotes the completion of an action (*eat up*, in the sense of *finish eating*; *split up*, in the sense of *cease being together*).

2.1 Translation Asymmetries

Bulgarian lacks phrasal verbs in the form in which they appear in English. A VPC is usually mapped to a single verb in Bulgarian which preserves the original meaning. For instance[1]:

(1) to *put off* the decision
da *otlozhi* reshenieto
to postpone decision-the

(2) to *take over* peacekeeping operations
da *poemat* miroopazvashtite operacii
to take-over peacekeeping-the operations

(3) to *set out* the priorities
da *opredeljat* prioritetite
to define priorities-the

This mapping is many-to-many in cases when the equivalent Bulgarian verb has a reflexive form, marked by the reflexive particles 'se' or 'si'.

(4) to *give up* the search for an agreement
da *se otkazhe* da tyrsi sporazumenie
to give-up-refl to look-for agreement

Another case of many-to-many mapping is the 'da'-construction in Bulgarian. It is used to denote complex verb tenses, modal verb constructions, and subordinating conjunctions. In the example below the preferred alignment is between 'break off' and 'da prekysne', *(to) interrupt*.

[1]Examples were extracted from the SeTimes corpus sentence alignments

(5) should break off negotiations
trjabva da prekysne pregovorite
should (to) interrupt negotiations-the

In the current work's experiments no additional efforts are made to improve the word alignments in cases of many-to-many mapping between the tokens in source and target sentences. The extent to which the translation system itself is able to correctly use a reflexive particle where needed, or build the correct verb phrase involving a 'da'-construction, is reflected in the manual evaluations.

3 Multiword Expressions in Real-Life Applications like Statistical Machine Translation

To date, considerable effort has been devoted to detecting MWE types and tokens and including them in NLP applications that involve some degree of semantic interpretation. Approaches for their identification use a variety of linguistic and distributional features, ranging from syntactic and semantic flexibility (Ramisch et al., 2008; Fazly et al., 2009), collocation (Pearce, 2002) and parsibility scores (Zhang et al., 2006), as well as word alignment information (de Medeiros Caseli et al., 2010; Morin and Daille, 2010; Tsvetkov and Wintner, 2010), usually combined with association measures, such as pointwise mutual information (Evert and Krenn, 2005; Tsvetkov and Wintner, 2011). For the automatic identification of PV types, syntactic and semantic flexibility combined with association measures have resulted in an F-score of 90.1% (Ramisch et al., 2008). For PV tokens, an F-score of 97.4% was obtained using syntactic and semantic information like the selectional preferences of the verb and of the PV (Baldwin and Kim, 2010).

When it comes to real-life applications like machine translation, research has mainly focused on incorporating even simple treatments for MWEs in order to show that such an incorporation may improve translation quality. Carpuat and Diab (2010) adopt two complementary strategies for MWE integration: a static strategy of single-tokenization that treats MWEs as word-with-spaces and a dynamic strategy that keeps a record of the number of MWEs in the source phrase. They have found that both strategies result in improvement of translation

quality, which suggests that SMT phrases alone do not model all MWE information. Improvements were also presented in (Pal et al., 2010), who apply preprocessing steps like single-tokenization along with prior alignment and transliteration for named entities and compound verbs. Morin and Daille (2010) obtained an improvement of 33% in the French–Japanese translation of MWEs with a morphologically-based compositional method for backing-off when there is not enough data in a dictionary to translate an MWE (e.g. *chronic fatigue syndrome* decomposed as *[chronic fatigue] [syndrome]*, *[chronic] [fatigue syndrome]* or *[chronic] [fatigue] [syndrome]*).

When translating from and to morphologically rich languages like German, where a compound is in fact a single token formed through concatenation, Stymne (2009) proposes to deal with productivity and data sparseness by splitting the compound into its single word components prior to translation. Then, after translation, she applies some post-processing like the re-ordering or merging of the components, respecting possible annotations about compound membership and headedness. The adopted strategy for performing merging based on part-of-speech matching resulted in improvements in quality.

Another approach for minimizing data sparseness is adopted by Nakov (2008), who generates monolingual paraphrases to augment the training corpus. The basis for generating paraphrases that are nearly-equivalent semantically (e.g. *ban on beef import* for *beef import ban* and vice-versa) are the parse trees. They are syntactically transformed by a set of heuristics, looking at noun compounds and related constructions. This technique generates an improvement equivalent to 33%-50% of that of doubling training data. These results indicate that strategies like these for maintaining some information about the source MWEs during the translation process may help improve the quality of the translations in SMT systems.

Additional information about MWEs can also be obtained by the asymmetries between languages, where an MWE in a source language does not always correspond to an MWE in another, as we have also mentioned in the previous section. In this work the particular focus is on phrasal verbs (PVs), whose potential for syntactic flexibility and semantic idiomaticity can lead to problems in SMT.

4 English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment

4.1 Language Resources

The SeTimes[2] corpus contains parallel news articles available in nine Balkan languages including Bulgarian, and in English. The original version of the corpus is distributed as part of OPUS[3] and is aligned automatically at the sentence level. Efforts have been made to improve the quality of these alignments semi-automatically, resulting in a data set of 151,718 sentence pairs (Simov et al., 2012). Two additional manually annotated parallel SeTimes datasets[4] (2848 sentences) are available as part of the EuroMatrixPlus Project (Simov et al., 2012). The parallel data used for this work's experiment is a combination of the corrected version of SeTimes, and these two manually annotated sets.

In addition to a parallel resource, a large monolingual corpus is necessary for the creation of an accurate language model. A sub-corpus of about 50 million words from the Bulgarian National Reference Corpus[5] was chosen for this task.

4.2 Subtasks

Figure 1 shows the pipeline of this work's experiment. The architecture includes three main subtasks: preprocessing and data preparation, PV identification, and translation with integrated PV knowledge.

The English part of the parallel data was preprocessed with TreeTagger (Schmid, 1994), which provides part-of-speech tag and lemma information for each word. Similar annotations were automatically produced for the Bulgarian data with the help of the BTB-LPP tagger (Savkov et al., 2012). This is a necessary preliminary step for both the PV identification module and for translation. The PV identification system detects PVs in running text using lexicon look-up. Therefore in order for all occurrences to be detected it needs to operate on the lemma, instead of word level. The translation step employs a factored translation model (Koehn and Hoang, 2007), a suitable choice for this language pair and translation direction due to the rich morphology of Bulgarian.

[2] http://www.setimes.com
[3] http://opus.lingfil.uu.se/
[4] http://www.bultreebank.org/EMP/
[5] http://webclark.org/

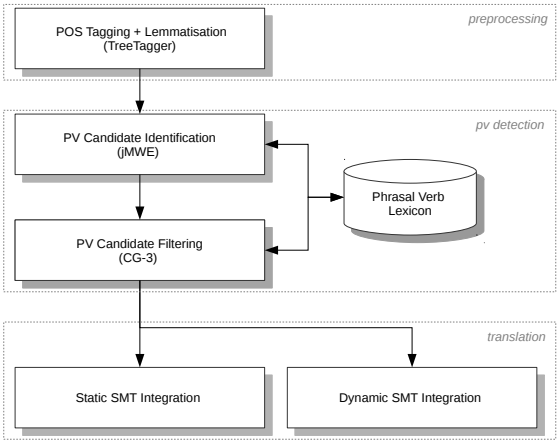

Figure 1: Pipeline of the experiment including phrasal verb detection and integration into the English part of the parallel corpora.

The PV detection step makes use of a lexicon of phrasal verbs, which was constructed from a number of resources. These include the English Phrasal Verbs section of Wiktionary[6], the Phrasal Verb Demon[7] dictionary, the CELEX Lexical Database (Baayen et al., 1995), WordNet (Fellbaum, 1998), the COMLEX Syntax dictionary (Macleod et al., 1998), and the gold standard data used for the experiments in (McCarthy et al., 2003) and (Baldwin, 2008). Most of these resources contain additional linguistic information about each PV, such as whether it is transitive or intransitive, separable or inseparable. This information was extracted together with the PVs where available and used to tackle the problem of ambiguous PP-attachments in the PV detection step.

PV candidates are detected in the source data with the help of the library for multiword expression detection jMWE (Kulkarni and Finlayson, 2011; Finlayson and Kulkarni, 2011). An additional module is employed as a post-processing step to filter out the spurious PV candidates. It is implemented in the form of a constraint grammar (Karlsson et al., 1995), and makes use of shallow parsing techniques, as well as the additional linguistic information extracted about the entries in the lexicon. The main idea behind the filtering mechanism is to define a number of positive contexts in which valid PV candidates would occur within a sentence. For example, valid contexts for a transitive separable phrasal verb are a noun phrase appearing between the verb and particle, or a noun phrase following the verb and particle. The grammar is thus able to mark cases like (b) as unsafe (in this case due to missing direct object).

take to, transitive, inseparable

(a) Peaceful demonstrators *took to* the streets this Saturday.

(b) The time it **took to* establish the full peacekeeping presence.

The information received from the PV identification step is used for two translation experiments. The two PV integration strategies are referred to as *static* and *dynamic*[8]. A baseline model, uninformed of the presence of PVs, is trained in addition to serve as basis for comparison between these techniques.

data set	number of sentences
test	800
development	100
tune	2000
train	the remaining (≈151K)

Table 1: Data sets created from the parallel corpus.

The parallel data was divided into development, tune, test, and training sets (Table 1). To better measure the influence of phrasal verb integration on translation quality, the test set sentences were chosen so that 50% of them (400 sentences) contain at least one detected PV occurrence. The rest of the sentences in the test set serve as means of establishing whether the PV integration has any negative effects when translating sentences without PVs, following the evaluations in (Kordoni et al., 2012). The development set was used for refining the constraint grammar for PV candidate filtering.

A phrase-based translation system was built with the following tools and settings: the Moses open source toolkit (Koehn et al., 2007) was used to build a factored translation model. The parallel data was aligned with the help of GIZA++ (Och and Ney, 2003). Two 5-gram language models were built with the SRI Language Modeling

[6]http://en.wiktionary.org/wiki/Category:English_phrasal_verbs
[7]http://www.phrasalverbdemon.com/

[8]terminology adopted from (Carpuat and Diab, 2010). The *dynamic* strategy is slightly altered to use binary features.

Toolkit (SRILM[9]) (Stolcke, 2002) on the preprocessed monolingual data from the Bulgarian National Reference Corpus to model word and part-of-speech tag n-gram information.

This choice of translation model is motivated by data sparsity issues due to the rich morphology of Bulgarian. When translating between a language with poor morphology and a highly inflected language, traditional translation models which use only word information often produce poor results because inflected forms of the same word are treated as separate tokens. A very large parallel resource is necessary to observe examples of translations for all inflected forms of the same word during training. To overcome this issue we use a factored model which operates on a more general representation than surface word forms, and is thus able to establish a better mapping between the source and target translation equivalents in the data. In the current experiment translation is carried out using lemma and part-of-speech information. English lemmas and part-of-speech tags are translated into their Bulgarian equivalents. The target word form is then produced in a *generation* step using the translated lemma and tag as input.

In the static integration constituent tokens of phrasal verbs are concatenated via underscores and are thus treated as single words. They can be seen as *static* expressions in the sense that their semantics becomes no longer derivable from the semantics of the tokens they consist of (Carpuat and Diab, 2010).

The static integration approach can enhance translation quality in several aspects. The technique is effective at improving alignments between source and target sentences, increasing the number of consistent examples of each expression in the training data (separable PVs in joined or split form obtain the same surface realization), and decreasing translation inconsistencies caused by ambiguous prepositional phrase (PP) attachments.

In the *dynamic* phrasal verb integration approach no modifications are made to the parallel data. The word alignment and training processes are not influenced externally in any way as well. Instead, a binary feature is included in the automatically extracted translation table of the system to indicate the presence of phrasal verb instances in the source English phrase.

[9]http://www-speech.sri.com/projects/srilm/

Incorporating this feature into the translation table helps improve translation quality in a more *dynamic* way in comparison with the *static* approach, in the sense that the translation system decides at decoding time how to segment and translate each input sentence (Carpuat and Diab, 2010). In the static approach, on the other hand, the treatment of each phrasal verbs as a unit is enforced due to their concatenation, and the approach is therefore more liable to errors in the PV detection process.

In the following section we give an in-depth analysis of the results obtained by the baseline, static and dynamic integration.

5 Evaluation Results

5.1 Phrasal Verb Identification Evaluation

The evaluations of the performance of the phrasal verb identification module were manually carried out on the test set consisting of 800 sentences, in half of which the PV detection system found at least one PV occurrence. The metrics used for this evaluation include *Precision*, *Recall* and F_1 score. In the context of the current experiment, *Precision* is defined as the amount of correct phrasal verbs identified by the module out of all discovered phrasal verbs. *Recall* is the amount of correct phrasal verbs out of all phrasal verbs instances present in the data, including the ones which the detection system has missed. F_1 score can be interpreted as the harmonic mean of Precision and Recall.

Manual evaluations revealed that the phrasal verb identification module managed to correctly detect 375 expressions out of 410 found in total. The system missed 28 PV occurrences. This results in Precision of 91%, Recall of 93%, and F_1 score of 92%.

The most common cause of errors were ambiguous PP-attachments. Recall was decreased mainly due to the restrictive nature of the constraint grammar filtering mechanism, and because of missing lexical entries in the PV lexicon.

5.2 Automatic Evaluation of Translation Quality

Table 2 presents the BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) scores obtained for the baseline system, and the static and dynamic integration strategies. The three experiments were evaluated once only for sentences with detected

PV instances (1), once for the part of the corpora with no detected PVs (2), and once for the whole data (3).

	with PVs (1)		no PVs (2)		all (3)	
	bleu	nist	bleu	nist	bleu	nist
baseline	0.244	5.97	0.228	5.73	0.237	6.14
static	0.246	6.02	0.230	5.76	**0.239**	6.18
dynamic	0.250	5.92	0.226	5.54	**0.244**	6.02

Table 2: Automatic evaluation of translation.

Sentences with phrasal verbs consistently receive higher BLEU and NIST scores than those without. The *static* integration strategy brings slight improvements in both scores for all three measurements. It can be safely concluded that it has no negative impact on translations of sentences without phrasal verbs. The best performing model according to BLEU is the *dynamic* one. However, it leads to a slight decrease in NIST for all experiments. In cases of sentences without PV instances, this approach gives a slight decrease in BLEU score, and a more noticeable one in terms of NIST.

The differences in BLEU and NIST scores for the two integration strategies suggest that they influence the translation process in different ways. The decrease in NIST over the baseline indicates that the dynamic system tends to use less informative n-grams. The static method, on the other hand, consistently obtains slightly higher NIST than the baseline.

To get a better insight on how the three models deal with the translation of phrasal verbs, we propose a more detailed discussion of the results in the following section.

5.3 Manual Evaluation of Translation Quality

The translations of each sentence in the test data which contains correctly identified phrasal verbs were considered, taking into account the phrasal verb itself and a limited context. The translations were divided into the following categories, following the evaluations in (Kordoni et al., 2012):

- *good* - correct translation of the phrasal verb, correct verb inflection;
- *acceptable* - correct translation of the phrasal verb, wrong inflection (also when a reflexive particle is missing, or a *da-construction* is not built correctly);
- *incorrect* - incorrect translation, which modifies the original sentence meaning;

The percentage of good, acceptable, and incorrect translations per integration approach is presented in Table 3. Only the correctly identified phrasal verb instances (375) and their contexts were taken into account.

	translation quality		
	good	acceptable	incorrect
baseline	0.21	0.41	0.39
static	0.25	0.51	0.24
dynamic	0.24	0.51	0.25

Table 3: Manual evaluation of translation

The evaluations confirm that the two integration strategies bring improvements in translation quality over the baseline. The best performance was achieved by the static approach, with 25% good and 51% acceptable translations, closely followed by the dynamic approach, with 24% good and 51% acceptable translations.

The evaluations further reveal that cases of separable PVs where the verb and particle(s) were not adjacent in the sentence are best handled with the static technique. It produced nearly twice as much *acceptable* translations compared to the other two. Even though the dynamic approach managed to handle several instances better than the baseline, overall it could not cope well with these expressions.

Table 4 summarizes the results obtained by the systems when taking into account the semantic properties of the translated expressions. PV instances in the data were divided into idiomatic and compositional, the latter including semi-compositional phrasal verbs such as *eat up*.

The static approach handles better idiomatic expressions than it does compositional ones. The opposite tendency is present for the baseline and dynamic model evaluations: the amount of acceptable translations they produce is higher for the compositional cases. Idiomatic expressions are best translated with the static approach. It produces 14% good and 26% acceptable translations. Compositional cases, on the other hand, are

handled best with the dynamic integration, which yields 12% good and 27% acceptable translations.

	translation quality					
	good		acceptable		incorrect	
	i+	i-	i+	i-	i+	i-
baseline	0.10	0.10	0.18	0.23	0.20	0.19
static	0.14	0.11	0.26	0.25	0.08	0.16
dynamic	0.12	0.12	0.25	0.27	0.11	0.14

Table 4: Manual evaluation of translation quality w.r.t semantic compositionality of the phrasal verbs (idiomatic: i+; compositional: i-).

The static approach outperforms the other two when dealing with separable verb-particle constructions and with idiomatic expressions. It is, however, most liable to errors in the PV detection process and relies on a wide-coverage phrasal verb dictionary for good results. In several examples errors were caused because the concatenated phrasal verb form was simply not found in the training data.

Even though the dynamic method achieved the highest BLEU score, its performance was not standing out during the manual evaluations. The only exceptions were some cases of compositional phrasal verbs. The performance of the dynamic approach was disappointing for cases of separable verb-particle constructions in a split form, where it did nearly as badly as the baseline.

6 Conclusion

The presented work was designed as an experimental evaluation of the significance of phrasal verb identification and analysis for the performance of an English-to-Bulgarian SMT system. The phenomenon of phrasal verbs is not observable in Bulgarian, and therefore an alignment asymmetry is introduced for the language pair. The phrasal verb constituents in the source language are usually aligned to a single verb equivalent in the target language. A module which employs lexicon look-up and shallow parsing techniques was developed to detect instances of phrasal verbs in the source English part of the parallel corpus. In order to minimize the risk of detecting spurious expressions, additional linguistic factors in terms of the transitivity and separability properties of the entries were brought into the detection process. This resulted into 92% F1-score of the detection module on the test set sentences.

Two integration strategies were used to incorporate information on the detected phrasal verb occurrences into a factored translation system. The first strategy encodes phrasal verbs as static units by concatenating their constituents via underscores. The second approach includes phrasal verb information into the translation table of the system in the form of a binary feature. Automatic and manual evaluations both showed that these approaches improve the translation quality over a standard baseline model. Manual evaluations further revealed that the different integration strategies have certain strengths and weaknesses associated with them, and therefore influence the translation process in a complementary way.

The evaluation results revealed that compositional phrasal verbs tend to be handled better with the dynamic strategy. The static one often led to loss of information when translating these cases, but performed better for sentences containing idiomatic phrasal verbs. This suggests the possibility for defining a targeted approach for phrasal verb integration. It would treat idiomatic phrasal verbs with the static, and compositional phrasal verbs with the dynamic technique, and thus combine the strengths of the two methods.

The targeted approach constitutes one possible way of future development for this work. There is room for improvement in the current integration pipeline. Minimizing errors in the PV identification task is just one of the goals which could be pursued. Besides the targeted approach, our research could be extended to include and compare additional integration strategies, such as the augmenting of the translation table with a bilingual phrasal verb dictionary. Set up in this way, the pipeline allows for other multiword phenomena to be studied with little additional effort for their integration. It would be interesting to investigate the translation of other semi-fixed multiword expressions which allow for discontinuous elements (e.g., *decomposable idioms* and *light verb constructions* (Sag et al., 2002)), and are thus often problematic to identify and interpret.

References

Baayen, R. H., R. Piepenbrock, and L. Gulikers. 1995. The celex lexical database (cd-rom).

Baldwin, Timothy and Su Nam Kim. 2010. Multiword expressions. In Indurkhya, Nitin and Fred J. Damerau, edi-

tors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

Baldwin, Timothy. 2008. A resource for evaluating the deep lexical acquisition of english verb-particle constructions. In *Proceedings of the LREC 2008 Workshop: Towards a Shared Task for Multiword Expressions*, MWE 2008, pages 1–2. European Language Resources Association.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment - Volume 18*, MWE '03, pages 65–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bick, Eckhard. 2009. Basic constraint grammar tutorial for cg3 (vislcg3).

Carpuat, Marine and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.*, HLT '10., pages 242–245, Stroudsburg, PA, USA. Association for Computational Linguistics.

de Medeiros Caseli, Helena, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation Special Issue on Multiword expression: hard going or plain sailing.*, pages 59–77.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research.*, HLT '02., pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Evert, Stefan and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language Special issue on MWEs*, pages 450–466.

Fazly, Afsaneh, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Comput. Linguist.*, pages 61–103.

Fellbaum, Christiane. 1998. Wordnet: An electronic lexical database.

Finlayson, Mark Alan and Nidhi Kulkarni. 2011. Detecting multiword expressions improves word sense disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World.*, MWE '11., pages 20–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jackendoff, Ray. 1997. *The Architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila. 1995. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text (Natural Language Processing, No 4).* Mouton de Gruyter, Berlin and New York.

Kim, Su Nam and Timothy Baldwin. 2007. Detecting compositionality of english verb-particle constructions using semantic similarity. In *Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 40–48.

Kim, Su Nam and Timothy Baldwin. 2010. How to pick out token instances of english verb-particle constructions. In *Journal of Language Resources and Evaluation (LRE) : Special Issue on Multiword Expressions: hard going or plain sailing?*, pages 97–113. Language Resources and Evaluation.

Koehn, Philipp and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.*, pages 868–876. Association for Computational Linguistics.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL).* Association for Computational Linguistics.

Kordoni, Valia, Carlos Ramisch, and Aline Villavicencio. 2012. Error analysis and the role of compositionality for high quality translation of phrasal verbs. Manuscript submitted for publication.

Kulkarni, Nidhi and Mark Alan Finlayson. 2011. jmwe: A java toolkit for detecting multi-word expressions. In *Proceedings of the 2011 Workshop on Multiword Expressions*, pages 122–124. Association for Computational Linguistics.

Li, Wei, Xiuhong Zhang, Cheng Niu, Yuankai Jiang, and Rohini Srihari. 2003. An expert lexicon approach to identifying english phrasal verbs. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1.*, ACL '03., pages 513–520, Stroudsburg, PA, USA. Association for Computational Linguistics.

Macleod, Catherine, Adam Meyers, and Ralph Grishman. 1998. Comlex english syntax lexicon.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Morin, Emmanuel and Béatrice Daille. 2010. Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation Special Issue on Multiword expression: hard going or plain sailing.*, pages 79–95.

Nakov, Preslav. 2008. Improved statistical machine translation using monolingual paraphrases. In *Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 338–342, Amsterdam, The Netherlands, The Netherlands. IOS Press.

Och, Franz Josef and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Pal, Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay, and Andy Way. 2010. Handling named entities and compound verbs in phrase-based statistical machine translation. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 46–54. Coling 2010 Organizing Committee.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.*, ACL '02., pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pearce, Darren. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on Language Resources and Evaluation, LAS*.

Ramisch, Carlos, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning.*, pages 49–56, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 47–54, Singapore, August. Association for Computational Linguistics.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing.*, CICLing '02., pages 1–15, London, UK. Springer-Verlag.

Savkov, Aleksandar, Laska Laskova, Stanislava Kancheva, Petya Osenova, and Kiril Simov. 2012. Linguistic processing pipeline for bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).*, Istanbul, Turkey. European Language Resources Association (ELRA).

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing.*

Simov, Kiril, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. Design and implementation of the bulgarian hpsg-based treebank. In *Erhard Hinrichs and Kiril Simov, editors, Journal of Research on Language and Computation.*, pages 495–522. Kluwer Academic Publishers.

Simov, Kiril, Petya Osenova, Laska Laskova, Stanislava Kancheva, Aleksandar Savkov, and Rui Wang. 2012. HPSG-based Bulgarian-English statistical machine translation. *Littera et Lingua*, Spring Issue.

Stolcke, Andreas. 2002. Srilm - an extensible language modeling toolkit. In *John H. L. Hansen and Bryan Pellom, editors, Proceedings of the Seventh International Conference on Spoken Language Processing.*, pages 901–904. International Speech Communication Association.

Stymne, Sara. 2009. A comparison of merging strategies for translation of German compounds. In *Proceedings of the Student Research Workshop at EACL 2009*, pages 61–69, Athens, Greece. Association for Computational Linguistics.

Tsvetkov, Yulia and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China. Coling 2010 Organizing Committee.

Tsvetkov, Yulia and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 836–845, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Villavicencio, Aline, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1034–1043, Prague, Czech Republic, June. Association for Computational Linguistics.

Zhang, Yi, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In *Proceedings of the COLING/ACL Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 36–44. Association for Computational Linguistics.