

# Multiword Expressions in Machine Translation

Valia Kordoni, Iliana Simova

Humboldt-Universität zu Berlin, Saarland University (Germany)  
Unter den Linden 6 10099 Berlin, Im Stadtwald 66143 Saarbrücken (Germany)  
kordonie@anglistik.hu-berlin.de, ilianas@coli.uni-saarland.de

## Abstract

This work describes an experimental evaluation of the significance of phrasal verb treatment for obtaining better quality statistical machine translation (SMT) results. 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.

**Keywords:** Multiword Expressions, Phrasal Verbs, Machine Translation

## 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.

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.

## 2. 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<sup>1</sup>:

- (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

<sup>1</sup>Examples were extracted from the SeTimes corpus sentence alignments

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

## 3. English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment

### 3.1. Language Resources

The SeTimes<sup>2</sup> 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<sup>3</sup> 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<sup>4</sup> (2848 sentences) are available as part of the EuroMatrix-Plus 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 mono-lingual 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<sup>5</sup> was chosen for this task.

### 3.2. 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

<sup>2</sup><http://www.setimes.com>

<sup>3</sup><http://opus.lingfil.uu.se/>

<sup>4</sup><http://www.bultreebank.org/EMP/>

<sup>5</sup><http://webclark.org/>

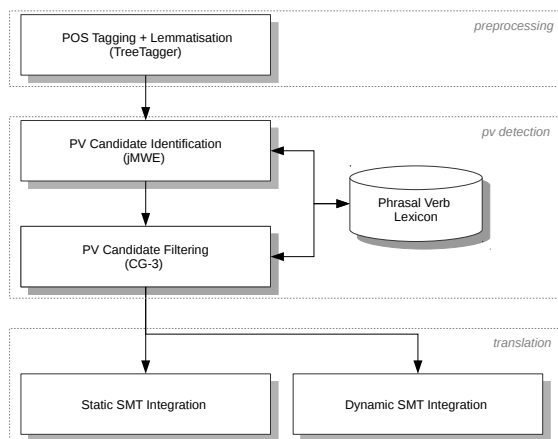


Figure 1: Pipeline of the experiment including phrasal verb detection and integration into the English part of the parallel corpora.

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.

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<sup>6</sup>, the Phrasal Verb Demon<sup>7</sup> 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 grammar is able to mark cases like (b) as unsafe (in this case due to missing direct object).

*take to*, transitive, inseparable

<sup>6</sup>[http://en.wiktionary.org/wiki/Category:English\\_phrasal\\_verbs](http://en.wiktionary.org/wiki/Category:English_phrasal_verbs)

<sup>7</sup><http://www.phrasalverbdemon.com/>

- (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*<sup>8</sup>. 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 ( $\approx 151K$ )

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 Toolkit (SRILM<sup>9</sup>) (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 tar-

<sup>8</sup>terminology adopted from (Carpuat and Diab, 2010). The *dynamic* strategy is slightly altered to use binary features.

<sup>9</sup><http://www-speech.sri.com/projects/srilm/>

get 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).

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.

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.

## 4. Evaluation Results

### 4.1. 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	<b>0.239</b>	6.18
dynamic	0.250	5.92	0.226	5.54	<b>0.244</b>	6.02

Table 2: Automatic evaluation of translation.

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.

### 4.2. 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 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

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.

## 5. Outlook

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.

## 6. References

- Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The celex lexical database (cd-rom).
- Baldwin, T. (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.
- Carpuat, M. and Diab, M. (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.
- Doddington, G. (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.
- Fellbaum, C. (1998). Wordnet: An electronic lexical database.
- Finlayson, M. A. and Kulkarni, N. (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.
- Karlssohn, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text (Natural Language Processing, No 4)*. Mouton de Gruyter, Berlin and New York.
- Koehn, P. and Hoang, H. (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, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.
- Kordoni, V., Ramisch, C., and Villavicencio, A. (2012). Error analysis and the role of compositionality for high quality translation of phrasal verbs. Manuscript submitted for publication.
- Kulkarni, N. and Finlayson, M. A. (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.
- Macleod, C., Meyers, A., and Grishman, R. (1998). Complex english syntax lexicon.
- McCarthy, D., Keller, B., and Carroll, J. (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.
- Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.
- Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (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, A., Laskova, L., Kancheva, S., Osenova, P., and Simov, K. (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, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*.
- Simov, K., Osenova, P., Laskova, L., Kancheva, S., Savkov, A., and Wang, R. (2012). Hpsg-based bulgarian-english statistical machine translation. *Littera et Lingua*, Spring Issue.
- Stolcke, A. (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.