How hard is it to automatically translate phrasal verbs from English to French?

Carlos Ramisch and **Laurent Besacier** and **Alexander Kobzar**
LIG-GETALP, BP 53
38041 Grenoble Cedex 9
France
{FirstName.LastName}@imag.fr

Abstract

The translation of English phrasal verbs (PVs) into French is a challenge, specially when the verb occurs apart from the particle. Our goal is to quantify how well current SMT paradigms can translate split PVs into French. We compare two in-house SMT systems, phrase-based and hierarchical, in translating a test set of PVs. Our analysis is based on a carefully designed evaluation protocol for assessing translation quality of a specific linguistic phenomenon. We find out that (a) current SMT technology can only translate 27% of PVs correctly, (b) in spite of their simplistic model, phrase-based systems outperform hierarchical systems and (c) when both systems translate the PV similarly, translation quality improves.

1 Introduction

For a long time, MT research has been struggling to deal with a certain number of unsolved problems, which reduce the usability and the utility of MT. One of these problems — a particularly hard one — is the translation of multiword expressions like noun compounds (*dry run, vacuum cleaner*), idioms (*set the bar high, French kiss*) and phrasal verbs (*make up, think through, sit down*).

A *multiword expression* is a combination of at least two lexical units that presents idiosyncratic behaviour at some level of linguistic analysis (Baldwin and Kim, 2010). Often, the lexical, syntactic and semantic idiosyncrasies of multiword expressions are at the root of translation problems, as exemplified in Table 2.

In the current dominant trend, *statistical machine translation* (SMT), transfer models are automatically learnt from sentence-aligned corpora. SMT paradigms have evolved from simple word-based models (Brown et al., 1993) to more sophisticated phrase-based models (Koehn et al., 2003) and hierarchical models (Chiang, 2007), where translation units are word sequences and trees instead of single words. Implicitly, these models capture some kinds of multiword expressions, like common noun compounds. However, due to their non-compositional semantics, unpredictable syntax, polysemous, productive and creative uses, many types of multiword expressions are not properly dealt with by state-of-the-art SMT systems.

English *phrasal verbs* (PVs) like *take off, give up* and *pull out* represent a particularly challenging class of multiword expressions for MT. The goal of this paper is to quantify how hard it is for current MT technology to translate these constructions. We focus on split PV occurrences because, as explained in Section 3, these constructions present a specific syntactic and semantic behaviour that makes them intuitively hard to model in current MT paradigms.

We want to evaluate the quality of PV translation in phrase-based and hierarchical English-French SMT systems. Therefore, we design and apply a generic evaluation protocol suitable to circumscribe a particular linguistic phenomenon (in our case, PVs) and manually annotate translation quality. Automatic evaluation measures such as BLEU and METEOR estimate the similarity between candidate and reference translations by comparing their n-grams. In our case, manual annotation is crucial, because these automatic metrics do not provide insights into the nature of errors. Our analysis aims to answer the questions:

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.

- What proportion of PVs is translated correctly/acceptably by each SMT paradigm?
- Which MT paradigm, phrase-based or hierarchical, can better handle these constructions?
- What are the main factors that influence translation quality of PVs?

2 Related work

One way to identify if a construction is a multiword expression is via word by word translation into another language: if the result is not successful, then the construction is probably a multiword expression (Manning and Schütze, 1999, p. 184). In other words, multiword expressions induce lexical and grammatical asymmetries between languages, as an expression in one language may be realized differently in another language.

It was not until recently that multiword expressions became an important research topic in SMT. Recent results show that incorporating even simple treatments for them in SMT systems can improve translation quality. For instance, Carpuat and Diab (2010) adopt and compare two complementary strategies: (a) they perform static retokenisation, representing expressions as words with spaces before word alignment, and (b) they add a feature to dynamically count the number of expressions in the source phrase. They use multiword Wordnet entries and experiment with an English-Arabic system, showing that both strategies result in improvement of translation quality in terms of automatic evaluation measures (BLEU, TER).

Other simplistic techniques that have been employed to integrate bilingual lexicons into standard SMT systems include (a) concatenating the lexicon to the training parallel corpus, and (b) artificially appending the lexicon (enriched with artificial probabilities) to the system's phrase table. This has been applied to Chinese-English terminology (Ren et al., 2009) and English-French nominal expressions (Bouamor et al., 2012). However, results are reported in terms of automatic measures and improvements are not always convincing.

For translating noun compounds from and to morphologically rich languages like German, where a compound is in fact a single token formed through concatenation, Stymne (2009) splits the compound into its single word components prior to translation. Then, after translation, post-processing rules are applied to reorder or merge the components. A different approach was proposed by Kim and Nakov (2011), who generate monolingual paraphrases of noun compounds to augment the training corpus (e.g. *beef import ban* $\rightarrow$ *ban on beef import*).

Probably Monti et al. (2011) present the most similar work to ours. They compile a parallel corpus of sentences containing several types of expressions, including PVs, and compare the outputs of rule-based and SMT systems. While their discussion provides insightful examples, it does not help quantify the extent to which multiword expressions pose problems to MT systems. Moreover, it is not possible to know the exact details of the MT paradigms used in their experiments.

Most of the published results to date focus on automatic evaluation measures and only deal with fixed constructions like noun compounds. The present paper presents two original contributions with respect to related work. First, we focus on a more flexible type of construction, phrasal verbs, which are not correctly dealt with by simple integration strategies (Carpuat and Diab, 2010; Stymne, 2009). Secondly, we base our findings on qualitative and quantitative results obtained from a large-scale human evaluation experiment. Moreover, we do not intend to improve a SMT system with multiword unit processing: our goal is rather to evaluate and quantify how hard it is to translate these constructions. We believe that this can help conceiving more linguistically informed models for treating multiword units in MT systems in the future, as opposed to heuristic trial-and-error strategies that can be found in the literature.

3 Phrasal verbs

Phrasal verbs are recurrent constructions in English. They are composed by a main verb (*take*) combined with a preposition (*take on* in *take on a challenge*) or adverb (*take away* in *I take away your books*). Even if "it is often said that phrasal verbs tend to be rather 'colloquial' or 'informal' and more appropriate to spoken English than written" (Sinclair, 1989, p. iv), PVs are pervasive and appear often in all language registers. PVs present a wide range of variability both in terms of syntax and semantics. Thus, they are challenging not only for NLP, but also for students learning English as a second language (Sinclair, 1989).

Syntactic characterisation Phrasal verbs can be intransitive, that is, taking no object (*the aircraft takes off*, *she will show up later*) or transitive (*he*

took off his shoes, *we made up this story*). Many PVs can appear in both intransitive and transitive configurations, having either related senses (*the band broke up*, *the government broke up monopolies*) or unrelated senses (*the aircraft takes off*, *he took off his shoes*). In this work, we will focus only on transitive PV occurrences.

In terms of syntactic behaviour of transitive PVs, one must distinguish two types of constructions: verb-particle constructions like *put off*, *give up* and *move on*, and prepositional verbs like *talk about*, *rely on* and *wait for*. In verb-particle constructions, the particle depends syntactically (and semantically) on the verb, while in prepositional verbs it depends on the object, constituting a PP-complement of a regular verb.

Moreover, as particles in English tend to be homographs with prepositions and adverbs (*up, out, in, off*), a verb followed by a particle may be syntactically ambiguous (*eat up [ten apples]*, *eat [up in her room]*, *eat [up to ten apples]*). This affects how they are to be identified, interpreted, and translated automatically, as explained in Section 4.

Semantic characterisation PVs can be described according to a three-way classification as (a) literal or compositional like *take away*, (b) aspectual or semi-idiomatic like *fix up*, and (c) idiomatic combinations like *pull off* (Bolinger, 1971). The first two classes capture the core meaning of particles as adding a sense of motion-through-location (*carry NP up*) and of completion or result (*fix NP up*) to the verb. Semi-productive patterns can be found in these combinations (e.g. verbs of cleaning + *up*). For idiomatic cases, however, it is not possible to straightforwardly determine their meanings by interpreting their components literally (e.g. *make out* → *kiss*).

Like simple verbs, PVs are often polysemous and their interpretation is not straightforward. Metaphor can change the sense and the interpretation (literal or idiomatic) of the PV, like in *wrap up the present* vs *wrap up the presentation*. While some PVs have limited polysemy (e.g. *figure out* and *look up* have only 1 sense in Wordnet), others can have multiple uses and senses (e.g. *pick up* has 16 senses and *break up* has 19 senses in Wordnet).

Many PVs seem to follow a productive pattern of combination of semantically related verbs and a given particle (Fraser, 1976), like verbs used to join material (*bolt, cement, nail* + *down*). While some verbs form combinations with almost every particle (*get, fall, go*), others are selectively combined with only a few particles (*book, sober* + *up*), or do not combine well with them at all (*know, want, resemble*). This productivity is specially high in spoken registers, as we verified in our experimental corpus (see Section 4).

	# sentences	
	Sys. 1	**Sys. 2**
Shared training set	137,319	137,319
PVs training set	1,034	1,037
Shared dev. set	2,000	2,000
PVs test set	1,037	1,034
Total	141,390	141,390

Table 1: Training, development and test set dimensions for MT systems 1 and 2.

4 Experimental setup

Our goal is to quantify the translation quality of PVs by current SMT paradigms. Therefore, we build phrase-based and hierarchical SMT systems from the same parallel English-French corpus. We also identify the sentences containing PVs on the English side, and then use them as test set for manual error analysis.

4.1 Parallel corpus and preprocessing

For all the experiments carried out in this work — extraction and translation of PVs — the English-French portion of the *TED Talks* corpus was used (Cettolo et al., 2012).[1] It contains transcriptions of the TED conferences, covering a great variety of topics. The colloquial and informal nature of the talks favours the productive use of PVs. Talks are given in English, and are translated by volunteers worldwide. The corpus contains 141,390 English-French aligned sentences with around 2.5 million tokens in each language.

Before feeding the corpus into the MT training pipeline, we performed tokenisation. Tokenisation was performed differently on both languages. Since we wanted to identify PVs in English automatically, we had to parse the English corpus. Therefore, we used the RASP system v2 (Briscoe et al., 2006) to generate the full syntactic analysis of the English sentences. Since the parser contains an embedded tokeniser, we ensured consistency by

[1] Available at the Web Inventory of Transcribed and Translated Talks: `https://wit3.fbk.eu/`

using this tokenisation as preprocessing for MT as well. On the French side, we applied the simplified tokeniser provided as part of the Moses suite.

After preprocessing, we performed automatic PV detection on the corpus, as described in Section 4.3. This resulted in a set of 2,071 sentences in the corpus which contain split PVs (henceforth *PV set*). We used around half of the PV set as test data, while the other half was kept as training data, included in the larger set of training sentences with no split PVs. However, since we wanted to maximise the amount of translated data to analyse, we built two similar MT systems (1 and 2) for each paradigm.[2] System 1 uses the first half of the PV set as training data and the second half as test, while for system 2 the sets are swapped. Table 1 summarises the data sets. Since the systems are comparable, we can concatenate the two test sets after translation to obtain 2,071 French sentences.[3] This ensures that training and test sets are disjoint and that the systems have seen enough occurrences to be able to learn the constructions. In the remainder of this paper, we make no distinction between systems 1 and 2.

4.2 MT systems

We compare SMT systems of two paradigms: a *phrase-based system* (PBS) and a *hierarchical system* (HS). The main difference between these two paradigms is the representation of correspondences in the translation model. While the PBS uses word sequences, the HS uses synchronous context-free grammars, allowing the use of non-terminal symbols in the phrase table. Intuitively, the HS should be more suitable to translate PVs because it can generalize the intervening words between the verb and the particle. In other words, while the PBS enumerates all possible intervening sequences explicitly (*make up*, *make it up*, *make the story up*, ...), the HS can replace them by a single variable (*make X up*).

Both PBS and HS were built using the Moses toolkit (Koehn et al., 2007) and standard training parameters.[4] The preprocessed training sets described in Table 1 were used as input for both systems. The corpus was word-aligned using GIZA++ and the phrase tables were extracted using the *grow-diag-final* heuristic. Language models were estimated from the French part of the parallel training corpus using 5-grams with IRSTLM. For the HS, the maximum phrase length was set to 5. The model weights were tuned with MERT, which converged in at most 16 iterations. The training scripts and decoder were configured to print out word alignment information, required to identify which part of a French translated sentence corresponds to a PV in English (see Section 5).

4.3 Phrasal verb detection

PVs were detected in three steps: automatic extraction, filtering heuristics and manual validation.

Automatic extraction As described in Section 4.1, we parsed the English corpus using RASP. It performs full syntactic analysis and generates a set of grammatical relations (similar to dependency syntax). The parser has a module for automatic PV detection. However, we are only interested in split PVs. Therefore, we used the `mwetoolkit` (Ramisch et al., 2010) to extract only sentences that follow the pattern `Verb + Object + Particle`, where:

- `Verb` is a content verb (POS starts with `VV`);
- `Object` is a sequence of at least 1 and at most 5 words, excluding verbs;
- `Particle` is a preposition or adverb tagged as `II`, `RR` or `RP` which depends syntactically on the verb with a `ncmod_part` relation.

Filtering heuristics The application of this pattern on the parsed corpus generates the PV set (2,071 sentences). Manual inspection allowed us to formulate further heuristics to filter the set. We removed 243 sentences that match one of the following rules around the identified PV:

- Verbs *go, walk, do, see* + locative words;[5, 6]
- Particles *about, well, at*;
- Locative words followed by the words *here* and *there*, or preceded by the word *way*;
- Expressions *upside down*, *inside out*, *all over*;
- Verbs with double particles.[7]

[2] In total, 4 MT systems were built.

[3] These were further cleaned, as described in Section 4.3.

[4] Described in more detail on the Moses online documentation, at `http://www.statmt.org/moses/?n=Moses.Baseline`.

[5] Prepositions or adverbs that indicate locations and/or directions: *up, down, in, out*

[6] Even though the rule removes some authentic PVs (*walk somebody out*), most of the time it matches regular verb+PP constructions wrongly parsed as PVs (*walk up the steps*).

[7] Even though these constructions are authentic PVs, the parser attaches the second particle to the verb instead of the first one (*walk out on somebody* as *walk on* instead of *walk out* + *PP*).

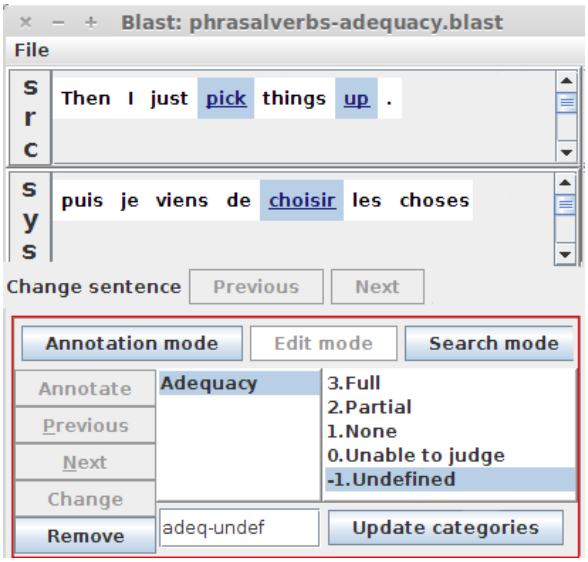

Figure 1: Annotation interface using BLAST.

Manual validation The extraction pattern and the filtering heuristics generate a precise set of sentences in a fully automatic manner. However, we require that the test set to be presented to our annotators contains 100% correctly identified PVs. Therefore, we manually inspected the resulting set of sentences and manually removed 266 of them. These were mainly due to parsing errors. The resulting set of sentences containing PVs has 1,562 sentences (705 different PVs). PV frequencies in this set vary from 1 to 44, and 637 PVs occur only once. Almost a half of all identified PVs, 452, were present in both training and test sets.

5 Evaluation protocol

The MT systems were used to translate the test set of English sentences containing PVs. For each English sentence, two corresponding translations in French were generated by the PBS and HS. We developed an evaluation protocol that allows human annotators to assess the quality of PV translation in the sentences in terms of adequacy and fluency.

5.1 Guidelines and annotation environment

An annotator was presented with a pair of sentences, the English source and the French target translated by one of the MT systems. If a sentence contains more than one PV, it is repeated in the annotation set, once for each PV. Even though a reference translation was available, we did not present it to avoid biases in the evaluation. Since we wanted to measure overall system performance, we did not perform comparative translation ranking (as in WMT, for instance), but this is intended as future work.

In order to avoid duplicated annotation effort, we only present once those sentences for which the PBS and the HS generate similar PV translations. This means that, for a given English sentence, its translations are considered as similar when the PV in it is aligned to the same number of French words and the concatenations of these words are identical in both translations. These translations are only presented once. On the other hand, since we also want to compare the systems, we select a set of highly dissimilar translations by picking up those whose longest common substring is shorter than half of the shortest translation. The dataset provided to annotators contains 250 similar sentences and 250 dissimilar sentence pairs, and the latter correspond to 500 translations (for dissimilar translations, each sentence pair is presented once, for the PBS and for the HS). In total, each annotator assessed 750 translations selected randomly from the test set of 1,562 sentences described in Section 4.3.

We ask annotators to focus only on the phrasal verb and its translation, ignoring the rest of the sentence. We use an adapted version of the BLAST system to provide a visual annotation interface (Stymne, 2011). The PV is highlighted, as well as its French counterpart, as shown on Figure 1. The French counterpart is identified thanks to the word alignment information output by the MT systems. There are two dimensions on which a translation is evaluated: adequacy and fluency.

Adequacy The annotator assessed a translated PV based on the extent to which the meaning of the original English PV is preserved in the French translation. The grade is based on how easy and precisely one can infer the intended meaning conveyed by the translation. The scale uses grades from 3 to 0, with 3 being the highest one.

- 3 - FULL: the highlighted words convey the same meaning as their English counterparts.
- 2 - PARTIAL: the meaning can be inferred without referring to the English sentence. The highlighted words sound clumsy, unnatural and/or funny, less relevant words might be missing or spurious words were added.
- 1 - NONE: the meaning is not conveyed in the translated sentence. In other words, the meaning of the French highlighted words cannot be

understood without reading and understanding the English sentence.

- 0 - UNABLE TO JUDGE: There is a problem with the source English sentence, which prevents the annotator from understanding it.[8]

Fluency The annotator assessed a translated PV based on its grammatical correctness in French, regardless of its meaning. The grade is based on how well the highlighted French words are inflected, specially regarding verb agreement in tense, number and gender. In this evaluation, the English sentence must be ignored. The scale uses grades from 4 to 1, with 4 being the highest one.

- 4 - FLUENT: the highlighted words in French show neither spelling nor syntax errors.
- 3 - NON-NATIVE: the verb form and/or its agreement with subject/object are wrong.
- 2 - DISFLUENT: the highlighted words make the sentence syntactically incoherent.
- 1 - INCOMPREHENSIBLE: the PV was not translated.

Four annotators, all of them proficient in English and French, participated in our human evaluation experiment. They were provided with detailed guidelines.[9] Annotators have access to a list of Wornet synsets and are instructed to consult online resources in case of doubts. In order to avoid bias towards either system, annotators are not informed which one was used to translate which sentence, and sentences are ordered randomly. If the PBS and the HS generate dissimilar translations for a source PV, they are presented consecutively. Fluency and adequacy are annotated separately in two passes.

5.2 Inter-annotator agreement

In order to validate the evaluation protocol, we calculated inter-annotator agreement,[10] following the methodology proposed by Artstein and Poesio (2008). In a first moment, a group of five volunteers annotated a pilot dataset of 156 sentences.

[8]Problematic source sentences were removed manually, but a small number of such cases accidentally remained in the test data.

[9]The guidelines, labels and datasets discussed here are available at `http://cameleon.imag.fr/xwiki/bin/view/Main/Phrasal_verbs_annotation`

[10]We report values of multi-π (Fleiss' κ), which estimates chance agreement from the overall category distribution.

	could **boil** *this poem* **down** *to saying*
PBS	*pourriez* **furoncle** *ce poème* **jusqu'** *à dire*
HS	*pourriez* **bouillir** *ce poème* **descendu** *à dire*
	he would **think** *it* **through** *and say*
Both	*il* **pense que** *ça* **à travers** *et dire*
	you couldn 't **figure** *it* **out**
HS	*vous ne pouvais pas le* **comprendre**
PBS	*vous ne pouviez pas le* **découvrir**
	Then we 'll **test** *some other ideas* **out**
Both	*puis nous allons* **tester** *certains autres idées*

Table 2: Examples of translated sentences.

Sentences annotated by at least one judge as UNABLE TO JUDGE were removed from adequacy data.

For fluency, the overall agreement is $\kappa = 0.50$. It seems easier to distinguish FLUENT translations from other classes (60% of agreeing pairs), than making distinctions between NON-NATIVE, DISFLUENT and INCOMPREHENSIBLE translations (42 to 45% of agreeing pairs). As for pairwise agreement, values range from $\kappa = .33$ to $\kappa = .72$, with one annotator being an outlier ($\kappa \leq .38$). If this annotator is removed, overall agreement is $\kappa = .61$, with the hardest class to distinguish being the intermediary NON-NATIVE (49% of agreeing pairs). This indicates a high level of coherence among annotators, given the complexity of the task.

For adequacy, annotation is harder and $\kappa = .35$, with pairwise agreement ranging from $\kappa = .23$ to $\kappa = .52$. While it seems intuitive to assess a translation as NONE (54% of agreeing pairs), the distinction between FULL and PARTIAL is more subjective (31% to 34% agreeing pairs). If these classes are merged, agreement raises to $\kappa = 0.47$. Even though these values are low, they are acceptable for our analysis. For future work, we intend to improve our guidelines and provide additional training to annotators.

6 Results

We analyse the results of manual annotation by four human judges on a set of 750 sentences, corresponding to 500 source sentences. In half of them, PVs were translated similarly by the HS and by the PBS. In the other half, they were translated differently, and thus included twice.

6.1 How does MT perform?

Our first question concerns the overall quality of translation, regardless of the fact that it was generated by the HS or by the PBS. Table 2 presents examples of translations showing that translation quality is poor. For instance, the PV *boil down*, which means *reduce* or *come down* and should be translated as *résumer*, was translated literally as *bouillir descendu (boil went down)* by the HS and as *furoncle jusqu' (furuncle until)* by the PBS. The second example, *think through*, should be translated as *repenser* or *refléchir*, but was translated literally as *penser à travers (think through)*, which makes no sense.

An automatic sanity check, based on BLEU score, was performed for both systems on the PV set (2,071 sentences) according to the protocol presented in Table 1. PBS and HS systems obtained 29.5 and 25.1 BLEU points respectively (to be compared with 32.3 for Google Translate). This automatic evaluation shows that the PBS is better than the HS system. Even though both systems are outperformed by Google, we consider them as acceptable for our experiment, considering the limited amount of training data used (TED corpus only).

On Table 3, the first column shows the average score obtained by the PV translations. In a scale from 1 to 3, the translations obtain an average of 1.73 for adequacy and, in a scale from 1 to 4, an average of 2.57 for fluency. This means that roughly half of the translations present some meaning and/or grammar problem that reduces their utility. In proportion to the scale, adequacy problems are slightly more frequent than fluency problems. In order to have a better idea of how serious this problem is, we plot in Figure 2 the proportion of each adequacy category in the dataset. The graphic shows that only 27% of the PVs are translated as a French verb which fully conveys the meaning of the English PV. Around 20% of the PVs are translated as a verb that is partly related to the original meaning, and the remainder 57% of translations are useless. This is a clear evidence that these constructions are not correctly dealt with by our SMT systems.

6.2 Comparison of both MT paradigms

Let us now compare the average scores obtained by each MT paradigm. As shown in the second and third columns of Table 3, the PBS seems to outperform the HS for fluency and adequacy. However, the difference between both systems for adequacy is not statistically significant ($p = 0.5236$).[11]

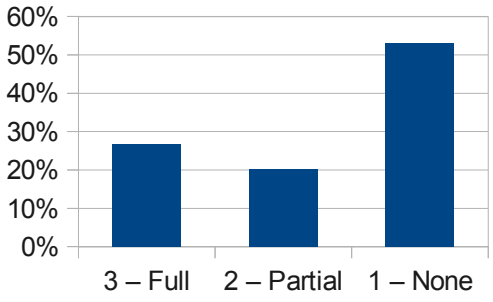

Figure 2: Proportion of translations judged as FULL, PARTIAL and NONE for adequacy.

However, in order to avoid the smoothing of category distribution generated by the presence of similar translations, we consider only those sentences for which different translations were generated. As explained in Section 5, we include 250 source sentences which have different translations by the PBS and by the HS. The average grade of each system on this set of different translations is shown in the last two columns of Table 3. In this case, the PBS performs significantly better than the HS in both fluency and adequacy.

An interesting finding of our analysis is shown in columns 4 and 5 of Table 3. We compared the average grades of sentences that were translated similarly by both systems with those translated differently. We found out that similar translations are of better quality (average grades 2.82 fluency and 2.07 adequacy) than different translations (average grades 2.44 fluency and 1.56 adequacy), and this difference is statistically significant ($p < 0.0001$). This result is a potentially useful feature in models to automatically estimate translation quality.

It is counter-intuitive that the PBS outperforms the HS in translating split PVs. These constructions have a flexible nature, and a PBS systems generally enumerate all possibilities of intervening material whereas the HS can efficiently represent gapped phrases and generalise using non-terminal symbols. We provide three hypotheses for this surprising outcome. First, it is possible that the size of our training corpus is not sufficient for the HS to learn useful generalisations (notably, the language model was trained on the French part of the parallel corpus only). Second, possibly the standard

[11] Statistical significance was calculated using a two-tailed t test for the difference in means.

	Overall	PBS	HS	Similar	Different	PBS-Diff.	HS-Diff.
Fluency (1-4)	2.57	2.67	2.46	2.82	2.44	2.63	2.25
Adequacy (1-3)	1.73	1.75	1.72	2.07	1.56	1.65	1.48

Table 3: Average grades obtained by the systems.

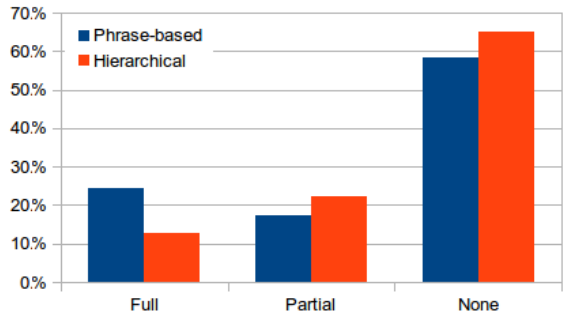

Figure 3: Proportion of different translations judged as FULL, PARTIAL and NONE for adequacy.

parameters of the HS should be tuned to our language pair and corpus. Third, most of the time the intervening word is the pronoun *it*, and this can be efficiently represented as two bi-phrases in the PBS, one for the joint form (*make up*) and another for the split form (*make it up*). Further investigation and a careful inspection of the phrase tables is needed in order to validate these hypotheses.

In both PBS and HS, the frequency of PVs in the training data is one possible factor that influences translation quality. In order to validate this hypothesis, we calculated the correlation (Kendall's τ) between the frequency of verb types and their average translation quality. The correlations range from $\tau = 0.17$ to $\tau = 0.26$, showing that, even though frequency is correlated with translation quality, it is not the only factor that explains our results. The influence on translation quality of other factors — such as polysemy, frequency of joint occurrences and verb-particle distance — will be investigated as future work.

7 Conclusions and future work

We presented a systematic and thorough evaluation of split PV translation from English into French. Therefore, we first identified sentences containing these constructions using a reusable pipeline (based on RASP + `mwetoolkit`) and applied heuristics and manual validation. Two SMT systems, a PBS and a HS were built on the TED corpus (spoken English) using standard parameters with Moses. These were used to generate translations which were then evaluated by human annotators following detailed guidelines.

Our main contribution is to show that, even though SMT is nowadays a mature framework, flexible constructions like PVs cannot be modelled appropriately. As a consequence, more than half of the translations have adequacy and/or fluency problems. The use of hierarchical systems does not seem to overcome these limitations, and generalisation over limited parallel data seems to be a bottleneck.

As future work, we would like to improve the general quality of our SMT systems. We noticed that, sometimes, the bad quality of other parts of the sentence prevented annotators from concentrating on the PV. Therefore, we would like to reproduce these experiments using a much larger parallel corpus as training data and much larger monolingual corpora for training language models.

We underline that the correct translation of PVs depends on their correct identification. There is still room for improvement in PV identification methods, as can be seen from the manual cleaning steps in the creation of our datasets. Even though automatic identification was out of the scope of this work, as future work we would like to study its impact on translation quality.

Finally, we would like to investigate other types of multiword units. On the one hand, joint PV instances and PVs with double particles (*look forward to*) are equally challenging for MT, and we would like to include them in future evaluations. On the other hand, there are many other complex expressions, like idioms and support-verb constructions, which are not correctly dealt with by current MT systems. We hope that this research can help designing better MT systems, capable of taking multiword expressions into account in an elegant manner.

Acknowledgements

We would like to thank Emmanuelle Esperança-Rodier for the help in writing the annotation guidelines, and all the volunteers who annotated the data sets. This research was partly

funded by the CAMELEON project (CAPES-COFECUB 707-11).

References

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comp. Ling.*, 34(4):555–596.

Baldwin, Timothy and Su Nam Kim. 2010. Multiword expressions. In Indurkhya, Nitin and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.

Bolinger, Dwight. 1971. *The phrasal verb in English*. Harvard UP, Harvard, USA. 187 p.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. 2012. Identifying bilingual multi-word expressions for statistical machine translation. In *Proc. of the Eigth LREC (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Briscoe, Ted, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In Curran, James, editor, *Proc. of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sidney, Australia, Jul. ACL.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comp. Ling.*, 19(2):263–311.

Carpuat, Marine and Mona Diab. 2010. Task-based evaluation of multiword expressions: a pilot study in statistical machine translation. In *Proc. of HLT: The 2010 Annual Conf. of the NAACL (NAACL 2003)*, pages 242–245, Los Angeles, California, Jun. ACL.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy, May.

Chiang, David. 2007. Hierarchical phrase-based translation. *Comp. Ling.*, 33(2):201–228.

Fraser, Bruce. 1976. *The Verb-Particle Combination in English*. Academic Press, New York, USA.

Kim, Su Nam and Preslav Nakov. 2011. Large-scale noun compound interpretation using bootstrapping and the web as a corpus. In Barzilay, Regina and Mark Johnson, editors, *Proc. of the 2011 EMNLP (EMNLP 2011)*, pages 648–658, Edinburgh, Scotland, UK, Jul. ACL.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 2003 Conf. of the NAACL on HLT (NAACL 2003)*, pages 48–54, Edmonton, Canada. ACL.

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 *Proc. of the 45th ACL (ACL 2007)*, pages 177–180, Prague, Czech Republic, Jul. ACL.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA. 620 p.

Monti, Johanna, Anabela Barreiro, Annibale Elia, Federica Marano, and Antonella Napoli. 2011. Taking on new challenges in multi-word unit processing for machine translation. In *Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation*, Barcelona, Spain, Jan.

Ramisch, Carlos, Aline Villavicencio, and Christian Boitet. 2010. Multiword expressions in the wild? the mwetoolkit comes in handy. In Liu, Yang and Ting Liu, editors, *Proc. of the 23rd COLING (COLING 2010) — Demonstrations*, pages 57–60, Beijing, China, Aug. The Coling 2010 Organizing Committee.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In Anastasiou, Dimitra, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim, editors, *Proc. of the ACL Workshop on MWEs: Identification, Interpretation, Disambiguation, Applications (MWE 2009)*, pages 47–54, Suntec, Singapore, Aug. ACL.

Sinclair, John, editor. 1989. *Collins COBUILD Dictionary of Phrasal Verbs*. Collins COBUILD, London, UK. 512 p.

Stymne, Sara. 2009. A comparison of merging strategies for translation of German compounds. In *Proc. of the Student Research Workshop at EACL 2009*, pages 61–69, Apr.

Stymne, Sara. 2011. Blast: A tool for error analysis of machine translation output. In *Proc. of the ACL 2011 System Demonstrations*, pages 56–61, Portland, OR, USA, Jun. ACL.