Machine Translation Summit XIV

2-6 September 2013, Nice, France

Workshop Proceedings:

MULTI-WORD UNITS IN MACHINE TRANSLATION AND TRANSLATION TECHNOLOGIES

Editors:

Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, Violeta Seretan

Workshop Proceedings for:
Multi-word Units in Machine Translation and Translation Technologies
(Organised at the 14th Machine Translation Summit)

Editors: Johanna Monti, Ruslan Mitkov, Gloria Corpas Pastor, Violeta Seretan

Published by:
The European Association for Machine Translation
Schützenweg 57
CH-4123 Allschwil / Switzerland

ISBN: 978-3-9524207-4-4

MT Summit 2013 Workshop Chair: Svetlana Sheremetyeva

Workshop organisers:

Johanna Monti (University of Sassari, Italy)
Ruslan Mitkov (University of Wolverhampton, United Kingdom)
Gloria Corpas Pastor (University of Málaga, Spain)
Violeta Seretan (University of Geneva, Switzerland)

Programme Committee members:

Iñaki Alegria (University of the Basque Country, Spain)
Giuseppe Attardi (University of Pisa, Italy)
Doug Arnold (University of Essex, United Kingdom)
Francis Bond (Nanyang Technological University, Singapore)
Bruno Cartoni (University of Geneva, Switzerland)
Jean-Pierre Colson (Université catholique de Louvain, Belgium)
Béatrice Daille (Nantes University, France)
Mona Diab (Columbia University, USA)
Gaël Dias (University of Caen Basse-Normandie, France)
Dmitrij O. Dobrovol'skij (Russian Academy of Sciences, Russia)
Annibale Elia (University of Salerno, Italy)
Thierry Fontenelle (Translation Centre for the Bodies of the European Union, Luxembourg)
Roxana Girju (University of Illinois at Urbana-Champaign, USA)
Barry Haddow (University of Edinburgh, United Kingdom)
Ulrich Heid (Universität Hildesheim, Germany)
Kyo Kageura (University of Tokyo, Japan)
Valia Kordoni (Humboldt University Berlin, Germany)
Koenraad Kuiper (University of Canterbury, New Zealand)
Guy Lapalme (University of Montreal, Canada)
Preslav Nakov (Qatar Computing Research Institute, Qatar Foundation, Qatar)
Pavel Pecina (Charles University in Prague, Czech Republic)
Carlos Ramisch (University of Grenoble, France)
Johann Roturier (Symantec Ltd., Ireland)
Gilles Sérasset (University of Grenoble, France)
Max Silberztein (University of Franche-Comté, France)
Dan Tufiş (Romanian Academy, Romania)
Agnes Tutin (University of Grenoble, France)
Michael Zock (Aix-Marseille University, France)

Invited speakers:

Gloria Corpas Pastor (University of Málaga, Spain)
Violeta Seretan (University of Geneva, Switzerland)

Workshop Programme
Tuesday, 3 September 2013

8.45 – 8.50 **Welcome address:** Ruslan Mitkov (University of Wolverhampton, United Kingdom)

8:50 – 9:30 **Invited talk**

All that Glitters is not Gold when Translating Phraseological Units
Gloria Corpas Pastor (University of Málaga, Spain)

9:30 – 10:30 Session 1

9:30 – 10:00 *Anaphora Resolution, Collocations and Translation*
Eric Wehrli and Luka Nerima

10:00 – 10:30 *A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora*
Oscar Mendoza Rivera, Ruslan Mitkov and Gloria Corpas Pastor

10:30 – 11:00 Coffee break

11:00 – 13:00 Session 2

11:00 – 11:30 *When Multiwords Go Bad in Machine Translation*
Anabela Barreiro, Johanna Monti, Fernando Batista and Brigitte Orliac

11:30 – 12:00 *Using a Rich Feature Set for the Identification of German MWEs*
Fabienne Cap, Marion Weller and Ulrich Heid

12:00 – 12:30 *MWU Processing in an Ontology-Based CLIR Model for Specific Domain Collections*
Maria Pia di Buono, Johanna Monti, Mario Monteleone and Federica Marano

---

12:30 – 14:30 Lunch break

14:30 – 15:10 Invited talk

On Translating Syntactically-Flexible Expressions
Violeta Seretan (University of Geneva, Switzerland)

15:10 – 16:10 Session 3

15:10 – 15:40 *How Hard is it to Automatically Translate Phrasal Verbs from English to French?*
Carlos Ramisch, Laurent Besacier and Oleksandr Kobzar

15:40 – 16:10 *Improving English-Bulgarian Statistical Machine Translation by Phrasal Verb Treatment*
Iliana Simova and Valia Kordoni

16:10 – 16:15 Closing

Table of Contents

Welcome by the MT Summit 2013 Workshop Chair

The biennial MT Summit conference has the unique mission of bringing together researchers in all areas of language processing as Machine Translation can both benefit from different computational techniques and approaches, as well as serve as a test bed for their viability.

It is my pleasure to welcome the attendees of the workshop on Multi-word Units in Machine Translation and Translation Technology in conjunction with the 14th MT Summit here in Nice. Developing strategies for detecting and using multi-word expressions in Machine Translation is of one of the most important areas of research that has a direct impact on improving translation accuracy.

I should like to take this opportunity to express my appreciation to Johanna Monti, Violeta Seretan, Gloria Corpas Pastor and Ruslan Mitkov, the organisers of the workshop, whose expertise and work provided the most comprehensive collection of papers to be presented and ensured fruitful exchange of professional knowledge and discussions.

I wish you all an excellent workshop and enjoyable stay in Nice, the queen of the French Riviera.

Svetlana Sheremetyeva

Foreword by the Workshop Organisers

It is a pleasure to welcome you to the MT Summit workshop on Multi-word Units in Machine Translation and Translation Technology. The workshop is dedicated to one of the open challenges of Machine Translation, a complex linguistic phenomenon, ranging from lexical units with a relatively high degree of internal variability to expressions that are frozen or semi-frozen. In spite of the recent positive developments in translation technologies, multi-word units still present unexpected obstacles to Machine Translation and translation technologies in general, because of intrinsic ambiguities, structural and lexical asymmetries between languages, and cultural differences. Multi-word unit (MWU) identification and translation problems are far from being solved and there is still considerable room for improvement.

The focus of this workshop is to address the MWU issue in a synergetic way, taking advantage of the recent developments in disciplines such as Linguistics, Translation Studies, Computational Linguistics, and Computational Phraseology.

These Proceedings bring together papers from researchers working on various aspects of MWU processing in Machine Translation and Translation Technologies. The papers are indicative of the current efforts of researchers and developers who are actively engaged in improving the state of the art of MWU translation. In particular, this workshop collects contributions concerning different types of MWUs, different aspects of MWU processing and, finally, evaluation of MWU translation. We anticipate that the presentations at this workshop will allow us to create an environment for interesting discussions and maybe even for the creation of new partnerships.

We hope you find the papers in this volume rewarding. We would like to thank all those who contributed papers to this workshop and the Programme Committee members for their valuable comments during the review process, which in turn helped the authors to improve their contributions. We would also like to thank the MT Summit 2013 Workshop Chair, Svetlana Sheremetyeva, for all her work and help with the organisation of the workshop.

Ruslan Mitkov, University of Wolverhampton
Johanna Monti, University of Sassari
Gloria Corpas Pastor, University of Málaga
Violeta Seretan, University of Geneva

ABSTRACT

Invited Talks 3 September 2013
Invited Talk 1

Date and Time: September 3rd
Speaker: Prof. Gloria Corpas Pastor
Organisation: University of Malaga (Spain)

Bio: Gloria Corpas Pastor, PhD is Professor in Translation and Interpreting at the University of Malaga (Spain) and Visiting Professor in Translation Technologies at the University of Wolverhampton (Great Britain). She is currently member of the advisory council of EUROPHRAS (European Society for Phraseology) and secretary of AIETI (Iberian Association for Translation and Interpreting). Prof. Corpas has taught translation and interpreting courses for both undergraduate and postgraduate programs since 1989. She has published a large number of papers and books on translation technologies, specialised translation of legal and scientific texts, corpus-based translation, lexicography, terminology and phraseology. She has also taken part and/or led many European and Spanish R&D projects. Prof. Corpas is co-author of a patent for a method to determine corpus representativeness (ref. no. P200695657). At present, she is Research Group Leader (Ref. No. HUM-106 "Translation and Lexicography", since 1997) and Project Leader for a number of R&D projects for the Spanish Ministry of Education and the Andalusian Ministry of Education. In addition, she plays an important role in University evaluation and curricula design within the Spanish Agency for Quality Assessment and Accreditation (ANECA). Prof. Corpas is part of national and international committees (AEN/CTN 174, CEN/BTTF 138, ISO TC37/SC2-WG6), whose goal is to implement a number of standardised regulations in the field of translation and interpreting.

Title: All that glitters is not gold when translating phraseological units
Abstract: *Phraseological unit* is an umbrella term which covers a wide range of multi-word units (collocations, idioms, proverbs, routine formulae, etc.). Phraseological units (PUs) are pervasive in all languages and exhibit a peculiar combinatorial nature. PUs are usually frequent, cognitively salient, syntactically frozen and/or semantically opaque. Besides, their creative manipulations in discourse can be anything but predictable, straightforward or easy to process. And when it comes to translating, problems multiply exponentially. It goes without saying that cultural differences and linguistic anisomorphisms go hand in hand with issues arising from varying degrees of equivalence at the levels of system and text. No wonder PUs have been considered *a pain in the neck* within the NLP community. This presentation will focus on contrastive and translational features of phraseological units. It will consist of three parts. As a convenient background, the first part will contrast two similar concepts: *multi-word unit* (the preferred term within the NLP community) versus *phraseological unit* (the preferred term in phraseology). The second part will deal with phraseological systems in general, their structure and functioning. Finally, the third part will adopt a contrastive

approach, with especial reference to translators' strategies, procedures and choices. For good or for bad, when it comes to rendering phraseological units, human translation and computer-assisted translation appear to share the same garden path.

ABSTRACT

Invited Talks: 3 September, 2013
Invited Talk 2

Date and Time: September 3rd
Speaker: Dr. Violeta Seretan
Organisation: University of Geneva (Switzerland)

Bio: Violeta Seretan is a Senior Researcher at the Faculty of Translation and Interpreting, University of Geneva. She received her PhD in Computational Linguistics from the University of Geneva in 2008. She has been a Lecturer at the Language Technology Laboratory in the Department of Linguistics of the University of Geneva (2008-2010), then a visiting researcher at Institute for Language, Cognition and Computation at the University of Edinburgh (2010-2011). Her research interests are in language analysis, computational lexicography, machine translation and language generation. She has authored a book and over 30 papers in international journals and conference proceedings in these areas.

Title: On Translating Syntactically-Flexible Expressions

Abstract: The performance of translation systems largely depends on their ability to identify the units of meaning in text. These units are not limited to single words, but, to a large extend, they are represented by multi-word expressions. Because of their non-compositionality, such expressions cannot be accounted for in a word-by-word basis, but have to be processed as a whole. A major challenge in processing them is, however, their syntactic flexibility: While theoretical studies describe multi-word expressions as relatively fixed, with the syntactic fixedness going hand in hand with the semantic opacity, evidence from corpus-based studies showed that there is a surprising range of variation, leading to the discontiguity of the composing items. This presentation will first look at the extent to which existing translation paradigms are able to cope with this discontiguity. Then, it will outline the findings of an empirical study showing that syntactic flexibility affects the translation performance, but to a degree which is dependent on the type of the systems (rule-based vs. statistical). The recent advances in hybrid machine translation, which integrate grammatical modelling in statistical machine translation, may provide a suitable solution to the flexibility challenge in translating multi-word expressions.

Anaphora Resolution, Collocations and Translation

Eric Wehrli
LATL-CUI
University of Geneva
Eric.Wehrli@unige.ch

Luka Nerima
LATL-CUI
University of Geneva
Luka.Nerima@unige.ch

Abstract

Collocation identification and anaphora resolution are widely recognized as major issues for natural language processing, and particularly for machine translation. This paper focuses on their intersection domain, that is verb-object collocations in which the object has been pronominalized. To handle such cases, an anaphora resolution procedure must link the direct object pronoun to its antecedent. The identification of a collocation can then be made on the basis of the verb and its object or its antecedent. Preliminary results obtained from the translation of a large corpus will be discussed.

1 Introduction

Collocation identification and anaphora resolution (henceforth AR) are widely recognized as major issues for natural language processing, and particularly for machine translation. An abundant literature has been dedicated to each of those issues (see in particular Mitkov (2002) for AR, Wehrli *et al.* (2010) and Seretan (2011) for collocation identification), but to the best of our knowledge their intersection domain – a collocation in which the base term has been pronominalized – has hardly been treated yet. This paper intends to be a modest contribution towards filling this gap, focusing on the translation from English to French of collocations of the type verb-direct object, with and without pronominalization of the complement. The paper is organized as follows. The next section will give a brief overview of the translation problems with respect to both collocations and anaphors. We will also show how current MT systems fail to handle successfully such cases. In section 3 our treatment of collocations and anaphora resolution will be presented, along with some preliminary results. Finally, in section 4, we will try to address the issue of the frequency of those phenomena, presenting the results of our collocation extraction system over a corpus of approximately 10'000 articles from the news magazine *The Economist* totalizing over 8'000'000 words.

2 Collocations in Translation

The importance of collocations in translation has long been recognized, both by human translators and by developers of MT systems. For one thing, collocations tend to be ubiquitous in natural languages. Furthermore, it is often the case that they cannot be translated literally, as illustrated below. One of the characteristic features of collocations is that the choice of the collocate may be quite arbitrary and therefore cannot be safely derived from the meaning of the expression, and for that matter be translated literally. Consider, for instance, the examples in (1)-(2):

(1)a. heavy smoker

b. French
*lourd fumeur
gros/grand fumeur "big/large smoker"

c. German
*schwerer Raucher
starker Raucher "strong smoker"

(2)a. John broke a record.

b. French
John a battu un record
"John has beaten a record"

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.

The adjective *heavy* in the collocation (1) *heavy smoker* cannot be translated literally into French or into German. Both of those languages have their own equivalent collocation, which in turn could not be translated literally into English. Similarly, the verbal collocate in a verb-object collocation can usually not be translated literally, as illustrated in (2). In most cases, a literal translation, though sometimes understandable, would be felt as "non idiomatic" or "awkward" by native speakers. Even though this state of affair does not apply to all collocations, it is widespread across languages and requires a proper treatment of collocations. Commercial MT systems usually have a good handling of collocations of the type "noun-with-spaces", such as adjective-noun, noun-noun, noun-preposition-noun, and the like. With respect to collocations which display a certain amount of syntactic flexibility and in which the two constituents can be arbitrarily far away from each other, commercial MT systems do relatively poorly, as illustrated in the few examples given at the end of the next section.

2.1 Translating collocations with Its-2

In this section, we describe how collocations are handled in the Its-2 translation system (cf. Wehrli et al. 2009a, 2009b), which is based on the Fips multilingual parser (cf. Wehrli, 2007). The proposed treatment relies on the assumption that collocations are "pervasive" in NL (cf. Jackendoff, 1997; Mel'cuk, 2003), which calls for a "light" and efficient treatment – perhaps in contrast to true idiomatic expressions, which are far less numerous and may require and justify a much heavier treatment[1].

Let us first consider again example (2), which involves a verb-object collocation, both in the source language (*break-record*) and in the target language (*battre-record* "beat record")

The structure assigned to this sentence by the Fips parser is identical to the structure of a non-collocational sentence such as

(3) Jean a mangé un biscuit
"Jean has eaten a cookie"

Ideally, therefore, we would like to say that the only difference between the two examples boils down to a lexical difference: the verb and the object head noun correspond to a collocation in (2), but not in (3). Based on this observation, we will strive to develop a transfer and generation process which will be identical for the two cases, except for the lexical transfer.

The general transfer algorithm of Its-2 recursively traverses the syntactic tree structure generated by the parser in the following order: head, left sub-constituents, right sub-constituents. Lexical transfer occurs during the transfer of a non-empty head. At that time, the bilingual dictionary is consulted and the target language item with the highest score among all the possible translations of the source language lexical item is selected. If a collocation is identified in the source sentence, as in our example, the lexical item associated with the verb *break* will also specify that collocation. In such a case, lexical transfer occurs on the basis of the collocation and not on the basis of the lexeme.

This procedure yields encouraging results, as illustrated by the following simple example of translation, which we compare with outputs from some commercial MT systems, both statistical and rule-based[2]. A few more examples, with sentences taken from the magazine *The Economist*, are given in the last section:

(4)a. The record that Paul set is likely to be broken.

b. Its-2
Le record que Paul a établi est susceptible d'être battu.

c. Google translate
L'enregistrement qui Paul ensemble est susceptible d'être rompu.

d. Systran
Le disque que l'ensemble de Paul est susceptible d'être cassé.

e. Reverso
Le rapport(record) que Paul met va probablement être cassé.

Example (4) contains two collocations, *to set a record* and *to break a record*. The first one occurs in a relative clause, while the latter is in the

[1] See Sag et al. 2002 for a thorough and enlightening discussion of multiword expressions.

[2] The commercial MT systems are Google-Translate (translate.google.fr), Systran (www.systranet.com) and Reverso (www.reverso.net), accessed between August 22 and August 29, 2012.

passive voice. As a result, in neither of them the direct object follows the verb. For that reason, the three commercial MT systems that we considered fail to identify the presence of those collocations and, thus, yield a poor translation. Its-2, thanks to the Fips parser, is quite capable of identifying verb-object collocations even when complex grammatical processes disturb the canonical order of constituents and correctly translate them by means of the equivalent French collocations *établir un record* and *battre un record*.

(5)a. The world record will be broken.

b. Its-2
Le record du monde sera battu.

c. Google translate
Le record du monde sera brisé.

d. Systran
Le record mondial sera cassé.

e. Reverso
Le record du monde sera cassé.

Example (5) also exhibits two collocations, *world record* and *to break a record*. The first one is of the "noun-with-spaces" variety, and therefore is well-translated by all the systems. The second one is in the passive form and, as in the previous example, commercial systems fail to recognize it.

2.2 Anaphora resolution

As a first step towards a proper treatment of anaphora, we have developed a simple procedure that allows the Fips parser to handle personal pronouns, by far the most widespread type of anaphora[3], restricted to 3rd person[4]. A second limitation of our AR procedure is that it only covers cases of anaphoric pronouns with antecedent within the same sentence or within the preceding sentence. As reported by Laurent (2001) on the basis of a French corpus, these two cases cover nearly 89% of the cases (67% and 22%, respectively). Roughly speaking, our AR procedure adopts the Lappin and Leass (1994) algorithm, adapted to the grammatical representations and other specificities of the Fips parser.

First, the AR procedure must distinguish between anaphoric and non-anaphoric occurrences. For English, this concerns mainly the singular pronoun *it*, which can have an impersonal reading, as in (6). Identifying impersonal pronouns is achieved by taking advantage of the rich lexical information available to our parser.

(6)a. It is raining.

b. It turned out that Bill was lying.

c. To put it lightly.

d. It is said that they have been cheated.

The next step concerns anaphors in the stricter sense of Chomsky's binding theory (cf. Chomsky, 1981), that is reflexive and reciprocal pronouns, which must be bound in their governing category. Our somewhat simplified interpretation of principle A of the binding theory states that a reflexive/reciprocal pronoun must be linked to (ie. agrees with and refers to) the subject of its minimal clause[5].

Finally, in the third step, we consider referential pronouns, such as personal pronouns (*he, him, it, she, her, them, etc.*), still using the insight of binding theory, which states according to principle B that pronouns must be free (ie. not bound) in their governing category. Here again, our simplified interpretation of principle B prevents a pronoun from referring to any noun phrase in the same minimal clause.

Note that the binding theory is not an AR method per se, in the sense that it does not say what the antecedent of a pronoun is. What it does, though, is to filter out possible, but irrelevant candidates. To illustrate, consider the simple sentences in (7), where the indices represent the coindexing relation between a pronominal element and its antecedent.

(7)a. Peter$_i$ watches himself$_i$ in the mirror.

b. Peter$_i$ watches him$_k$ in the mirror.

c. *Peter$_i$ watches him$_i$ in the mirror.

[3]According to Tutin (2002), personal pronouns range from 60% to 80% of anaphoric expressions, based on a large, well-balanced French corpus. Russo et al. (2011) report relatively similar results for English, Italian, German and French.

[4]First and second person pronouns are left out, since they do not have any linguistic antecedent. Rather, their interpretation is usually set by the discourse situation.

[5]The minimal clause containing a constituent X is the first sentential node (tensed or untensed) which dominates X in the phrase structure.

Sentence (7a) is well-formed because the anaphor *himself* is bound by the subject *Peter*. Given principle A of binding theory, we can conclude that the only possible antecedent of *himself* is *Peter*. Following the same reasoning, binding theory validates (7b) and rules out (7c). Since *him* is a pronoun, it cannot be bound (ie. find its antecedent) within the same minimal clause. Therefore, it cannot refer to *Peter*.

Our implementation of a simple but efficient AR procedure makes use of a stack of noun phrases, restricted to argument noun phrases, that the parser stores for each analysis and maintains across sentence boundaries. When a pronoun is read, the parser first determines whether it is a reflexive/reciprocal pronoun, in which case by virtue of principle A it must co-refer to the subject of its minimal clause, or a 3rd person pronoun. In the latter case, the parser will distinguish between referential and non-referential *it*, as discussed above. As we mentioned, that distinction can be made on the basis of the lexical and grammatical information available to the parser, in connection with the grammatical environment of the pronoun. For referential 3rd person personal pronouns, the procedure selects all the noun phrases stored on the stack which agree in person, number and gender with the pronoun. If more than one is selected, preference goes first to the subject arguments and second non subject arguments, a heuristic inspired in part by the Centering theory (cf. Grosz et al., 1986, 1995; Kibble, 2001). Needless to say, the procedure sketched above is merely a first attempt at tackling the AR problem.

3 Results and final remarks

The examples discussed above are all simple sentences constructed for the purpose of the present research. Let us now turn to "real" sentences taken respectively, from the July 2, 2002 and from the February 7, 2004 issues of *The Economist*.

Consider the English collocation *to make a case*, as illustrated by the examples (8-9). A literal translation into French of this collocation would give something like *faire un cas*, which is hardly understandable and certainly fails to convey the meaning of that collocation. A more appropriate translation would use the collocation *présenter un argument*. In the first example, the collocation occurs in a *tough*-movement construction, a peculiar grammatical construction in which an adjective of the *tough*-classs (*tough, difficult, easy, hard, fun, etc.*) governs an infinitival complement whose direct object cannot be lexically realized, but is understood as the subject of the sentence – in our example the phrase *such a case*[6]. Following a standard generative linguistics analysis of that construction, we assume that the direct object position of the infinitival verb is occupied by an abstract anaphoric pronoun linked to the subject noun phrase.

We can observe that Google-translate chooses a literal translation of the collocation (8a), while Its-2 correctly identifies the presence of the collocation and translates it appropriately with the corresponding French collocation *présenter un argument*.

(8)a. Such a case would not be at all difficult to make.

b. Google-translate
Un tel cas ne serait pas du tout difficile à faire.

c. Its-2
Un tel argument ne serait pas du tout difficile à présenter.

In our second example (9), the collocation *make a case* occurs twice (*making this case*, *makes it*). Notice that in the second occurrence, the base term of the collocation has been pronominalized, with its antecedent in the previous sentence. Thanks to the AR procedure, Its-2 correctly identifies the collocation and translates it appropriately (9c), which is not the case for Google-translate (9b).

(9)a. Every Democrat is making this case. But Mr Edwards makes it much more stylishly than Mr Kerry.

b. Google-translate
Chaque démocrate rend ce cas. Mais M. Edwards, il est beaucoup plus élégant que M. Kerry.

c. Its-2
Chaque démocrate présente cet argument. Mais M. Edwards le présente beaucoup plus élégamment que M. Kerry.

[6] See Chomsky (1977) for a detailed analysis of this construction.

To measure the accuracy of our collocation identification procedure as well as the impact of the anaphora resolution algorithm, we parsed a corpus taken from *The Economist* totalizing over 8'000'000 words (463'173 sentences). 14'663 occurrences (tokens) of verb-object collocations were identified, corresponding to 553 types[7]. In 68 cases, the direct object had been pronominalized, as in the next two examples, where the source sentence(s) is given in the (a) section in which both the collocation (verb + pronoun) and the antecedent of the pronoun are emphasized. The (b) section gives the Its-2 translation with the anaphora procedure turned off, the (c) section the Its-2 translation with the AR procedure turned on, and the (d) section, the translation obtained with Google-translate.

(10)a. The golden **rule** also turns slithery under close inspection.
On an annual basis, the government is **breaking it**.

b. [-AR] Sur une base annuelle, le gouvernement **le casse**.

c. [+AR] Sur une base annuelle, le gouvernement **l'enfreint**.

d. [Google] Sur une base annuelle, le gouvernement est **le casser**.

The best result is (c), the only one where the collocation *break-rule* is correctly identified thanks to the AR procedure which connects the direct object pronoun to the subject of the preceding sentence *golden rule*. The translation of that collocation yields the French verb *enfreindre* rather than *casser*.

(11)a. In Spain the **target** is mainly symbolic, since companies will not face financial penalties if they do not **meet it**.

b. [-AR] En Espagne la cible est principalement symbolique, depuis que les sociétés n'affronteront pas des pénalités financières si ils ne **le rencontrent** pas.

c. [+AR] En Espagne la cible est principalement symbolique, depuis que les sociétés n'affronteront pas des pénalités financières si elles ne **l'atteignent** pas.

d. [Google] En Espagne, la cible est surtout symbolique, puisque les entreprises ne seront pas passibles de sanctions financières si elles ne **répondent** pas.

In that last example, the source sentence contains two pronouns, *they* referring to *companies* and *it* referring to *target*. In (c), both of them have been correctly handled by the AR procedure and with the latter the collocation *meet-target* has been identified, yielding the correct collocation translation *atteindre(-cible)*.

Although not very frequent, collocations with a direct object pronoun should not be overlooked if one aims at a high-quality translation, as illustrated by the examples (10-11). Extending the collocation lexicon and the AR procedure to a larger set of pronouns, as we intend to do in future work is likely to increase the number of pronominalized collocations detected by the system.

[7]The most frequent collocations are *to take place* (529 occurrences), *to make sense* (407), *to play a role* (323), *to make money* (304) and *to make a difference* (266). Among the collocations with pronominalized objects, the most frequent are *to spend money* (7) and *to solve a problem* (5).

Acknowledgements

Part of the research described in this paper has been supported by a grant from the Swiss National Science Foundation (grant No. 100012-113864/1).

4 References

Chomsky, N. 1977. "On Wh-Movement", in Peter Culicover, Thomas Wasow, and Adrian Akmajian, eds., *Formal Syntax*, New York, Academic Press, 71-132.

Chomsky, N. 1981. *Lectures on Government and Binding*, Foris Publications.

Grosz, B.. A. Joshi & S. Weinstein, 1995. "Centering: A Framework for Modeling the Local Coherence of Discourse", *Computation Linguistics*, 21:2, 203-225.

Grosz, B. & C. L. Sidner, 1986. "Attention, intention, and the structure of discourse", *Computational Linguistics*, 12:3, 175?204.

Jackendoff, R. 1997. *The Architecture of the Language Faculty*, Cambridge, Mass., MIT Press.

Kibble, R. 2001. "A Reformulation of Rule 2 of Centering Theory", in *Computational Linguistics*, 27:4, Cambridge, Mass., MIT Press.

Lappin, Sh. & H. Leass, 1994. "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics* 20:4, 535-561.

Laurent, D. 2001. *De la résolution des anaphores*, Rapport interne, Synapse Développement.

Mel'cuk, I. 2003. "Collocations : définition, rôle et utilité", in F. Grossmann and A. Tutin, eds., *Les collocations : analyse et traitement*, Amsterdam, De Werelt, pp. 23-32.

Mitkov, R. 2002. *Anaphora Resolution*, Longman.

Russo, L., Y. Scherrer, J.-Ph. Golman, S. Loaiciga, L. Nerima & E. Wehrli, 2011. "Etudes inter-langues de la distribution et des ambiguités syntaxiques des pronoms", Montpellier, TALN.2011.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger, 2002. "Multiword expressions: A pain in the neck for NLP", in *Proccedings of Computational Linguistics and Intelligent Text Processing (CICLing-2002)*, Lecture Notes in Computer Science, 2276, 1-15.

Seretan, V. 2011. *Syntax-Based Collocation Extraction*, Springer Verlag.

Tutin, A. 2002. "A Corpus-based Study of Pronominal Anaphoric Expressions in French", in *Proceedings of DAARC 2002*, Lisbonne, Portugal.

Wehrli, E. 2007. "Fips, a 'deep' linguistic multilingual parser" in *Proceedings of the ACL 2007 Workshop on Deep Linguistic processing*, pp. 120-127, Prague, Czech Republic.

Wehrli, E., Nerima, L., and Scherrer Y., 2009a. "Deep linguistic multilingual translation and bilingual dictionaries", *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 90-94, Athens, Greece.

Wehrli, E., Seretan, V., Nerima, L., and Russo, L., 2009b. "Collocations in a rule-based MT system: A case study evaluation of their translation adequacy", *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation*, pp. 128-135, Barcelona, Spain.

Wehrli, E., V. Seretan and L. Nerima (2010). "Sentence Analysis and Collocation Identification" in *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications* (MWE 2010), pp. 27-35, Beijing, China.

A Flexible Framework for Collocation Retrieval and Translation from Parallel and Comparable Corpora

Oscar Mendoza Rivera, Ruslan Mitkov and Gloria Corpas Pastor
Research Group in Computational Linguistics, University of Wolverhampton
{o.mendozarivera, r.mitkov}@wlv.ac.uk, gcorpas@uma.es

Abstract

This paper outlines a methodology and a system for collocation retrieval and translation from parallel and comparable corpora. The methodology was developed with translators and language learners in mind. It is based on a phraseology framework, applies statistical techniques, and employs source tools and online resources. The collocation retrieval and translation has proved successful for English and Spanish and can be easily adapted to other languages. The evaluation results are promising and future goals are proposed. Furthermore, conclusions are drawn on the nature of comparable corpora and how they can be better exploited to suit particular needs of target users.

1 Introduction

Multiword expressions (MWEs) are lexical units made up of several words in which at least one of them is restricted by linguistic conventions. One example is the expression *fast food*, in which the word *fast* is arbitrary, as it cannot be replaced with synonyms, such as *quick*, *speedy* or *rapid*. It is thought that a significant part of a language's vocabulary is made up of these expressions: as noted by Biber *et al.* (1999), MWEs account for between 30% and 45% of spoken English and 21% of academic prose, while Jackendoff (1997) goes as far as to claim that their estimated number in a lexicon is of the same order of magnitude as its number of single words. Furthermore, these numbers are probably underestimated: they appear in all text genres, but specialised domain vocabulary, such as terminology, "overwhelmingly consists of MWEs" (Sag *et al.*, 2002, p. 2).

Collocations represent the highest proportion of MWEs (Lea and Runcie, 2002; Seretan, 2011). As such, collocation retrieval has sparked interest in the NLP community (Smadja, 1993; Sag *et al.*, 2002; Lü and Zhou, 2004; Sharoff *et al.*, 2009; Gelbukh and Kolesnikova, 2013). Several methods have been adopted to measure the association strength of collocations, which has achieved favourable results with increases in accuracy (Seretan, 2011). However, a much more limited number of studies have dealt with post-processing of collocations from the perspective of their practical use. Collocation translation, for instance, while a natural follow-up to collocation extraction in this trail of research, still poses a problem for computational systems (Seretan, 2011). Furthermore, while several collocation resources have been put together, such as the multilingual collocation dictionary *MultiCoDiCT* (Cardey *et. al*, 2006), approaches to collocation retrieval and translation lack, in general, the solid theoretical basis of phraseology (Corpas Pastor, 2013).

To address this problem, the present paper describes the development and implementation of a computational tool to allow language learners and translators to retrieve collocations in a source language (SL) and their translations in a target language (TL) from bilingual parallel and comparable corpora. The project focuses on English and Spanish, but the methodology is designed to be flexible enough to be applied to other pairs of languages.

The remainder of this paper is organised as follows: Section 2 discusses the phraseology basis of our project and presents two collocation typologies (one in English and one in Spanish) as well as a comparative grammar. Section 3 provides a brief review of existing techniques for the extraction and translation of collocations. Section 4 presents a new methodology and outlines the implementation of a computational tool based on it. Finally, Section 5 details the results from the experiments set up to evaluate our system, and discusses opportunities for future work.

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.

2 Phraseology

Collocations are compositional and statistically idiomatic MWEs (Baldwin and Kim, 2010). Like idioms, collocations belong to the set phrases of a language. Unlike them, while the meaning of an idiom is mostly incomprehensible if not previously heard (*to **pay** through the nose*, *cold **turkey***), collocations are compositional: their meanings can be deduced from the meaning of their component words (*to **pay** attention*, *roast **turkey***). However, these are arbitrary. For example, the expression *I **did** my homework* is correct in English, but the expression *I **made** my homework* is not. The choice of using the verb *to do* and not the verb *to make* in this particular example can be thought of as an arbitrary convention. In addition, some collocates exhibit delexical and metaphorical meanings (*to **make** an attempt, to **toy** with an idea*). Similarly, collocations are cohesive lexical clusters. This means that the presence of one or several component words of a collocation in a phrase often suggests the existence of the remaining component words of that collocation. This property attributes particular statistical distributions to collocations (Smadja, 1993). For example, in a sample text containing the words *bid* and *farewell*, the probability of the two of them appearing together is higher than the probability of the two of them appearing individually.

2.1 Typologies of collocations

Hausmann (1985) argued the components of a collocation are hierarchically ordered: while the *base* can be interpreted outside of the context of a collocation and can therefore be considered as semantically autonomous, the *collocatives* depend on the base in order to get their full meaning. He also presented a typology of collocations in English based on their syntax (see Table 1). Similarly, Corpas Pastor (1995, 1996) studied, classified, and contrasted collocations for Spanish and English and has proposed her own typology of collocations in these two languages (see Tables 1 and 2). These tables show the base of collocations in bold and use abbreviations borrowed from the tagset of TreeTagger (Schmid, 1994): *VB* stands for verb, *NN* for noun, *RB* for adverb, *JJ* for adjective, and *IN* for preposition. Furthermore, these typologies have been helpful in the development of this project's underlying methodology to extract collocations (see Sections 4.1 and 4.2).

Type	Examples
1. VB + **NN** *(direct object)*	*to express concern, to bid farewell*
2. NN or JJ + **NN**	*traumatic experience, copycat crime*
3. NN + *of* + **NN**	*pinch of salt, pride of lions*
4. RB + **JJ**	*deadly serious, fast asleep*
5. **VB** + RB	*to speak vaguely, to sob bitterly*
6. VB + IN + **NN**	*to take into consideration, to jump to a conclusion*
7. VB + **NN** *(subject)*	*to break out <war>, to crow <a cock>*

Table 1: Typology of collocations in English

Type	Examples
1. VB + **NN** *(direct object)*	*conciliar el sueño, entablar conversación*
2. **NN** + JJ or NN	*lluvia torrencial, visita relámpago*
3. NN + *de* + **NN**	*grano de arroz, enjambre de abejas*
4. RB + **JJ**	*profundamente dormido, estrechamente relacionado*
5. **VB** + RB	*trabajar duro, jugar sucio*
6. VB + IN + **NN**	*tomar en consideración, poner a prueba*
7. VB + **NN** *(subject)*	*ladrar <un perro>, estallar <una guerra>*

Table 2: Typology of collocations in Spanish

2.2 Transfer rules

Bradford and Hill (2000) studied the comparison between the grammar of English and Spanish. Based on their work, we have developed a set of transfer rules (see Table 3) between these two languages which help us translate collocations (see Section 4.4).

English	Spanish
VB + **NN**	VB + **NN**
NN or JJ + **NN**	**NN** + JJ or NN
NN + *of* + **NN**	NN + *de* + **NN**
RB + **JJ**	RB + **JJ**
VB + RB	**VB** + RB
VB + IN + **NN**	VB + IN + **NN**

Table 3: English-Spanish syntax comparison

It is worth noting that these transfer rules are designed to aid us in our own approach to the task

of syntactic processing, but they are not all-inclusive. In fact, as is often the case, there are exceptions to the rules. For example, collocations in English such as *copycat crime* (*delito inspirado en uno precedente* or *que trata de imitarlo,* in Spanish) and *to commit suicide* (*suicidarse* in Spanish) cannot be translated using the proposed approach.

3 Related Work

This section presents a brief review of existing techniques for the extraction and translation of collocations. It starts by outlining collocation extraction and then moves to translation.

3.1 Collocation retrieval

Early work on collocation extraction focused on statistical processing. Choueka *et al.* (1983) developed an approach to retrieve sequences of words occurring together over a threshold in their corpora. Similarly, Church and Hanks (1989) proposed a correlation method based on the notion of mutual information. Smadja (1993), however, highlighted the importance of combining statistical and linguistic methods. In recent years, advances have been made (Ramisch *et al.*, 2010; Seretan, 2011), many of them advocating rule-based and hybrid approaches (Hoang, Kim and Kam, 2009), and based on language-specific syntactic structures (Santana *et al.*, 2011) or machine learning of lexical functions (Gelbukh and Kolesnikova, 2013).

3.2 Parallel corpora

Classic approaches to translation using parallel corpora exploited the concepts of alignment and correspondence at sentence level (Brown *et al.*, 1991; Gale and Church, 1993). Two methods were developed: length-based and translation-based (Varga *et al.*, 2005). Collocation translation using parallel corpora has also been approached using transfer systems that rely on generative grammars, because of the notion that the base of a collocation determines its collocatives (Wehrli *et al.*, 2009) and the assumption that source and target MWEs share their syntactic relation (Lü and Zhou, 2004).

3.3 Comparable Corpora

Parallel resources are generally scarce and in many cases not available at all. The wider availability of comparable texts offers new opportunities to both researchers and translators. While these do not allow for bridging between languages (Sharoff *et al.*, 2009), research suggests (Rapp, 1995) that a word is closely associated with words in its context and that the association between a base and its collocatives is preserved in any language. Fung and Yuen (1998), for instance, argued that the first clue to the similarity between a word and its translation is the number of common words in their contexts. Similarly, Sharoff *et al.* (2009) proposed a methodology that relies on similarity classes.

4 System

The system[1] employs the following three language-independent tools: TreeTagger to POS-tag corpora, the MWEToolkit (Ramisch *et al.*, 2010) to extract collocations according to specific POS-patterns, and Hunalign (Varga *et al.*, 2005) to align corpora at sentence level. Furthermore, it connects online to *WordReference* and uses it as a multilingual translation dictionary and thesaurus. Figure 1 illustrates the architecture of the system; its main modules will be described in greater detail in the following paragraphs.

4.1 Candidate selection module

This module processes the SL corpus in order to format it to comply with the input requirements of the modules that follow in the system pipeline. It represents the linguistic component of the hybrid approach to collocation retrieval. It makes use of both TreeTagger and the MWEToolkit to perform *linguistic pre-processing* in the form of lemmatisation and POS-tagging on the input data, as well as *POS-pattern definition*.

Linguistic processing aims at transforming the input data from a stream of alphanumeric characters to sequences of words, which can be grouped in *n-grams*. It is important to work with lemmas instead of inflected words in order to identify collocations; otherwise, for example, collocations such as *committing murder* and *committed murder* would be treated separately, even though they are obviously the same (whose lemma is *commit murder*). The system relies on TreeTagger to annotate text sentences with both lemma and POS-tagging information. Its output is then transformed into the XML format (see Figure 2) by running a Python script, part of MWEToolkit.

[1] Consisting of a series of Python scripts which handle text and XML representations, and implemented using the wxPython development environment for Mac OSX.

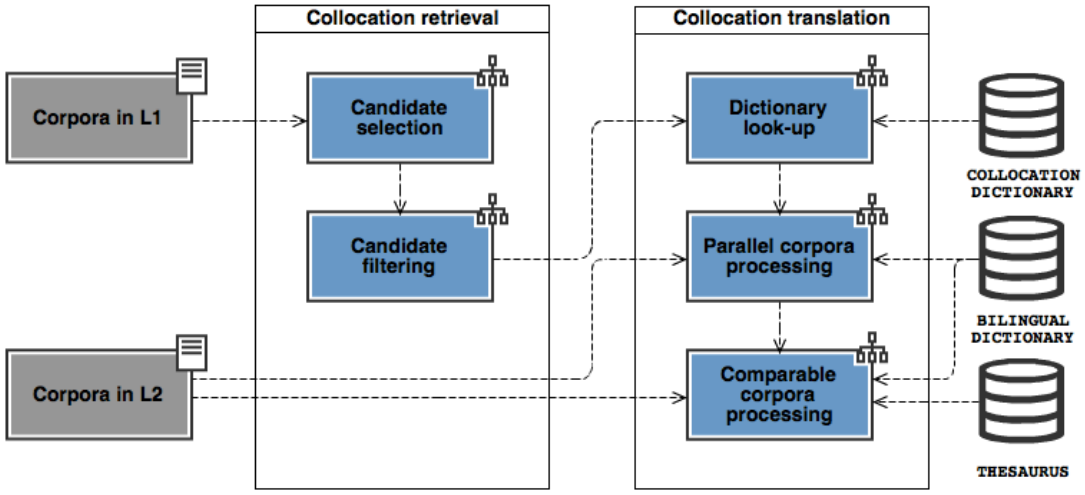

Figure 1: Architectural scheme of the system

POS-pattern definition aims at applying syntactic constraints on collocation candidates. This stage is language-dependent: as long as a language can be POS-tagged and a typology of the most commonly occurring collocations exists for it, POS-patterns can be defined. This task is simplified because the MWEToolkit supports the definition of syntactic patterns of collocations to extract. These can include repetitions, negation, and optional elements, much like regular expressions (see Figure 3, a definition of the English POS-pattern *NN or JJ + NN*). When retrieving collocations, each sentence in the corpus is matched against this set of patterns, and all n-grams which do not comply with any of them are ignored. Patterns that correspond exactly to the typologies of collocations in English and Spanish presented above have been defined (see Section 2.2).

```
<s s_id="0">
   <w surface="Harry" pos="NP" lemma="Harry"/>
   <w surface="unwrapped" pos="VBD" lemma="unwrap"/>
   <w surface="his" pos="PP$" lemma="his"/>
   <w surface="chocolate" pos="NN" lemma="chocolate"/>
   <w surface="frog" pos="NN" lemma="frog"/>
   <w surface="." pos="SENT" lemma="."/>
</s>
```

Figure 2: Sample XML output of TreeTagger

```
<pat>
   <pat repeat="+">
      <either>
         <pat> <w pos="JJ"/> </pat>
         <pat> <w pos="NN"/> </pat>
      </either>
   </pat>
   <w pos="NN"/>
</pat>
```

Figure 3: Example of POS-pattern definition

4.2 Candidate filtering module

This module computes collocation candidates and assigns a weight to each of these according to its probability of representing a collocation. It corresponds to the statistical component of our hybrid approach to collocation retrieval and relies on the MWEToolkit to perform n-gram selection and statistical processing. The toolkit receives two XML files as input: a representation of all sentences in the corpus with all words described by linguistic properties (see Figure 2), and a set of user-defined POS-patterns (see Figure 3). It performs *n-gram selection* by matching each sentence in the corpus against all defined POS-patterns, producing a set of collocation candidates. Once candidates have been extracted, it performs *statistical processing* by computing the frequencies of each candidate's word components from the SL corpus. This information is used to calculate a log-likelihood score for each candidate. Candidates are then ranked according to their scores. Figure 4 presents a sample collocation candidate in English. As can be observed, the toolkit not only extracts the lemma form of a collocation (*lemon drop*), but also the different surface forms it appears in (*lemon drops*).

```
<cand candid="1305">
  <ngram>
    <w lemma="lemon" pos="NN"> <freq value="9" /> </w>
    <w lemma="drop" pos="NN"> <freq value="18" /> </w>
    <freq value="6" />
  </ngram>
  <occurs>
    <ngram>
      <w surface="lemon" lemma="lemon" pos="NN" />
      <w surface="drop" lemma="drop" pos="NN" />
      <freq value="4" />
    </ngram>
    <ngram>
      <w surface="lemon" lemma="lemon" pos="NN" />
      <w surface="drops" lemma="drop" pos="NN" />
      <freq value="2" />
    </ngram>
  </occurs>
</cand>
```

Figure 4: Sample collocation candidate

4.3 Dictionary look-up module

This module connects to the online translation dictionary *WordReference* to attempt a direct translation in TL of a collocation in SL. *WordReference* translation entries include two tables: one for one-word direct translations (*principal translations*), and another for translations of MWEs (*compound forms*). Furthermore, the dictionary lists its translation entries in order from the most common to the least common. A Python script was written to handle the connection to the *WordReference* API. Our task is to look at the *compound forms* table and attempt to find a match for our collocation. If such a match is found, its translation from the HTML is extracted, and presented to the user. If no match is found, then translation will be based on the bilingual corpora presented by the user as input, triggering the *parallel corpora* or the *comparable corpora module* accordingly.

4.4 Parallel corpora module

This module first employs Hunalign to align the input corpora. Next, after *syntactic processing* and *semantic processing*, transformational rules are applied in order to identify the TL translations of all collocations extracted from SL. A sample output of Hunalign is presented in Figure 5: the first column refers to a sentence number in the SL corpus, the second column refers to a sentence number in the TL corpus, and the third column represents a confidence value, or the estimated certainty of the SL-TL pairing.

```
101     96      0.336927
102     97      0.583117
103     98      0.228
104     99      0.229412
105     100     0.226056
```

Figure 5: Sample Hunalign output

Semantic processing consists in identifying the base of a collocation in SL and finding its translation in TL. The POS-tags of the components of the collocation (see Figure 4) will help completely determine its base. This is because the POS-pattern of the collocation should adhere to one of the set of POS-patterns defined previously (see Figure 3). Next, the components representing the base for collocations will be identified following their linguistic model (see Section 2.1). Finally, *WordReference* is employed to retrieve the first three translation entries that match the POS-tag of our base from its *principal translations* table.

Similarly, *syntactic processing* consists in finding the translations of the collocatives in the TL corpus. It requires the output of both the *candidate filtering* module, which is an XML file containing a set of SL collocations (see Figure 4) and that of Hunalign presented above (see Figure 5). It also requires, as input, the TL corpus, which is a translation of the SL corpus. We implemented an algorithm that first reads the SL corpus and finds all sentences where a collocation appears, and then performs these tasks for each of the retrieved SL sentences:

- Read the output of Hunalign and match the SL sentence to its TL counterpart, where the translation of the collocation should appear.
- Expand this TL sentence to a window of five sentences to be extracted and analysed, to make up for any Hunalign precision error.
- For each of the translations in TL of the collocation's base, obtained during semantic processing, go through our window of sentences, one sentence at a time, and look for the presence of the translation within it. If a match is found, it means the translation of the collocatives in TL should also be present within the sentence.
- POS-tag the matching TL sentence using TreeTagger.
- Apply a transfer rule (see Table 3) to obtain the translation of the collocatives in TL.

4.5 Comparable corpora module

This module computes similarity classes in order to find the TL translations of all extracted SL collocations via *query expansion*, *query translation*, and *context generalisation*.

Query expansion produces a generalisation of the SL collocation's context by computing two different similarity classes, one centred on the base of the collocation, and another on its context (two open-class words that appear before it, and two after it). Computing similarity classes requires the use of a thesaurus. For English, we use WordNet, and obtain the first five synsets of the same POS-tag of any given open-class word. As for Spanish, *WordReference* is made use of. Our first similarity class, the one centred on the base of the collocation, will thus consist of up to six words, the original base itself and up to five synonyms. Correspondingly, our second similarity class will consist of up to 24 words: the four context words we retrieved, and up to five synonyms for each of them.

Next in the pipeline process is *query translation*, which computes a translation class, an expansion of the target language translations of the words that make up our original similarity class. Here again, we rely on *WordReference* as our de facto bilingual dictionary and thesaurus. For each of our two similarity classes, we iterate through all of their words, look up each via the *WordReference* API and retrieve up to five translation entries that match their POS-tags, and then further expand these by retrieving up to five thesaurus entries for each. This means that our first translation class, the one centred on the base of the collocation, will contain up to 30 translations for each of the (up to) six words of its similarity class, which totals up to 180 words. Similarly, our second translation class, centred on context words, will contain up to 720 words.

Finally, *context generalisation* aims at finding TL translations of a SL collocation by comparing context similarities. We first determine the POS-pattern of our SL collocation, and then see if any of the words in the translation class of its base corresponds with the base of any of the TL collocations of the same POS-pattern. If a match is found, we compute a similarity class for the context of the matched TL collocation and we see if it has any elements in common with the context of the SL collocation. If it does, we present it to the user as a potential translation of the collocation from the original text.

5 Evaluation

The choice of our experimental corpora was made completely on the basis of the profiles of the target users of our system: language learners and translators. Reading in a target language is an integral component of any language-learning process. We chose *Harry Potter and the Philosopher's Stone* and its translation into Spanish, *Harry Potter y la Piedra Filosofal*, to exemplify this. Similarly, professional translators usually specialise in a certain domain of translation, and therefore must translate technical terminology on a regular basis. Thus, we chose the *Ecoturismo corpus*[2], a collection of multilingual parallel and comparable corpora on tourism and tourism law, as it represents a real-life example of the technical documents a translator works with.

5.1 Experimental setup

Two bilingual annotators, fluent in English and Spanish, reviewed the output of our system after processing both experimental corpora. They assigned a score to the translations the system offered for each collocation according to a five-point scale (with 5 representing an excellent translation). Precision and recall are estimated from these scores for each case study.

5.2 Experimental results

100 English collocations were retrieved from the *Harry Potter* corpus. 12 collocations were successfully translated directly, using *WordReference*, such as *to talk rubbish*, *to speak calmly*, *fast asleep*, and *to lean against the wall*. Out of the remaining 88 collocations, 10 could not be translated at all, and 78 were translated using our approach to processing parallel corpora. Table 4 summarises these results (*A* stands for annotator, *WR* for *WordReference*, and *AVG* for average).

A	WR	1	2	3	4	5	AVG
#1	12	0	0	11	16	51	4.51
#2		0	0	8	17	53	4.58

Table 4: Parallel corpora result scores

As it can be observed, we obtained a high average score of 4.55 for the quality of translations retrieved from parallel corpora. Moreover, only 10 collocations out of the original 100 could not be translated, yielding an equally high score for recall, of 90%.

Similarly, 100 Spanish collocations were retrieved from the *Ecotourism* corpus. 15 of them were translated using *WordReference*; all of these were of the Spanish POS-patterns *NN + JJ or NN + de + NN*, such as *transporte público*, *asistencia técnica*, and *viaje de negocios*. Out of the 85 remaining collocations, 15 could not be translated at all, and the other 70 received translation suggestions found in the comparable corpora. Table 5 summarises the results.

A	WR	1	2	3	4	5	AVG
#1	15	7	14	19	17	13	3.21
#2		8	15	21	15	11	3.09

Table 5: Comparable corpora result scores

Despite the rather low average score of 3.15 for the quality of translations, we managed to provide translation suggestions to 85% of the collo-

[2] Compiled in the framework of the R&D project *Espacio Único de Sistemas de Información Ontológica y Tesaurus sobre el Medio Ambiente: ECOTURISMO* (Spanish Ministry of Education, FFI2008-06080-C03-03).

cations. We can conclude that by imposing flexible constraints on the matching process performed during the task of context generalisation, we obtain average translations for a high number of collocations. These constraints refer to the size of our context window and the number of thesaurus entries we retrieve for each original word during query expansion. Improving our precision score would mean strengthening these constraints, but this would also result in a lower recall. Moreover, in this particular case, recall of the output is more relevant than precision because our suggested translations, even if not always excellent, might offer translators a useful hint for correctly translating collocations.

5.3 Discussion and future work

Against the background of the limitations of the current version of our system, we propose the following future improvements. First, we exploit the nature of collocations as cohesive lexical clusters, but disregard the linguistic property of semantic idiomaticity that differentiates them from other MWEs, such as idioms. Our system cannot, therefore, differentiate between collocations and other MWEs in terms of compositionality. Secondly, we would like to provide better integration between the stages of collocation extraction and collocation translation. Currently, the former relies on TreeTagger and the MWEToolkit, while the latter makes use of Hunalign. This means that all users would also have to have access to these three tools; this poses no significant problem because all of them are open source, and readily available online, but it would be simpler to integrate the tasks performed by these tools into our system in order to increase its ease of use. Finally, we would like to investigate the use of the web as a corpus to find proficient ways of using information offered by search engines.

The expected final users of our system correspond to one of two groups: professional translators and language learners. However, as aforementioned, further fine-tuning of the system might be worthwhile in order to better address the specific needs of these particular user groups. Working with comparable corpora is not highly reliable because of its noisy nature. We opted to impose flexible constraints on the matching process performed during the last stage of comparable corpora processing, context generalisation, in order to increase the recall of our system. As stated before, this would be better suited to translators, who could benefit from the translation suggestions offered by our system to find the most adequate translation of a collocation. Language learners, however, are probably more interested in learning very precise translations for several collocations, rather than translation suggestions for a large number of collocations. A way forward would be to adjust the comparable corpora algorithm so it can impose stronger constraints during the task of context generalisation, to the benefit of language learners.

Future research goals could include (1) providing better integration between the different stages of the project, (2) finding a way to further exploit the use of the web as a corpus to aid in the processes of collocation retrieval and translation, (3) demonstrating the flexibility of our framework by adjusting our system to work with several other languages, and (4) tailoring the constraints imposed by our system to better meet the needs of our final users.

Acknowledgements

This project was supported by the European Commission, Education & Training, Erasmus Mundus: EMMC 2008-0083, Erasmus Mundus Masters in NLP & HLT programme.

References

Baldwin, T., and Kim, S. N. (2010). Multiword Expressions. In: *Handbook of Natural Language Processing*, second edition. Boca Raton, FL.

Biber *et al.* (1999). *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Bradford, W., and Hill, S. (2000). *Bilingual Grammar of English-Spanish Syntax*. University Press of America.

Brown P., Lai J., and Mercer R. (1991). Aligning Sentences in Parallel Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Canada, pp. 169-176.

Cardey, S., Chan, R. and Greenfield, P. (2006). The Development of a Multilingual Collocation Dictionary. In: *Proceedings of the Workshop on Multilingual Language Resources and Interoperability*, Sydney, pp. 32-39.

Choueka, Y., Klein, T., and Neuwitz, E. (1983). Automatic Retrieval of Frequent Idiomatic and Collocational Expressions in a Large Corpus. In: *Journal for Literary and Linguistic Computing*, 4(1): pp. 34-38.

Church, K. W., and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In: *Proceedings of the 27th annual meeting on Association for Computational Linguistics*, pp. 76-83.

Corpas Pastor, G. (1995). *Un Estudio Paralelo de los Sistemas Fraseológicos del Inglés y del Español*. Málaga: SPICUM.

Corpas Pastor, G. (1996). *Manual de Fraseología Española*. Madrid, Gredos.

Corpas Pastor, G. (2013). Detección, Descripción y Contraste de las Unidades Fraseológicas mediante Tecnologías Lingüísticas. Manuscript submitted for publication. In *Fraseopragmática,* I. Olza and E. Manero (eds.). Berlin: Frank & Timme.

Fung, P., and Yuen, Y. (1998). An IR Approach for Translating New Words from Nonparallel, Comparable Texts. In: *Proceedings of the 17th International Conference on Computational Linguistics*, pp. 414-420.

Gale W., and Church K. (1993). A Program for Aligning Sentences in Bilingual Corpora. In: *Journal of Computational Linguistics*, 19: pp. 75-102.

Gelbukh A., and Kolesnikova O. (2013). Expressions in NLP: General Survey and a Special Case of Verb-Noun Constructions. In *Emerging Applications of Natural Language Processing: Concepts and New Research*, S. Bandyopadhyay, S. K. Naskar, and A. Ekbal (eds.). Hershey: Information Science Reference. IGI Global.1-21.

Hausmann, F. (1985). Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: *Lexikographie und Grammatik*, (Lexicographica, series maior 3), ed. H. Bergenholtz and J. Mugdan. Tübingen: Niemeyer. 175-186.

H.H. Hoang, S.N. Kim, M.Y. Kan, A Re-examination of Lexical Association Measures, In *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP*, Singapour: ACL and AFNLP. 31-39.

Jackendoff, R. (2007). *Language, Consciousness, Culture: Essays on Mental Structure.* The MIT Press.

Lea D. and Runcie, M. (2002). *Oxford Collocations Dictionary for Students of English*. Oxford University Press.

Lü, Y. and Zhou, M. (2004). Collocation Translation and Acquisition Using Monolingual Corpora. In: *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04)*, pp. 167-174.

Ramisch, C., Villavicencio, A., and Boitet, C. (2010). MWEToolkit: A Framework for Multiword Expression Identification. In: *Proc. of LREC'10 (7th International Conference on Language Resources and Evaluation)*.

Ramisch, C. (2012). A Generic Framework for Multiword Expressions Treatment: from Acquisition to Applications. In: *Proceedings of ACL 2012 Student Research Workshop*, pp. 61-66.

Rapp, R. (1995). Identifying Word Translations in Nonparallel Texts. In: *Proceedings of the 35th Conference of the Association of Computational Linguistics*, pp. 321-322. Boston, Massachusetts.

Sag, I. et al. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (COCLing-2002)*, pp. 1-15.

Santana, O. *et al*. (2011). Extracción Automática de Colocaciones Terminológicas en un Corpus Extenso de Lengua General. In: *Procesamiento del Lenguaje Natural*, (47): pp. 145-152.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of International Conference on New Methods in Language Processing*. Manchester, UK.

Seretan, V. (2011). *Syntax-Based Collocation Extraction (Text, Speech and Language Technology)*, 1st Edition. Springer.

Sharoff, S., Babych, B., & Hartley, A. (2009). "Irreferragable answers" using comparable corpora to retrieve translation equivalents. In: *Language Resources and Evaluation*, 43(1): pp. 15-25.

Sinclair, J., and Jones, S. (1974). English Lexical Collocations: A study in computational linguistics. In: *Cahiers de lexicologie*, 24(2): pp. 15-61.

Smadja, F. (1993). Retrieving collocations from text: Xtract. In: *Computational Linguistics*, 19(1): pp. 143-177.

Varga, *et al.* (2005). Parallel corpora for medium density languages. In*: Proceedings of the RANLP 2005*, pp. 590-596.

Wehrli, E., Nerima, L., and Scherrer, Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. In: *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 90-94.

When Multiwords Go Bad in Machine Translation

Anabela Barreiro
L^2F - INESC-ID
Rua Alves Redol, 9
1000-029 Lisbon, Portugal
anabela.barreiro@inesc-id.pt

Johanna Monti
UNISS
Via Roma 151
07100 Sassari, Italy
jmonti@uniss.it

Brigitte Orliac
Logos Institute
1636 Pelots Point road
North Hero, VT 05474, USA
orliac.brigitte@gmail.com

Fernando Batista
L^2F - INESC-ID and ISCTE-IUL
Rua Alves Redol, 9
1000-029 Lisbon, Portugal
fernando.batista@inesc-id.pt

Abstract

This paper addresses the impact of multiword translation errors in machine translation (MT). We have analysed translations of multiwords in the OpenLogos rule-based system (RBMT) and in the Google Translate statistical system (SMT) for the English-French, English-Italian, and English-Portuguese language pairs. Our study shows that, for distinct reasons, multiwords remain a problematic area for MT independently of the approach, and require adequate linguistic quality evaluation metrics founded on a systematic categorization of errors by MT expert linguists. We propose an empirically-driven taxonomy for multiwords, and highlight the need for the development of specific corpora for multiword evaluation. Finally, the paper presents the Logos approach to multiword processing, illustrating how semantico-syntactic rules contribute to multiword translation quality.

1 Introduction

Multiwords play a crucial role in natural language processing. The lack of formalization or inadequate processing of multiwords triggers problems with the syntactic and semantic analysis of sentences where these multiwords occur and reduces the performance of natural language processing systems. Multiwords are essential in MT, and their incorrect generation has a severe impact on the understandibility and quality of the translated text.

The most common sources of errors in multiword processing is fragmentation. Since a multiword embeds semantic meaning as a whole, fragmentation of any part of a multiword leads to incorrect translation. Currently, with a few exceptions, most MT systems present severe weaknesses at effectively addressing the lack of compositionality of multiwords. In fact, an analysis of translations provided by freely available MT demonstrates that the translation of multiwords is a problem area for RBMT and SMT. RBMT systems fail for lack of multiword coverage, while SMT systems fail for not having linguistic (semantico-syntactic) knowledge to process them, leading to serious structural problems.

This paper describes an evaluation exercise that consists in the linguistic analysis and error categorization of the problems encountered in multiword translations performed by the OpenLogos RBMT and the Google Translate SMT systems for the English-French, English-Italian and English-Portuguese language pairs. The different types of translation errors were post-edited and categorized linguistically by MT expert linguists of the respective target languages. We used a corpus of 150 sentences containing an average of about 5 multiwords per sentence. Based on this corpus, we developed a multiword taxonomy that can be used to evaluate multiwords in any type of system, independently of the approach. This paper also presents the OpenLogos solution to the problem of multiword processing in MT.

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.

The remaining of the paper is organized in the following way: Section 2 describes the main properties of multiwords and stresses the need to evaluate multiwords from a linguistic point-of-view. Section 3 describes the state-of-the-art of the RBMT and the SMT approaches to multiword processing. Section 4 describes the corpus and the multiword taxonomy used to categorize the errors found in the translations provided by the OpenLogos and Google Translate MT systems. Section 5 presents some quantitative results and the analysis of the most important problems encountered in the French, Italian, and Portuguese translations of the multiwords in our corpus by the two systems. Section 6 underlines a unique feature of the OpenLogos machine translation system, namely the semantico-syntactic rules used to improve multiword translation precision. Finally, Section 7, presents the main conclusion and points to future work.

2 Multiwords

A multiword (short for multiword unit) is a group of two or more words in a language lexicon that generally conveys a single meaning. Multiwords are abundant in language, but until recently they have been given little focus by traditional theoretical linguistics. Grammars describe them inconsistently, and they are not formalized adequately in dictionaries or applied successfully to MT. The most critical problems in multiword processing is that they often have unpredictable, non-literal translations. Literal translations of idiomatic multiwords are often not understandable because the meaning of the multiword cannot be derived simply from the meaning of the individual constituents that make up the single unit. Multiwords may have different degrees of compositionality varying from free combinations to frozen expressions, and their morphosyntactic properties allow, in some cases, a number of variations with the possibility of constituent dependencies, even when the constituents are distant of each other in the sentence. These problems, along with the difficulty of including all multiwords in dictionaries, make some approaches incapable of processing them correctly.

Multiwords can be classified into three main categories: lexical units, frozen and semi-frozen expressions, including proverbs, and lexical bundles (Barreiro, 2010). Some multiword expressions do not fit into any of these three major types. For example, institutionalized utterances, such as *let's go!*, or *if you will*, sentence frames such as *as follows*, and non-contiguous text frames such as *on one side... on the other*, classified independently as compound adverbs, can also be seen as special types of multiword. Idioms, such as [*to*] *purr like a cat* and *for goodness' sake* are semi-frozen or frozen expressions that can fit in one or another class. Many semi-frozen expressions correspond to variable types of support verb construction, such as *take a seat* or *play a* [*very important*] *role*, which are further characterized by the possible insertion of external elements (inserts) inside the support verb construction. Section 4 presents a multiword taxonomy that takes into account contiguous (adjacent) and non-contiguous (remote) multiwords.

3 State-of-the-Art MT Approaches to Multiword Processing

Several authors have pointed out the importance of a correct processing of multiwords so that they can be translated correctly by MT systems (cf. (Sag et al., 2001), (Thurmair, 2004), (Rayson et al., 2010), (Monti, 2013), among others). Solutions to resolve multiword translation problems vary from (i) using generative dependency grammars with features (Diaconescu, 2004); (ii) grouping bilingual multiwords before performing statistical alignment (Lambert and Banchs, 2006); and (iii) paraphrasing them (Barreiro, 2010). The combination of different multiword processing solutions will contribute to a more successful MT approach. Sections 3.1 and 3.2 describe multiword processing in the RBMT and the SMT approaches respectively.

3.1 Multiword Processing in RBMT

In RBMT, the identification of multiwords is based on two main different approaches: the lexical approach and the compositional approach. In the lexical approach, multiwords are considered single lemmata and lemmatized in the system dictionaries. This approach is particularly suitable for contiguous compounds, which can be easily lemmatised.

In the compositional approach, multiword processing is obtained by means of part-of-speech tagging and syntactic analysis of the different components of a multiword. This approach is particularly useful for translating compound words not coded in the system dictionary, but it is also com-

monly used for translating verbal constructions. According to this approach, the single elements of a multiword are looked up in the system dictionary and analysed according to the information coded in them. Once the different constituents of multiwords have been identified and disambiguated, a rule is applied to properly translate the combination of the different words in a single unit of meaning.

3.2 Multiword Processing in SMT

In SMT, the problem of multiword processing is not specifically addressed. The traditional approach to word alignment following IBM Models (Brown et al., 1993) shows many shortcomings concerning multiword processing, especially due to the inability of this approach to handle many-to-many correspondences.

In the current state-of-the-art phrase-based SMT systems (Koehn et al., 2003), the correct translation of multiwords occurs only if the constituents of multiwords are marked and aligned as parts of consecutive phrases in the training set and they are not treated as special cases. Phrases are defined as sequences of contiguous words (n-grams) without any or with limited linguistic information. Some word combinations are, in fact, linguistically meaningful (e.g., *will stay*), but many of them have no linguistic significance at all (e.g., *that he*). Multiword processing and translation in SMT started being addressed only recently, and different solutions have been proposed that consider multiword errors either as a problem of automatically learning and integrating translations or as a word alignment problem (Barreiro et al., 2013).

Current approaches to multiword processing are moving towards the integration of phrase-based models with linguistic knowledge, and scholars are starting to use linguistic resources, either hand crafted dictionaries and grammars or data-driven ones, in order to identify and process multiwords as single units. The most widely used methodology consists in identifying possible monolingual multiwords (Wu et al., 2008) (Okita et al., 2010), among others. (Ren et al., 2009), instead, have underlined that the integration of bilingual domain multiwords in SMT could significantly improve translation performance. Other solutions are based on the incorporation of machine-readable dictionaries and glossaries, treating these resources as phrases in the phrase-based table (Okuma et al., 2008), and on the identification and grouping of multiwords prior to statistical alignment (Lambert and Banchs, 2006).

The identification and disambiguation of multiwords have also been considered a problem of word sense disambiguation (WSD) and proposals have been made to integrate WSD in SMT. Methods in this research area range from (i) supervised methods that make use of annotated training corpora, (ii) semi-supervised or minimally supervised methods that rely on small annotated corpora as seed data in a bootstrapping process, (iii) word-aligned bilingual corpora, or (iv) unsupervised methods that work directly from raw unannotated corpora. A more detailed description and analysis of the different approaches to multiword processing in SMT can be found in (Monti, 2013).

4 Corpus and Multiword Taxonomy

The corpus used in this research task contains 150 English sentences extracted randomly from an existing corpus of sentences gathered from the news and the internet. Each multiword under evaluation was annotated in the context of its sentence and classified according to the taxonomy presented in Table 1. The corpus was divided into three sets of 50 sentences each, and each set was then translated into French, Italian, and Portuguese respectively, using the OpenLogos and the Google Translate MT systems. The purpose of our study was not to compare and evaluate systems, but to assess and measure the quality of multiword unit translation independently of the two systems considered. Three native linguists, who are also MT experts, reviewed 50 sentences each for the three target languages, and evaluated the multiword translations for each of these languages (one evaluator for each language), classifying the translations according to a binary evaluation metrics: OK for correct translations and ERR for incorrect ones. After classifying the multiword translations, evaluators were asked to provide a more comprehensive evaluation of multiword translations according to the different types of multiword. None of the systems was specifically trained for the specific task, as the texts were not domain specific.

5 Quantitative Results

The results obtained in this study shed some light on the demand for higher precision multiword

VERBS (V)
Compound Verb (COMPV)
Contiguous (COMPV) *can learn; may have been done*
Non-contiguous (NON-CONT COMPV) *have [already] shown*
Support Verb Construction (SVC)
Nominal (NSVC) *make a presentation*
Adjectival (ADJSVC) *be meaningful*
Non-contiguous nominal (NON-CONT NSVC) *have [particularly good] links*
Non-contiguous adjectival (NON-CONT ADJSVC) *be ADV selective*
Prepositional nominal (PREPNSVC) *give an illustration of*
Prepositional adjectival (PREPADJSVC) *be known as; be involved in*
Non-contig prep nominal (NON-CONT PREPNSVC) *be the ADV cause of*
Non-contig prep adj (NON-CONT PREPADJSVC) *fall [so far] short of*
Prepositional Verb (PREPV)
Contiguous (PREPV) *deal with*
Non-contiguous (NON-CONT PREPV) *give N to*
Phrasal Verb (PHRV)
Contiguous (PHRV) *closing down*
Non-contiguous (NON-CONT PHRV) *make N up*
Prepositional (PREPPHRV) *slow down to*; *stand up to*
Non-contig prep (NON-CONT PREPPHRV) *mix N up with*
Other Verbal Expression (VEXPR)
Contiguous (VEXPR) *in trying to*
Non-contig (NON-CONT VEXPR) *hold N in place*
NOUNS (N)
Compound Noun (COMPN)
Common compound noun (*union spokesman*)
Domain term (*constraint-based grammar*)
Prepositional Noun (PREPN)
Simple (PREPN) (*interest in*)
Compound (COMPPREPN) *right side of*
ADJECTIVES (ADJ)
Compound Adjective (COMPADJ) *cost-cutting*
Prepositional Adjective (PREPADJ) *famous for; similar to*
ADVERBS (ADV)
Compound Adverb (COMPADV) *in a fast way; most notably; last time*
Prepositional Adverb (PREPADV) *in front of*
DETERMINERS (DET)
Compound Determiner (COMPDET) *certain of these*
Prepositional Determiner (PREPDET) *most of*
CONJUNCTIONS (CONJ)
Compound Conjunction (COMPCONJ) *in order to; as a result of; rather than*
PREPOSITIONS (PREP)
Compound Preposition (COMPPREP) *as part of*
OTHER EXPRESSIONS (OTHER)
Named Entity (NE) *Economic Council*
Idiom (IDIOM) *get to the bottom of the situation*
Lexical Bundle (BUNDLE) *I believe that; as much if not more than*

Table 1: Categories of multiword in our corpus

translation. Section 5.1 shows the global performance of each system with regards to multiwords, and Section 5.2 highlights system performance with regards to multiword type, presenting some indicators on which types of multiword are more problematic for each system, without any intention to compare multiword performance between systems.

5.1 Overall Performance by Language Pair

Multiwords occur very frequently in our corpus, often several times within the same sentence. For example, the English sentence *Witnesses said the speeding car may have been playing tag with another vehicle when it veered into the southbound lane occupied by Lopez' truck shortly before 8 p.m. Sunday* contains the following 4 multiwords: (i) the compound verb within the idiomatic prepositional support verb construction *may have been playing tag with*; (ii) the prepositional verb construction *veered into*; (iii) the nominal compound *southbound lane* and (iv) the double temporal expression (time + date) *8 p.m. Sunday*. Table 2 represents the total of multiwords found in the sentences translated for each language pair by the OpenLogos and Google Translate MT systems.

5.1.1 English-French

For French, a total of 196 multiwords were found in the 50 sentences analysed, representing an average of 3,92 multiwords per sentence. 110 of these multiwords were translated correctly and 86 were translated incorrectly. From the 88 multiwords found in the sentences translated by OpenLogos, 40 were translated correctly and 48 were translated incorrectly. From the 108 multiwords found in sentences translated by Google Translate, 70 were translated correctly and 38 were translated incorrectly.

5.1.2 English-Italian

For the Italian language, a total of 225 multiwords occurred in the 50 sentences analysed, representing an average of 4,5 multiwords per sentence. 95 of those were translated correctly and 130 were translated incorrectly. From the 119 multiwords found in the sentences translated by OpenLogos, 36 were translated correctly and 83 were translated incorrectly. From the 106 multiwords found in sentences translated by Google Translate, 59 were translated correctly and 47 were translated incorrectly.

System	Lang pair	OK	ERR	Total
OL	EN-FR	40	48	88
	EN-IT	36	83	119
	EN-PT	60	96	156
	Total	136	227	363
GT	EN-FR	70	38	108
	EN-IT	59	47	106
	EN-PT	67	47	114
	Total	196	132	328

Table 2: Number of correct (OK) and incorrect (ERR) multiword translations per language pair and per MT system

EN-FR	OL		GT	
Type	Ok	Error	Ok	Error
VERB	17	21	27	12
COMPN	8	10	13	18
NE	6	4	16	4
EN-IT	OL		GT	
Type	Ok	Error	Ok	Error
COMPN	14	39	26	21
VERB	10	12	6	15
NE	2	8	14	2
EN-PT	OL		GT	
Type	Ok	Error	Ok	Error
VERB	30	21	11	23
COMPN	28	12	18	17
NE	11	26	9	9

Table 3: OL and GT performance for the 3 most frequent types of multiword in our corpus

5.1.3 English-Portuguese

For the Portuguese language, the 50 sentences contained a total of 270 multiwords, representing an average of 5,4 multiwords per sentence. Overall, 47% of all multiwords were translated correctly (127 counts of OK), 53% were translated incorrectly (143 counts of ERR) by both systems. From the 156 multiwords found in the sentences translated by OpenLogos, 60 (38,5%) were translated correctly and 96 (61,5%) were translated incorrectly. From the 114 multiwords found in sentences translated by Google Translate, 67 (58,5%) were translated correctly and 47 (41,5%) were translated incorrectly.

5.2 Performance on Multiword Type

Table 3 shows the performance of the OpenLogos and Google Translate systems when translating the most frequent types of multiword.

5.2.1 English-French

For the English-French language pair, the largest category of multiword errors involved compound nouns. Incorrectly translated general language or domain-specific compound nouns represented 32,5% of all multiword errors. Some examples include *hit-run driver*, *cause-and-effect relationship*, *wage and price control legislation*, *compact digital audio disk*, *recession velocity*, and *nuclear fuel cycle*, among others. The second largest category of multiword errors were support verb constructions, representing 18,6% of all multiword errors (e.g., [*to*] *go on strike*, [*to*] *bring order* (nominal) or [*to*] *be* [*directly*] *related*, [*to*] *be* [*a bit*] *misleading* (adjectival)). Half of the errors in the support verb construction category involved non-contiguous expressions, such as [*to*] *gather* [*new*] *evidence*, and [*to*] *have* [*wide*] *applicability*. Another fairly large number of multiword errors (13,9%) involved prepositional verb constructions such as [*to*] *serve as*, or [*to*] *generalize upon*, with non-contiguous expressions representing more than half of all prepositional verb constructions errors ([*to*] *protect* [*the public*] *from*, or [*to*] *roll* [*three times*] *down*). Finally, incorrectly translated named entities accounted for 9,3% of the total number of multiword errors (*Rocky Mountain News*, *Christian Broadcasting Network*, *South Platte River*).

5.2.2 English-Italian

The most common mistranslations concerned general language or domain-specific compound nouns, which represented 46% of all multiword errors. Some examples include *windfall profits tax*, *court file*, *115 Vac receptacle*, *Party-State*, among others. The second largest critical area concerned the translation of multiword verbs, representing 16% of all multiword errors. Within this area, prepositional verbs mistranslations were the most common ones, corresponding to 9% of multiword errors. Examples are [*to*] *deal with* and [*to*] *rest upon*. Errors concerning this type of verb constructions were mostly related to non-contiguous constructions like *being acquired [automatically] from* and *has not patterned [its labor contract] after [that of its largest competitor]*. Support verb construction errors also occurred, accounting for 2% of all multiword errors, including adjectival support verb constructions, such as *[to] seem clear*. While these two categories, compound nouns and verb constructions, accounted for the lion's share of multiword errors, other critical areas included (i) named entities (3%), such as *Capitol Hill*, *Esprit's Compulog Net*, (ii) compound adverbs (3%), such as *in short*, and finally (iii) id-

iomatic expressions which were almost all incorrectly translated and included expressions such as *idle pipe dreams*.

5.2.3 English-Portuguese

The most frequent multiword error type occurring in sentences translated from English into Portuguese was multiword verbs. We counted 83 different structures of the verb subtypes, of which more than 50% (44) were translated incorrectly by the two machine translation systems. Within multiword verbs, errors with prepositional verb constructions accounted for 6,2% of all multiword errors. Examples of such expressions are: [to] *focus on*, [to] *veer into* or [to] *merge with*. Many prepositional verb constructions were non-contiguous, such as *stopped* [*momentarily*] *along*, and [*to*] *pay* [*Disney*] [*$100 million*] *for*. Support verb construction errors also occurred frequently accounting for 4,8% of all multiword errors. This category included contiguous support verb constructions, such as *give an illustration of* and non-contiguous support verb constructions, such as *has* [*particularly good*] *links with*. The second largest category of multiword errors were compound nouns. Incorrectly translated general language or domain-specific compound nouns represented 31% of all multiword errors. Some examples include *island nation*, *southbound lane*, *top player*, *hybrid constraint-based grammars*, *machine learning*, and the prepositional compound noun *right side of*, among others. Finally, incorrect translations of named entities represented the third most common problem in Portuguese, with 35 errors in both systems.

6 OpenLogos Approach to Multiword Processing in Machine Translation

One of the most intelligent approaches to multiword processing in RBMT is carried out by the former Logos system, now OpenLogos (Scott, 2003) (Scott and Barreiro, 2009) (Barreiro et al., 2011). The question of how to represent natural language inside a computer was answered in the OpenLogos system by the Semantico-syntactic Abstraction Language, known as SAL[1]. SAL is an abstract hierarchical language (consisting of supersets, sets and subsets) that represents the driving force of the translation process. The first activity that the system performs on a natural language sentence is to convert it to a SAL sentence before parsing can take place. SAL combines both the lexical and the compositional approaches in order to process different types of multiword.

The underlying philosophical principle of the OpenLogos system is to merge the syntactic and semantic information into SAL, so semantic knowledge is available at different stages of the translation process to help in the resolution of ambiguities at every linguistic level, including the lexicon. At the end of the process an abstract, formal and semantico-syntactic SAL representation of the source language is obtained, and subsequently translated into the target language.

The main linguistic knowledge bases of the OpenLogos system are (i) dictionaries; (ii) semantico-syntactic rules for analysis, transfer and generation; and (ii) Semantic Table (henceforth SEMTAB) rules. The SEMTAB database contains thousands of language-pair specific transformation rules that provide special analysis, formalization, and translation of words in context.

An important function of SEMTAB is to disambiguate the meaning of words by seeing them in their semantico-syntactic context. SEMTAB rules are invoked after dictionary look-up and during the execution of target transfer rules (TRAN rules) in order to solve various ambiguity problems, including: (i) verb dependencies, such as the different argument structures of the verb *speak* (eg., *speak to*, *speak about*, *speak against*, *speak of*, *speak on*, *speak on N (radio, TV, television, etc.)*, *[speak over N1 (air) about N2]*; and (ii) multiwords of different nature.

In the processing of multiwords, SEMTAB context-sensitive semantico-syntactic rules play a very important role in complementing the dictionary, capturing the nuances of words that cannot be discerned at the pure syntactical level. For example, SEMTAB comprehends the different meanings of the verb *raise* on the basis of its objects: *raise an issue*, *raise a child*, *raise vegetables/crops*, *raise the roof*, *raise the rent*, etc. In *raise a child*, the verbs object is semantically marked as [Animate + Human]. When *raise* is combined with any other noun with the same semantic properties, SEMTAB effects an appropriate target transfer that overrides the default dictionary transfer for this verb. In *raise vegetables/crops*,

[1]freely available at `https://www.l2f.inesc-id.pt/~abarreiro/openlogos-tutorial/new_A2menu.htm`.

the verb's object is semantically marked as [Mass + Edible]. In *raise the rent*, the verb's object is semantically marked as [Measurement + Abstract measured by units (such as Euros)], and so on and so forth.

In conjunction with the semantic robustness provided by SAL, SEMTAB also gives OpenLogos the unusual powerful ability to process multiwords morpho-syntactically. Rules in SEMTAB are conceptual and deep-structure rules, which means that a single deep-structure rule can match a variety of surface structures, regardless of word order, passive/active voice construction, etc.. So, in the case of the verb *raise*, one single rule is applied to the following different surface structures: (i) *he raised the rent* [V+Object]; (ii) *the raising of the rent* [Gerund]; (iii) *the rent, raised by* [Participial ADJ]; and (iv) *a rent raise* [Process or Predicate Noun].

To sum up, SEMTAB provides the linguistic (semantico-syntactic) knowledge that is currently missing in SMT models. SEMTAB's structural analysis ability in combination with the rich word selection in the transfer powered by sophisticated SMT methods, which allow to extract knowledge from large amounts of parallel corpora, can be an effective solution to improve translation quality.

7 Conclusions

Currently, multiword processing still represents one of the most significant linguistic challenges for MT systems. In our study, the translation of multiwords by the OpenLogos and the Google Translate systems proves that a significant amount of work still needs to be done to successfully resolve the multiword translation problem. Literal translations of multiwords lead to unclear or incorrect translations or total loss of meaning. Adequate identification and analysis of source language multiwords is a challenging task, however, it is the starting point for higher quality translation.

We explained how the SEMTAB rules of the OpenLogos system can contribute to the translation of multiwords and influence the performance of any type of MT system with reference to any language pair. Due to length limitations, we did not discuss how linguistic knowledge, such as that provided by the OpenLogos SEMTAB, can be applied to a SMT system, but in the future, we aim to demonstrate how a multiword error by Google Translate can be corrected by OpenLogos (and how this correction can be applied in the system) and how a multiword error in OpenLogos can be fixed in a statistical system.

When the research community is able to combine the linguistic precision provided in the OpenLogos approach to the coverage provided in the SMT approach in resolving the multiword problem, an important evolution will take place in the MT field. The successful integration of semantico-syntactic knowledge in SMT represents an important solution for achieving high quality MT. The accomplishment of this task requires a combination of expertise in MT technology and deep linguistic knowledge. Independently of how the integration is implemented, we have no doubts that linguistic understanding and representation of multiwords will improve the state-of-the-art MT significantly and it is a necessary condition for enabling internet users and the general public to communicate more freely and more understandably across different languages.

Acknowledgements

This work was supported by Fundação para a Ciência e Tecnologia (Portugal) through Anabela Barreiro's post-doctoral grant SFRH/BPD/91446/2012 and project PEst-OE/EEI/LA0021/2013.

Autorship contribution is as follows: Anabela Barreiro is author of the Abstract, Sections 1, 2, 4, 5.1.3, 5.2, 5.2.3, and 7; Johanna Monti of Sections 3, 3.1, 3.2, 5.1.2, 5.2.2, and 6; Brigitte Orliac of Sections 5.1.1 and 5.2.1; and Fernando Batista of Sections 5 and 5.1.

References

Barreiro, Anabela, Bernard Scott, Walter Kasper, and Bernd Kiefer. 2011. Openlogos rule-based machine translation: Philosophy, model, resources and customization. *Machine Translation*, 25(2):107–126.

Barreiro, Anabela, Luísa Coheur, Tiago Luis, Angela Costa, Fernando Batista, Joao Graça, and Isabel Trancoso. 2013. Multiword and semantico-syntactic unit alignments. *Language Resources and Evaluation*, (submitted).

Barreiro, Anabela. 2010. *Make it Simple with Paraphrases: Automated Paraphrasing for Authoring Aids and Machine Translation*. Lambert Academic Publishing.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, M. J. Goldsmith, J. Hajic, R. L. Mercer, and S. Mohanty.

1993. But dictionaries are data too. In *Proceedings of the HLT*.

Diaconescu, Stefan. 2004. Multiword expression translation using generative dependency grammar. In *EsTAL*, pages 243–254.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lambert, Patrik and Rafael Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multi-Word-Expressions in a Multilingual Context*, EACL '06, pages 9–16, April 3rd.

Monti, Johanna. 2013. *Multi-word Unit Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation*. Ph.D. thesis, University of Salerno, Salerno, Italy.

Okita, Tsuyoshi, Alfredo Maldonado Guerra, Yvette Graham, and Andy Way. 2010. Multi-word expression-sensitive word alignment. In *Proceedings of the 4th Workshop on Cross Lingual Information Access*, pages 26–34, Beijing, China, August. Coling 2010 Organizing Committee.

Okuma, Hideo, Hirofumi Yamamoto, and Eiichiro Sumita. 2008. Introducing a translation dictionary into phrase-based smt. *IEICE - Trans. Inf. Syst.*, E91-D(7):2051–2057, July.

Rayson, Paul, Scott Songlin Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? *Language Resources and Evaluation*, 44(1-2):1–5.

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*, MWE '09, pages 47–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Scott, Bernard and Anabela Barreiro. 2009. OpenLogos MT and the SAL representation language. In Pérez-Ortiz, Juan Antonio, Felipe Sánchez-Martínez, and Francis M. Tyers, editors, *Proceedings of the First International Workshop on Free-Open-Source Rule-Based Machine Translation*, pages 19–26, Alicante, Spain. Departamento de Lenguajes y Sistemas Informáticos - Universidad de Alicante.

Scott, Bernard (Bud). 2003. The logos model: An historical perspective. *Machine Translation*, 18(1):1–72, March.

Thurmair, G. 2004. Multilingual content processing. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)*.

Wu, Hua, Haifeng Wang, and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 993–1000, Stroudsburg, PA, USA. Association for Computational Linguistics.

Using a Rich Feature Set for the Identification of German MWEs

Fabienne Cap & Marion Weller
IMS
Universität Stuttgart
Pfaffenwaldring 5B
70174 Stuttgart
cap@ims.uni-stuttgart.de
wellermn@ims.uni-stuttgart.de

Ulrich Heid
IwiSt
Universität Hildesheim
Lübeckerstrasse 3
31141 Hildesheim
heid@uni-hildesheim.de

Abstract

Due to the formal variability and the irregular behaviour of MWEs on different levels of linguistic description, they are a potential source of errors for many NLP applications, e.g. Machine Translation. While most of the known approaches to MWE identification focus on one dimension of irregular behaviour, we present an approach that combines morpho-syntactic features (extracted from dependency parsed text) with semantic opacity features (approximated using word alignments). We trained supervised classifiers with different feature sub-sets and show that the combination of morpho-syntactic and semantic opacity features yields best overall results.

1 Introduction

The task of automatically identifying multiword expressions (MWEs) has gained considerable interest in NLP research in the past years (Sag et al., 2002).

Due to the formal variability and the irregular behaviour of MWEs on different levels of linguistic description, they are a potential source of errors for many NLP applications: consider for example Machine Translation, where MWEs with (partially) opaque semantics can hardly ever be transferred word by word to the target language:

(1)

zur Sprache bringen =	lit. "to bring to speech"
	idiom. "to address sth."

We present a method to identify German MWEs of the type preposition+noun+verb based on a rich feature set comprising morpho-syntactic features extracted from monolingual data and cross-lingual features obtained from word-aligned parallel data (DE-EN, DE-SE). The used features aim at modelling characteristic properties of MWEs, namely *fixedness* in terms of their disposition for variation with respect to e.g. number or type of determiner (morpho-syntactic features) and *irregular translational behaviour*, e.g. a broad variation in translational equivalents (cross-lingual features). Our experiments show that combining these different types of features leads to an improved classification accuracy.

Our approach consists of three main steps:

1. extraction of syntactically related multiword constructions
2. collect, sum and average feature values of all their occurrences
3. train a classifier on a hand-crafted dataset to distinguish unseen MWEs from regular combinations.

The remainder of this paper is structured as follows: In Section 2, we briefly describe our objectives and the specificities of MWE extraction for German. We give an overview of the morpho-syntactic and cross-lingual features and explain how we extract them in Section 3. Then, in Section 4, we describe the data and show how feature values are integrated to train the classifier. The experiments are presented in Section 5 and the results are discussed in Section 6. We report on related work in Section 7 and finally, we conclude in Section 8.

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.

2 Background

2.1 Objectives and State of the Art

Our approach to the identification of German MWEs (of the type preposition+noun+verb) makes use of their morpho-syntactic properties and their semantic transparency vs. opacity. We classify MWEs on the level of lexical types into idiomatic ones vs. trivial (non-idiomatic) word combinations.

We start from the widely shared assumption that idiomaticity is often correlated with morpho-syntactic fixedness, e.g. Bannard (2007), Fazly and Stevenson (2006), Weller and Heid (2010) and that idiomaticity implies an element of non-compositionality, e.g. Baldwin et al. (2003). The former type of features is observable in morpho-syntactically analysed data; for the latter, which is not directly observable in monolingual corpora, we follow Villada Moirón and Tiedemann (2006) and Fritzinger (2010) and induce transparency vs. opacity from bilingual corpora.

We operate on MWE types, not on tokens; we consider individual occurrences and then sum up and average the feature values observed for one MWE type. Obviously, not all co-occurrences of lexemes that can be part of an MWE neccessarily can be interpreted as being idiomatic (cf. work on token-based analysis, e.g. Cook et al. (2008), Fritzinger et al. (2010)). However, in lack of respective hand-crafted data, a classification on token level, as in e.g. Diab and Bhutada (2009) is beyond the scope of the present paper.

2.2 Specificities of MWE Extraction for German

German has a relatively rich inflectional morphology, both in the nominal and verbal domain. Strong morpho-syntactic preferences in a word combination may thus indicate idiomatisation (i.e. MWE status). German also has a relatively free constituent order and, despite its morphological richness, substantial syncretism in nominal morphology (Evert et al., 2004); as a consequence, POS-pattern based approaches to MWE extraction tend to have low recall. As suggested by e.g. Seretan (2011), we thus use dependency parsing (Schiehlen, 2003) and extract MWE candidates from the parse output.

In this paper, we concentrate on the extraction of verb+PP collocations. Examples are *zur Sprache bringen* (lit.: "to bring to language", idiom.: to raise), cf. Example (1) above. We expect our results to be transferable to other MWE patterns, e.g. verb+direct object, verb+subject, adjective+noun, etc. Some of the candidates identified as verb+PP collocations may be part of larger patterns, such as *den Wind **aus** den **Segeln nehmen*** ("**to take** the wind **out of** so.'s **sails**").

3 Preprocessing: Feature Collection

In this section, we describe how all occurrences of the MWE candidates and their features are extracted. Later, in Section 4.2, we describe how the feature values of all occurrences of lexically identical MWE candidates are averaged and integrated into the classifier.

3.1 Candidate Extraction

As German allows for a flexible constituent order, the components of an MWE need not always occur adjacently[1]. Consider the following example sentence, where the verbal component of *im Raum stehen* (lit. "stand in the room", idiom. "to be dealt with") occurs 4 words to the left of the preposition and the noun:

(2)

Also Thus	**steht** stands	das the	Gerücht rumor	weiter still	**im** in the	**Raum** room
Thus	the rumor	is still	to be	dealt	with	

Thus the rumor is still to be dealt with

A deep syntactic analysis is thus required in order to reliably extract candidate triples, regardless of the actual constituent order or the distance of their component words. We use FSPAR (Schiehlen, 2003), a finite-state based dependency parser providing good lexical coverage and a full morpho-syntactic analysis (including POS, lemma, gender, number, case, compound splitting). Based on this annotation, the morpho-syntactic fixedness features (cf. Section 3.2) are extracted. While the dependency-parsed representation allows for the extraction of different syntactic patterns, we focus on preposition-noun-verb triples in the present paper.

3.2 Morpho-Syntactic Features

MWEs often exhibit a certain degree of fixedness with respect to morphological or syntacti-

[1] However, the words of semantically opaque MWEs mostly **do** occur adjacently and we use this adjacency as an additional fixedness feature, as described in Section 3.2 below.

	name	description	type
A	refl	verb having a reflexive pronoun	M
	n-adj	noun taking an adjectival modifier	S'
	det-fus	noun with fused prep+determiner	M
	neg	verb negated	S'
	vorf	expression occurring in the *vorfeld*	S
B	num	number of the noun	M
	det	determiner of the noun	M
C	adja	adjacency of component words	S

Table 1: Overview of morpho-syntactic features. M = morphological, S = syntactical, S' = syntactical in a broader sense.

cal variability (Sag et al., 2002). For example, the verb+PP constructions *hinter+Ohr+schreiben* (lit.: "behind+ear+write") has its idiomatic reading only if the number of the noun is plural (*Ohr**en***), the noun has a definite determiner (*die*) and the verb is reflexive (*sich*):
sich etw. hinter die Ohren schreiben
(idiom.: "to make sure to remember").

For German, such morpho-syntactic features can help to identify MWEs. A complete list of the features we use is given in Table 1.

Our feature set comprises morphological features (M), syntactic features (S), and features which are syntactically motivated in a broader sense (S'). We distinguish 3 different groups of morpho-syntactic features, depending on the possible values of the features: the first group (A) contains features for which we count their presence vs. absence regardless of the actual value. These features are represented as the ratio of the majority value to the total number of occurrences. For example, the *neg* feature indicates how often an expression occurs negated, but does not contain information about the type of negation (e.g. negation particle(s), verbal negation, negation of the noun).

In contrast, the values of the features of the second group (B) are summed up, i.e. we count how often the noun of a candidate expression occurred in *singular* vs. *plural* number or, in cases where the nouns take a determiner, how often it is a *definite* vs. *indefinite* or *quantifying* determiner.

Finally, the feature of the third group (C) indicates the adjacency of the expression's components: their sentence positions are summed and then divided by the position of the noun[2], with adjacent expressions (without any intervening words) scoring exactly 3. An adjacency score equal or close to 3 is regarded as indicator for an MWE.

While most features can be straightforwardly applied for many languages, the *fus*-feature (= preposition and determiner are melted into a fused form, e.g. *zur* = *zu+der* ="to the") is to be found in only a few languages. In comparison to Romance languages, where the fusion of certain preposition+article combinations is mandatory (de+le=du), the fusion of German articles and prepositions is optional in many cases and can thus be used as indicator for idiomaticity (strong preference for being fused or not fused). This is illustrated in Example (3), which is an invalid variation of the sentence given in (2): for the MWE *in+Raum+stehen*, the fusion of preposition and article is required. In contrast, for the regular combinations in Example (4), variation is possible.

(3)

*Also	steht	das	Gerücht	weiter	**in dem**	Raum
Thus	stands	the	rumor	still	in the	room

(4)

Im	Zimmer	steht	eine	Topfpflanze
In dem	Zimmer	steht	eine	Topfpflanze
In the	room	stands	a	potted plant

The *vorfeld*-feature applies to a syntactic characteristic of German only: there are different sentence structures of German (verb-initial vs. verb second vs. verb final sentences) and – contrary to PPs of fully compositional constructions – PPs of idiomatic MWEs only rarely occur in sentence initial (= *vorfeld*) position, even though gramatically possible, see Example (5).

(5)

?Im	Raum	steht	das	Gerücht	also	weiter
In the	room	stands	the	rumour	thus	still

3.3 Cross-Lingual Features

For our cross-lingual features, we adapt two metrics of (Villada Moirón and Tiedemann, 2006), both approximating the semantic opacity of MWEs using word alignment data: ***translational entropy*** (*te*) and the ***proportion of default alignments*** (*pda*). Both are measures of how "regular" and similar to non-idiomatic cases the translations of the respective lexical combinations are.

Translational entropy indicates the degree of variety of the candidate's translational equivalences. Regular combinations with a transparent or compositional semantics mostly have one or only very few different translations. In contrast,

[2]Example calculation of adjacency score: preposition at sentence position 5 + noun at 6 + verb at 7 = 18 / 6 (position of the noun) = 3.

semantically opaque MWEs show more different translations, i.e. much variation in equivalents: the lack of a respective (likely idiomatic) counterpart in the target language leads to translational variation which is recognisable in a broader variation of word alignments (and thus higher *te* scores). We use the following formula[3] to derive *translational entropy* scores from word alignments:

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) \log P(t|s)$$

where "T_s" is the set of all translation links from the source word "s" into different target words "t".

The *proportion of default alignments* indicates how often the words of a candidate expression have been translated literally. First, the four most frequent translational equivalences for each word of the corpus are collected (= default alignments). Then, the proportion of these default alignments among all alignments of a candiate expression is calculated (Villada Moirón and Tiedemann, 2006):

$$pda(S) = \frac{\sum_{s \in S} \sum_{d \in D_s} align\_freq(s,d)}{\sum_{s \in S} \sum_{t \in T_s} align\_freq(s,t)}$$

where "T_s" is the set of all translation links of "s" (the source word), "D_s" contains the word's default alignments and "$align\_freq(x, y)$" is the frequency of translation links from word x to word y in the context of the triple "S". Semantically opaque MWEs lead to low *pda* scores.

We calculate *te* and *pda* scores based on automatically generated word alignments using GIZA++ (Och and Ney, 2003) of the German section of Europarl to English, French and Swedish, following Fritzinger (2010) who showed that averaged scores based on alignments from several language pairs are more reliable than single language pair scores[4].

4 Experimental Setup

We use Conditional Random Fields (CRFs, Lafferty et al. (2001)) for the classification of verb+PP triples into MWEs vs. regular combinations[5] In this section we go into more detail about our data set, and we explain how features are extracted and transformed into a CRF-suitable format and how the classification accuracies of the CRFs are evaluated.

[3]Taken from Melamed (1997).

[4]This is in line with Lefever et al. (2013), who use data from several language pairs for WSD.

[5]Note however that we do not exploit the full potential of CRFs: for type-based MWE extraction, we do not take any sequential features into account.

group	interval	#all	#train	#dev	#test
high	>39	2,272	1,818	227	227
mid	18–39	2,367	1,893	237	237
low	4–17	2,124	1,700	212	212
		6,763	3,774	676	676
thereof MWEs:		862	697	75	90

Table 2: Distribution of data: 3 different frequency intervals, randomly extract 80% of each interval for training, 10% for development and 10% for testing.

4.1 Data

We start from a set of 10,276 preposition-noun-verb triple types, extracted from Europarl version 3 (Koehn, 2005). These are manually annotated as MWEs[6] (937) vs. regular combinations (9,339). From this data set, we use only triples that occur at least 4 times in order to get reliable fixedness scores. We consider this threshold necessary as low occurrence frequencies can lead to an inaccurate representation of morpho-syntactic preferences. Assume, for example, a regular combination with f=2 which randomly occurs with the same values of number, article, etc., even though variation is possible and to be expected in a larger set of occurrences. The cross-lingual features are based on word alignment which is a purely statistical method and thus to a certain extent inaccurate. As non-recurring alignment is used as indicator for idiomaticity, infrequent candidate triples do not provide a sufficient basis for reliable alignment statistics. For comparison, see Evert (2005) who proposes a threshold of f>=5. We set the threshold to f=4, which reduces the data to 6,763 triples, whereof 862 are MWEs and 5,901 are regular combinations.

The set of triples is divided into three different frequency intervals (high, mid and low-frequent) and of each interval, we randomly extract 80% for training, and 10% for developement and testing respectively, without allowing for overlap between these three sets, cf. Table 2 for details.

As can be seen, there are much more regular combinations than MWEs in each of the sets. In order to not work on a data set with an artificial distribution of MWEs vs. regular combinations, we decided to not balance the sets with regard to the number of MWEs they contain.

[6]Without considering different levels of opacity or fixedness.

triple	all	sg	pl	bkt.
an Ball bleiben	12	12 (100%)	0 (0%)	10
aus Auge verlieren	431	91 (21%)	340 (79%)	7
auf Gedanke bringen	7	4 (57%)	3 (43%)	5

Table 3: Example of how feature values (here: number feature) are grouped into suitable buckets (bkt.) for CRF training.

4.2 Features

Feature values are considered as strings in CRFs. In order to be able to abstract over the training data and predict idiomaticity on the (unseen) development and test sets, the features need to be represented in a suitable format.

For the morpho-syntactic fixedness features (except *adjacency*) given in Table 1 above, and for each triple type: we (1) add up the values of all occurrences of one triple, (2) take the percentage of the most frequent value and (3) pack that into buckets incrementing in 10% steps, rounding down to the next smallest bucket. For clarification, we give some calculation examples in Table 3.

The values of the adjacency feature are spread around 3.0; for each lexical triple, we sum and average the values of all occurrences, round them to one decimal and calculate the absolute value of their distance to 3.0 (using increments of 0.1), with low bucket scores indicating high fixedness[7].

Translational entropy values (of all three language pairs DE-EN, DE-FR, DE-SE) are summed and averaged for each distinct triple. Depending on the triple these range between 0.045 and 4.406, with higher scores indicating opaque semantics. They are packed into buckets of 0.5 increment.

Finally, the proportion of default alignment values range between 0 and 1. They are summed, averaged and packed into buckets of 0.1 increment.

In addition to these features, we also use the lexical form of the verb+PP triples themselves, because even those can be indicators for MWEs. This holds particularly for nouns, as can be seen from Table 4, where we give the most frequent nouns occuring in MWEs vs. regular combinations (both lists are derived from our training data, cf. Section 4.1).

[7]For example: averaged value is 2.78; rounded it is 2.8, distance to 3.0 is 0.2 (name of the bucket).

MWE	regular
Weg (way)	Jahr (year)
Hand (hand)	Bereich (range)
Auge (eye)	Rahmen (framework)
Tisch (table)	Bericht (report)
Leben (life)	Land (country)
Seite (side/page)	Herr (mister)

Table 4: Lists of most frequent nouns in verb+PP constructions, derived from the training data.

4.3 Evaluation

The accuracy of the different CRF classifiers is evaluated using *precision*, *recall* and *f-score*. These are calculated with regard to the number of valid MWEs from the 10% subsets of our manually annotated data found by the respective CRF.

$$Precision = \frac{\#correct-found}{\#all-found}$$

$$Recall = \frac{\#correct-found}{\#to-be-found}$$

$$F-Score = 2 * \frac{Precision*Recall}{Precision+Recall}$$

Note that a majority of 88% of the triples from the training data are regular combinations, while only roughly 12% are MWEs. To get an impression of the overall classification accuracy, we thus also calculate the percentage of all correct classifications (regardless if MWE or not).

5 Experiments

We use the developement set to experiment with different feature combinations and then, in a second series of experiments apply these combinations on the previously unseen test set.

We compare the results of our experiments to the following baselines[8]:

- **guess** use the frequency distribution of MWEs vs. regular triples derived from the training data (12% vs. 88%) to classify the data;
- **all triv** as there are many more regular triples than MWEs, classify everything as being regular;
- **pnv** use the constituent lemmas of the triples to train a CRF classifier.

[8]Note that only baseline nr.3 relies on a CRF for the classification; we used PERL scripts to realise baselines (1)+(2).

(a) Results on **development** set (75 MWEs to be found).

exp	TP	FP	FN	prec.	rec.	f-score
guess	8	69	67	10.39	10.67	10.53
pnv	21	15	54	58.33	28.00	37.84*
m-s.	26	19	49	57.78	34.67	43.33*
c-l.	27	11	48	71.05	36.00	47.79*
all	40	13	35	75.47	53.33	62.50***
best	42	12	33	77.78	56.00	65.12***

(b) Results on **test** set (90 MWEs to be found).

exp	TP	FP	FN	prec.	rec.	f-score
guess	11	73	79	13.10	12.22	12.64
pnv	25	8	65	75.76	27.78	40.65*
m-s.	41	14	49	74.54	45.56	56.55*
c-l.	47	11	43	81.03	52.22	63.51**
all	48	12	42	80.00	53.33	64.00**
best	46	10	44	82.14	51.11	63.01**

Table 5: F-scores with respect to the number of MWEs to be found. Statistical significance is calculated using chi-square, with * = significant at 0.001 level wrt. **guess**, ** = significant at 0.1 level wrt. **pnv**, *** = significant at 0.05 level wrt. **pnv**

We trained CRF classifiers for the following feature combinations (all include the preposition, noun and verb of the triples):

- **m-s** use all morpho-syntactic features: *refl, n-adj, det-fus, neg, vorf, num, det, adja*, cf. Table 1 above;
- **c-l** use the cross-lingual features *translational entropy* and *proportion of default alignments*;
- **all** use all morpho-syntactic features and all cross-lingual features;
- **best** use the "best" combination of features: by (1) adding each of the features independently to the current feature set, then (2) calculating the percentage of correct classifications on the development data and (3) permanently adding the best performing feature to the feature collection (4) repeat from (1) until performance drops. This lead to the following feature combination: *pnv+te+adja+fus+refl*;

We are aware that our "best" combination does not neccessarily represent the global maximum of all possible feature combinations. However, we believe it is a reasonable local maximum, given that the calculation of all combinations is too costly (in terms of both time and computing resources) to be realised.

(a) Results on **development** set
(676 triples, whereof 75 mwes / 601 literals).

exp.	# correct			% correct
	mwe	lit	all	
guess	8	532	540	79.89%
all triv	0	601	601	88.91%
pnv	21	586	607	89.79%
m-s	26	582	608	89.94%
c-l	27	590	617	91.27%
all	40	588	628	92.89%
best	42	589	631	93.34%

(b) Results on **test** set
(676 triples, whereof 90 mwes / 586 literals).

exp.	# correct			% correct
	mwe	lit	all	
guess	11	513	524	77.51%
all triv	0	586	586	86.69%
pnv	25	578	603	89.20%
m-s	41	572	613	90.68%
c-l	47	575	622	92.01%
all	48	574	622	92.01%
best	46	576	622	92.01%

Table 6: Percentage of correct classifications on the development and test sets .

6 Results

The accuracies of the different classification models in terms of f-scores (wrt. MWEs to be found) are given in Table 5, where "TP" (= true positives) designates valid MWEs identified by the classifier, "FP" (= false positives) are regular triples that were erroneously identified as MWEs and "FN" (= false negatives) are MWEs that should have been identified but were not found by the classifier.

The results in Table 5 show that any combination of features outperforms both baselines ("guess", "pnv"): while we can see a moderate improvement for the use of morpho-syntactic ("m-s") and cross-lingual ("c-l") features when used independently, the combinations of morpho-syntactic and cross-lingual features (cf. "all" and "best") lead to even higher f-scores with a statistically significant improvement with respect to the two baselines.

Similarly, the percentage of correct classification decisions in Table 6 shows that combining morpho-syntactic and cross-lingual features results in higher prediction accuracy.

The results in Table 6 and Table 5 confirms that features of different and independent dimensions (morpho-syntax vs. semantics) benefit from each other, and as a consequence, that their combination leads to an improved classification into MWEs and

approach	lang.	pattern?		classification			identification			
		yes	no	rank.	sup.	unsup.	frq.	m-s.	sem.	wa.
(Smadja, 1993)	EN	X		X			X			
(Bannard, 2007)	EN	X		X				X		
(Baldwin et al., 2003)	EN	X		X					X	
(Villada Moirón and Tiedemann, 2006)	NL	X		X						X
(Weller and Fritzinger, 2010)	DE	X		X				X		X
(Ramisch et al., 2010)	PT		X			X	X			X
(Fothergill and Baldwin, 2011)	JA		X		X			X	X	
(Tsvetkov and Wintner, 2011)	HE		X			X	X	X	X	X
present paper	DE	X			X			X		X

Table 7: Non-exhaustive overview of different approaches dealing with the identification of MWE types. The approaches are classified according to the following categories: *pattern?* (= is it restricted to MWEs of a certain syntactic pattern), *classification* (= ranking according to association measures or supervised, unsupervised classification), *identification* (= aspect of the MWE that is used for their identification: *frq.* = collocational behaviour, *m-s.* = morpho-syntactic features, *sem* = semantic features, *wa* = word alignment, i.e. translational behaviour.

regular triples.

Overall, the classification performance is similar on the development and test set, indicating that the built classifiers are robust and not over-fitting. However, the combination of all features seems to be more stable than the "best" combination obtained by searching for a local maximum on the development set.

7 Related Work

The task of automatically identifying MWEs has gained much attention in the NLP research community in the past. The approaches that emerged are just as multi-dimensional as the phenomenon of MWEs itself, each tackling one (or more) specific characteristics of MWEs. Consider Table 7 for a partial overview.

There are three of these characteristics that have been repeatedly implemented by different researchers to identify MWEs: i) word association measures, ii) morpho-syntactic fixedness, iii) semantic opacity.

Approaches based on word association measures exploit estimated vs. observed co-occurrence frequencies of an MWE's content words and the expression as a whole to identify valid MWEs (e.g. Church and Hanks (1990)). Such approaches proved to work well, but their performance can easily be enhanced by additionally checking for syntactic consistency of the MWEs. This can be realised either by restricting the candidate list to a certain syntactic MWE pattern beforehand (Evert and Krenn, 2001) or by filtering out syntactically inconsistent MWEs after having identified highly associated word pairs (Smadja, 1993).

Many types of MWEs exhibit a certain degree of morpho-syntactic fixedness: they do not allow for morphological or syntactic variation when used idiomatically. While some approaches investigate different types of syntactic variation (e.g. Bannard (2007) for English or Weller and Heid (2010) for German), others combine syntactic fixedness with limited lexical variability to identify MWEs (Fazly and Stevenson, 2006).

Finally, there are approaches tackling the opaque semantics of MWEs: they are based on the assumption that the semantics of the expression as a whole cannot be derived from the semantics of its constituent words. While Baldwin et al. (2003) use Latent Semantic Analysis for this task, Villada Moirón and Tiedemann (2006) present an approach that approximates the MWE's semantics by deriving translational equivalences from parallel text. While Villada Moirón and Tiedemann (2006) use word alignment only for ranking MWE candidates identified separately by means of syntactic patterns in parsed data, other approaches, e.g. Zarrieß and Kuhn (2009), de Caseli et al. (2010) use word alignment as basis for MWE extraction itself.

More recently, some approaches came up with combinations of features addressing different characteristics of MWEs. (Ramisch et al., 2010) combine word association measures with alignment-based approaches and use Bayesian Networks to predict MWEs, while (Weller and Fritzinger, 2010) combine morpho-syntactic fixedness with translational equivalences. In contrast, (Fothergill and Baldwin, 2011) combine morpho-syntactic fixedness with lexical hypernyms and (Tsvetkov

and Wintner, 2011) present a very feature-rich approach (using Bayesian Networks) that combines collocational behaviour with morpho-syntactic fixedness and translational equivalences[9].

Table 7 shows where our approach ranges in relation to the work just discussed. It is most comparable to the one of Weller and Fritzinger (2010). However, they use less morpho-syntactic features and their evaluation is restricted to different rankings (of 200 candiate triples) in order to find an optimal feature combination. While such rankings are useful to identify MWE candidates for lexicographical applications, the CRF models trained in the present paper allow for a more robust MWE identification that is easier to integrate into higher order applications.

To our knowledge, our approach allows to extract the most detailed morpho-syntactic data on MWEs for German, taking into account the rather intricate specificities of German morphology and syntax.

8 Conclusion and Future Work

We presented an approach for the identification of MWEs using morpho-syntactic fixedness (derived from deep syntactic analysis) and cross-lingual features (derived from automatic word alignment). We showed that combinations of these two feature sets, which both address different aspects of MWEs, clearly outperform the baselines, as well as the independent use of any of the feature sets in a supervised classification task.

Our approach could be applied prior to post-editing of SMT output, providing a comparatively accurate highlighting of MWEs (which are known to be potential sources of SMT errors).

In the future, we plan to investigate an even more fine-grained combination of features, e.g. in more linguistically motivated combinations. To give an example, the German verb+PP *in Gang kommen* means "to be set in motion" when the noun appears in singular form without a determiner, while the same lemmas used in plural form with a definite article *in die Gänge kommen*, bears the meaning "to get organised". Occurrences for which always the same two feature values are modified should have an additional impact.

As soon as manually annotated data on token-level become available, our approach can easily be trained on them. Moreover, we can then extend it to use sequential features, where appropriate.

So far, we focused on verb+PP constructions, but we plan to extend our approach to MWEs of different patterns in the future. Moreover, we intend to apply the CRF models we trained on the Europarl corpus to verb+PP constructions extracted from other domains.

[9]Note however that – in lack of available parallel corpora for Hebrew – they approximate the translational equivalences by combining dictionary entries with corpus lookups.

References

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment (ACL 2003)*, pages 89–96.

Bannard, Colin. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, pages 1–8.

Church, Kenneth Ward and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2008. The vnc-tokens dataset. In *Proceedings of the Workshop: Towards a shared task for multiword expressions (LREC 2008)*, pages 19–22.

de Caseli, Helena Medeiros, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Diab, Mona T. and Pravin Bhutada. 2009. Verb noun construction mwe token supervised classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (ACL-JICNLP 2009)*, pages 17–22.

Evert, Stefan and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 188–195.

Evert, Stefan, Ulrich Heid, and Kristina Spranger. 2004. Identifying morphosyntactic preferences in collocations. In *Proceedings of the 4th international conference on language resources and evaluation (LREC 2004)*, pages 907–910.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart, PhD dissertation.

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *Proceedings of the 11st Conference of the European Chapter of the ACL (EACL 2006)*, pages 337–344.

Fothergill, Richard and Timothy Baldwin. 2011. Fleshing it out: a supervised approach to mwe-token and mwe-type classification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 911–919.

Fritzinger, Fabienne, Marion Weller, and Ulrich Heid. 2010. A survey of idiomatic preposition-noun-verb triples on token level. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)*, pages 2908 –2914.

Fritzinger, Fabienne. 2010. Using parallel text for the extraction of german multiword expressions. *Lexis: E-Journal in English Lexicology*.

Koehn, Phillip. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit (MT Summit 2005)*, pages 79–86.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML 2001)*.

Lefever, Els, Veronique Hoste, and Martine De Cock. 2013. Five languages are better than one: An attempt to bypass the data acquisition bottleneck for wsd. In *Proceedings of the 14th international conference on intelligent text processing and computational linguistics (CICLing 2013)*, pages 343–354.

Melamed, I. Dan. 1997. Measuring semantic entropy. In *ACL-SIGLEX Workshop Tagging Text with Lexical Semantics: Why, What and How*, pages 41–46.

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

Ramisch, Carlos, Helena de Medeiros Caseli, Aline Villavicencio, André Machado, and Maria José Finatto. 2010. A hybrid approach for multiword expression identification. In *Computational Processing of the Portuguese Language*, pages 65–74. Springer.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2002)*, pages 1–15.

Schiehlen, Michael. 2003. A cascaded finite state parser for german. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL 2003)*.

Seretan, Violeta. 2011. *Syntax-Based Collocation Extraction*. Text, Speech and Language Technology. Springer.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(3):143–177.

Tsvetkov, Yulia and Shuly Wintner. 2011. Identification of multi-word expressions by combining multiple linguistic information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 836–845.

Villada Moirón, Begoña and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Workshop on Multiword-Expressions in a multilingual context (EACL 2006)*, pages 33–40.

Weller, Marion and Fabienne Fritzinger. 2010. A hybrid approach for the identification of multiword expressions. In *Online Proceedings of the SLTC 2010 Workshop on Compounds and Multiword Expressions*.

Weller, Marion and Ulrich Heid. 2010. Extraction of German multiword expressions from parsed corpora using context features. In *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)*.

Zarrieß, Sina and Jonas Kuhn. 2009. Exploiting translational correspondences for pattern-independent mwe identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30.

Multi-word processing in an ontology-based Cross-Language Information Retrieval model for specific domain collections

Maria Pia di Buono
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
mdibuono@unisa.it

Johanna Monti
UNISS
Via Roma 151
Sassari, Italy
jmonti@uniss.it

Mario Monteleone
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
mmonteleone@unisa.it

Federica Marano
UNISA
Via Giovanni Paolo II, 132
84084 Fisciano (SA)
fmarano@unisa.it

Abstract

This paper proposes a methodological approach to CLIR applications for the development of a system which improves multi-word processing when specific domain translation is required. The system is based on a multilingual ontology, which can improve both translation and retrieval accuracy and effectiveness. The proposed framework allows mapping data and metadata among language-specific ontologies in the Cultural Heritage (CH) domain. The accessibility of Cultural Heritage resources, as foreseen by recent important initiatives like the European Library and Europeana, is closely related to the development of environments which enable the management of multilingual complexity. Interoperability between multilingual systems can be achieved only by means of an accurate multi-word processing, which leads to a more effective information extraction and semantic search and an improved translation quality.

1 Introduction

Cross-language Information Retrieval (CLIR) applications aimed at accessing information on the web in several languages is attracting many important players in the Information Retrieval (IR) field, such as Google and Microsoft. Typically in CLIR applications, information is searched by means of a query expressed in the user's mother tongue. This query is automatically translated in the desired foreign language and the results are translated back in the user's mother tongue.

This process is based on two different translation stages: query translation and document translation. The query translation concerns the translation in the desired foreign language of the query expressed in the user's mother tongue, whereas the document translation is the back translation in the user's language of the relevant documents found by means of the translated query. Translation is usually based on bilingual or multilingual Machine Readable Dictionaries (MRD), Machine Translation (MT) and parallel corpora.

CLIR applications are often used in domain specific collections, such as the Europeana Connect, which is aimed at facilitating multilingual access to Europeana.eu, an internet portal that acts as an interface to millions of books, paintings, films, museum objects and archival records that have been digitized throughout Europe, regardless of the users' native language.

In Europeana Connect, indeed, users can submit queries in their native language and are able to retrieve documents in other languages and ob-

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.

tain information about objects from many sources across all European countries. The retrieved information is translated back into the user's language by means of MT.

Figure 1 shows a typical Europeana item description in English. The text contains several compound terms (highlighted in the text). Compound terms belong to multi-word units (MWU), which designate a wide gamut of lexical constructions, composed of two or more words with an opaque meaning, i.e. the meaning of a unit is not always the result of the sum of the meanings of the single words that are part of the unit.

MWUs are not always easy to identify since co-occurrence among the lexemes forming the units may vary a great deal. In domain specific texts compound terms, mainly noun compounds, are very frequent. In all languages there is indeed a close relationship between terminology and multi-words and, in particular, word compounds. In fact, word compounds account in some cases for 90% of the terms belonging to a domain specific language.

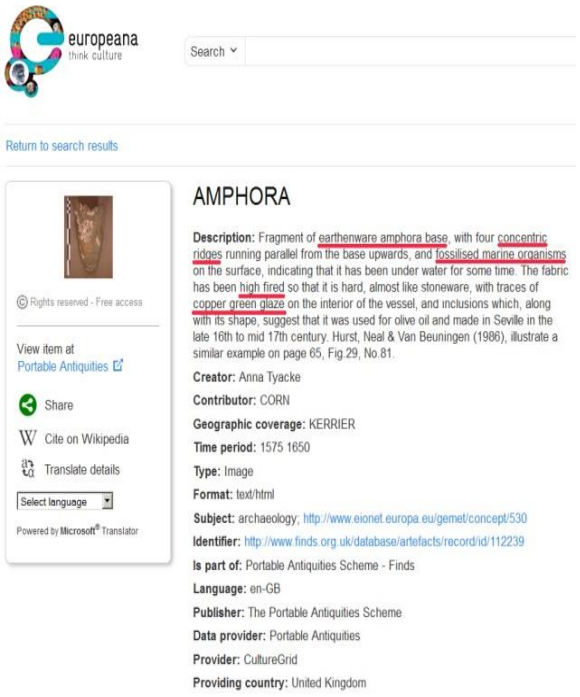

Figure 1: Europeana item description

CLIR success clearly depends on the quality of translation and therefore inaccurate or incorrect translations may cause serious problems in retrieving relevant information. A very frequent source of mistranslations in specific domain texts, as clearly emerges from the example in Figure 2, is, indeed, represented by MWUs, and in particular terminological word compounds.

Contrary to generic simple words, terminological word compounds are mono-referential, i.e. they are unambiguous and refer only to one specific concept in one special language, even if they may occur in more than one domain. Their meaning, similar to all compound words, cannot be directly inferred by a non-expert from the different elements of the compounds because it depends on the specific area and the concept it refers to.

Figure 2 is the result of the automatic translation into Italian of the item description in Figure 1. Almost all MWU translations powered by Microsoft Translator, the MT system used in Europeana, are wrong, such as *earthenware amphora base* translated with **anfora di terracotta base* instead of *piede di anfora in terracotta* or *high fired* translated with **alto sparato* instead of *cotta ad alte temperature.*

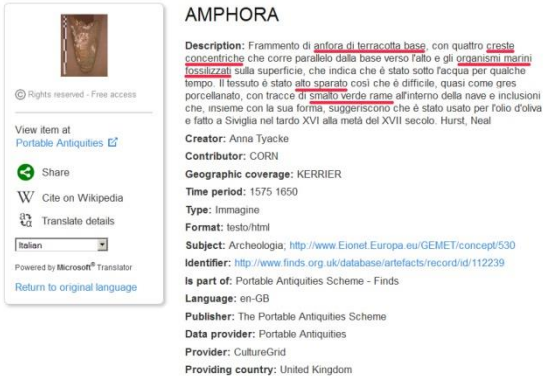

Figure 2: Europeana item description translated by Microsoft Machine Translation

Processing and translating these different types of compound words is not an easy task since their morpho-syntactic and semantic behavior is quite complex and varied according to the different types and their translations are practically unpredictable.

The main contribution of this paper is the experimentation of an ontology-based CLIR system designed to overcome the current limitations of the state-of the-art CLIR, in specific domain collections, and in particular to take into account a proper processing and translation of MWUs. This experiment has been set up for the Italian/English

language pair and can be easily extended to other language pairs.

The remaining of this paper is organized as follows. The next section briefly explains the related work in the area of CLIR. Section 3 describes the methodology used in the experiment. Then, section 4 is devoted to system overview, and, in particular, presents the data modeling and the system architecture extension. Section 5 introduces the feasibility study together with the description of the electronic dictionaries, the semantic annotation and the translation process. Finally, conclusions and future work are described in section 6.

2 Related work

Approaches to CLIR are either based on bilingual or multilingual Machine Readable Dictionaries (MRD), Machine Translation (MT), parallel corpora and finally ontologies.

Hull & Greffenstette (1996), Oard & Dorr (1996), Pirkola (1998) and more recently Oard (2009) provide comprehensive descriptions of these approaches.

Both MRD-based and MT-based CLIR are the prevalent models but they show several weaknesses especially with regard to domain-specific contexts because they are not able to solve translation problems associated to MWUs, a very frequent and productive linguistic phenomenon in languages for special purposes (LSPs). Both approaches in most cases produce literal translations of the single constituents of MWUs which do not represent appropriate translation solutions for this type of lexical constructions. MWUs, in fact, have to be considered as single meaning units. For instance, the Italian translation of the compound adjective "*high fired"* is *cotto ad alte temperature* which cannot be obtained by the literal translation of the single constituents of this MWU.

Translation errors mainly depend on lack of coverage and quality of the systems and various techniques have been proposed to reduce the errors due to the presence of MWU used during query translation. Among these techniques, phrasal translation, co-occurrence analysis, and query expansion are the most popular ones.

Concerning phrasal translation, techniques are often used to identify multi-word concepts in the query and translate them as phrases. Hull & Grefenstette (1996) showed that the performance achieved by manually translating phrases in queries is significantly better than that of a word-by-word translation using a dictionary. Davis and Ogden (1997) used a phrase dictionary extracted from parallel sentences in French and English to improve the performance of CLIR. Ballesteros and Croft (1996) performed phrase translation using information on phrase and word usage contained in Collins MRD. More recently, Gao et al. (2001) propose that noun phrases are recognized and translated as a whole by using statistical models and phrase translation patterns and that the best word translations are selected based on the cohesion of the translation words. Finally, Saralegi & de Lacalle (2010) use a simple matching and translation technique based on a bilingual MWU list to detect and translate them.

Co-occurrence statistics is used to identify the best translation(s) among all translation candidates using text collections in the target language as a language model, assuming that correct translations occur more frequently than wrong ones (Maeda et al., 2000; Ballesteros and Croft, 1998; Gao et al., 2001; Sadat et al., 2001).

As for query expansion techniques, Ballesteros & Croft (1996 and 1997) assume that additional terms that are related to the primary concepts in the query are likely to be relevant and that phrases in query expansion via local context analysis and local feedback can be used to reduce the error associated with automatic dictionary translation.

Concerning MT-based CLIR, MWU identification and translation problems are far from being solved. MWU processing and translation in SMT started being addressed only very recently and different solutions have been proposed so far, but basically they are considered either as a problem of automatically learning and integrating translations, of word alignment or word sense disambiguation (WSD) (Monti, 2013).

Current approaches to MWU processing move towards the integration of phrase-based models with linguistic knowledge and scholars are starting to use linguistic resources (LRs), either handcrafted dictionaries and grammars or data-driven ones, in order to identify and process MWUs as single units.

A first possible solution is the incorporation of MRDs and glossaries into the SMT system, for which there are several straightforward approaches. One is to introduce the lexicon as phrases in the phrase-based table. Unfortunately, the words coming from the dictionary have no

context information. A similar approach is to introduce them to substitute the unknown words in the translation, but this poses the same problem as before.

Another solution for overcoming translation problems in MT and in SMT in particular is based on the idea that MWUs should be identified and bilingual MWUs should be grouped prior to statistical alignment (Lambert and Banchs, 2006). In their work, bilingual MWU were grouped as one unique token before training alignment models.

More recently, Ren et al. (2009) have underlined that experiments show that the integration of bilingual domain MWUs in SMT could significantly improve translation performance. Wu et al. (2008) propose the construction of phrase tables using a manually-made translation dictionary in order to improve SMT performance. Finally, Bouamor et al. (2011) affirm that integration of contiguous MWUs and their translations improves SMT quality and propose a hybrid approach for extracting contiguous MWUs and their translations in a parallel corpus.

Other solutions try to integrate syntactic and semantic structures (Chiang, 2005; Marcu et al., 2006; Zollmann & Venugopal, 2006), but the solutions undoubtedly vary according to the different degrees of compositionality of the MWU.

Very recently, identification and disambiguation of MWUs are being considered as a problem of Word Sense Disambiguation (WSD), i.e. the identification and the selection of the proper meaning of a word in a given context when it has multiple meanings, and several approaches to integrate WSD in SMT have been proposed (Carpuat & Wu, 2007; Carpuat & Diab, 2010 among others).

The problem is here to select the most appropriate translation in TL to a given lexical unit in the SL. Some scholars refer to this problem also as word translation disambiguation (WTD), such as for instance Yang and Kirchoff (2012).

Ontologies are also used in CLIR and are considered by several scholars a promising research area to improve the effectiveness of Information Extraction (IE) techniques particularly for technical-domain queries. Volk et al. (2003) use ontologies as interlingua in cross-language information retrieval in the medical domain and show that the semantic annotation outperforms machine translation of the queries, but the best results are achieved by combining a similarity thesaurus with the semantic codes. Yapomo et al. (2012) perform ontology-based query expansion of the most relevant terms exploiting the synonymy relation in WordNet.

3 Methodology

Our approach to CLIR is based on Lexicon-Grammar (LG) devised by the French linguist Maurice Gross during the ‘60s (Gross, 1968, 1975 and 1989).

LG presupposes that linguistic formal descriptions should be based on the examination of the lexicon and the combinatory behaviors of its elements, encompassing in this way both syntax and lexicon. Nowadays, the LG methodology is being adopted by a wide research community both for Indo-European languages (French, Italian, Portuguese, Spanish, English, German, Norwegian, Polish, Czech, Russian, Bulgarian and Greek) and other ones (Arabic, Korean, Malagasy, Chinese, Thai...).

LG linguistic framework is based on the analysis of the so-called “simple sentence” [1], the smallest linguistic meaning context, by applying rules of co-occurrence and selection restriction.

LG scholars have been studying MWUs for years now and LG research in this field is indebted to the transformational and distributional concepts developed by Harris (1957, 1964 and 1982).

Thanks to these abovementioned research studies, LG range of analysis concerns lexicon, and especially the concept of MWU as “meaning unit”, “lexical unit” and “word group”, for which LG identifies four different combinatorial behaviors (De Bueriis and Elia, 2008).

Linguistic resources (LRs) developed according to the LG framework are used in Natural Language Processing (NLP) applications and are useful to achieve effective Information Retrieval (IR) systems (Marano F., 2012) and translation processes.

In the field of CLIR, the LRs developed according to the LG methodology can be used to overcome the shortcomings of statistical approaches to MT such as in *Google Translate* or

[1] In LG, a simple sentence is a context formed by a unique predicative element (a verb, but also a name or an adjective) and all the necessary arguments selected by the predicate in order to obtain an acceptable and grammatical sentence. For a detailed definition of simple sentence refer to Gross (1968).

Bing by Microsoft concerning MWU processing in queries, where the lack of context represent a serious obstacle to disambiguation. The same resources can also be used for domain-adaptation purposes in SMT, thus improving the translation quality in the document translation phase in specific domain contexts.

The main linguistic resources developed by LG researchers concerning MWUs are (i) matrix tables describing the syntactic-semantic properties of lexical entries, (ii) morphologically and semantically tagged electronic dictionaries, (iii) local grammars in the form of Finite State Automata (FSA) [2] and Finite State Transducers (FST)[3].

3.1 LG Methodology to Assess the Translation Quality

The quality of translations is guaranteed, from the beginning, by developing highly formalized LRs according to morphological, syntactical and semantic criteria. Often using smart translation technologies involves the deterioration of Translation Quality (TQ). In LG methodology, instead, we take advantage of well-formed LRs to keep a high level of TQ, since from the beginning, we use a supervised approach carried out by highly skilled linguists during the proper setting of the resources.

Assessing the quality of resources before they are translated prevents from subsequent checks on translated resources, though evaluation *ex post* of TQ results is necessary in any case.

According to LG a valid evaluation methodology should be based on a hybrid approach that encompasses both human and automatic evaluation.

The process is composed of two cycles. The first cycle can be outlined as follows (i) a query expressed in a Source Language (SL) is the input of the CLIR application, (ii) the CLIR system produces sample queries (i.e. sample texts) in the Target Language (TL), (iii) the resulting translated queries are examined by humans (Linguists, Translators, Terminologists/Domain Experts) to evaluate their quality. The human judgments are based on common criteria of TQ – i.e. adequacy and fluency – and are expressed using a Likert scale with scores 1-5 (for instance using the following judgments: 1. Strongly disagree, 2. Disagree, 3. Neither agree nor disagree, 4. Agree, 5. Strongly agree), (iv) only texts which obtain scores 4-5 become "validated" and "supervised" texts which represent the gold standard, (v) this gold standard is the training set for the Automatic Evaluation process, that can be carried out using METEOR[4] and GTM[5], the most suitable methods according to our opinion, as well as other ones[6].

During the second cycle, human evaluation is skipped and the SL queries are directly used as input for automatic evaluation.

It is necessary to periodically repeat the first cycle in order to enrich the training set and to increase the quality cycle.

4 System overview

We propose an architecture, which, when applied to a given language, maps data and metadata exploiting the morpho-syntactic and semantic information stored both in electronic dictionaries and FSA/FSTs (presented in 5.2 and 5.3). Furthermore, this architecture can also map linguistic tags (i.e. POS) and structures (i.e. sentences, MWU) to domain concepts.

The first step performed by our system is a linguistic pre-processing phase which formalizes (i.e. converts) natural language strings into reusable linguistic resources. During this first phase we also extract information from free-form user queries, and match this information with already available ontological domain conceptualizations. As described in Fig. 3, prior to the execution of a query against a knowledge base it is necessary to apply the Translation and the Transformation routines. We can see that the system is based on two workflows which are carried out simultaneously but independently.

[2] Finite-State Automata (FSA) are a special case of Finite-State Transducers that do not produce any result (i.e. they have no output). Typically, FSA are used to locate morpho-syntactic patterns in corpora and extract the matching sequences to build indices, concordances, etc.

[3] Finite-State Transducers (FSTs) are graphs that represent a set of text sequences and then associate each recognized sequence with an analysis result. The text sequences are described in the input part of the FST; the corresponding results are described in the output part of the FST. Typically, a syntactic FST represents word sequences and then produces linguistic information (its phrasal structure, for example).

[4] http://www.cs.cmu.edu/~alavie/METEOR/.

[5] http://nlp.cs.nyu.edu/GTM/.

[6] BLEU and NIST (based only on precision measure), F-Measure (based also on recall).

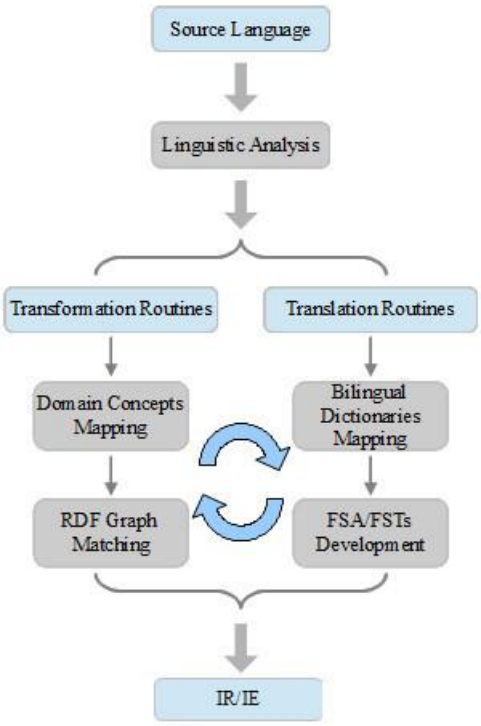

Figure 3: System workflow

The benefits of keeping separate these two workflows are (i) the development of an architecture with a central multilingual formalization of the lexicon, in which there is no specific target language, but each language can be at the same time target and source language, (ii) the development of extraction ontologies and SPARQL/SERQL adaptation systems which could represent a standard not only for our multilingual electronic dictionaries, but also for any lexical and/or language data-base for which translation is required.

With this dual-structure system, it is easier to successfully achieve the CLIR process since the results are given explicitly in the target language chosen by the user and the translation process is separated from the matching with the RDF triples.

5 Feasibility study

To test the feasibility of our architecture, we are carrying out a translation experiment from Italian into English, using all ontological and semantic constraints defined for the Italian model.

We have chosen the Archaeological domain to test the applicability of our approach. This choice allows us to demonstrate that the modularity of our architecture may be applied to a domain which is variable by type and properties and is semantically interlinked with other domains.
In the next paragraphs, we will present the LRs developed for our study, together with the description of the semantic annotation and the translation routines used in query translation.

5.1 Electronic dictionaries

An electronic dictionary is a lexical database homogeneously structured, in which the morphologic and grammatical characteristics of lexical entries (gender, number and inflection) are formalized by means of distinctive and non-ambiguous alphanumeric tags (Vietri et al. 2004).

All the electronic dictionaries, developed according to the LG descriptive method, form the DELA[7] system, which is used as the linguistic knowledge base in NLP applications. DELA electronic dictionaries are of two types: (i) simple word dictionaries, which include semantically autonomous lexical units formed by character sequences, delimited by blanks, such as *home* and *chair*, (ii) compound word dictionaries, which include lexical units composed of two or more simple words with a non-compositional meaning, such as *nursing home* and *rocking chair*. Terminological compound words (the most common obstacle in CLIR applications) are lemmatized in compound word electronic dictionaries[8].

The following example represents an excerpt from the Italian/English compound word dictionary of Archaeological Artefacts:

anfora di terracotta, N + NPN + FLX=C41 + DOM=RA1 + EN=earthenware amphora, N+AN+FLX=EC3
cerchi concentrici, N + NA + FLX=C601 + DOM=RA1 + EN=concentric ridges, N+AN+FLX=EC4

[7] Dictionnaire Électronique of LADL (Laboratoire d'Automatique Documentaire et Linguistique).

[8] Our domain dictionaries cover about 180 different semantic tags. The most important dictionaries are those of Informatics (54,000 entries ca.), Medicine (46,000 entries ca.), Law (21,000 entries) and Engineering (19,000 entries ca.). Subset tags are also foreseen for those domains that include specific subsectors. This is the case of Archaeological Artefacts dictionary (9,200 entries ca.), for which a generic tag RA1 is used, while more explicit tags are used for object type, subject, primary material, method of manufacture, object description.

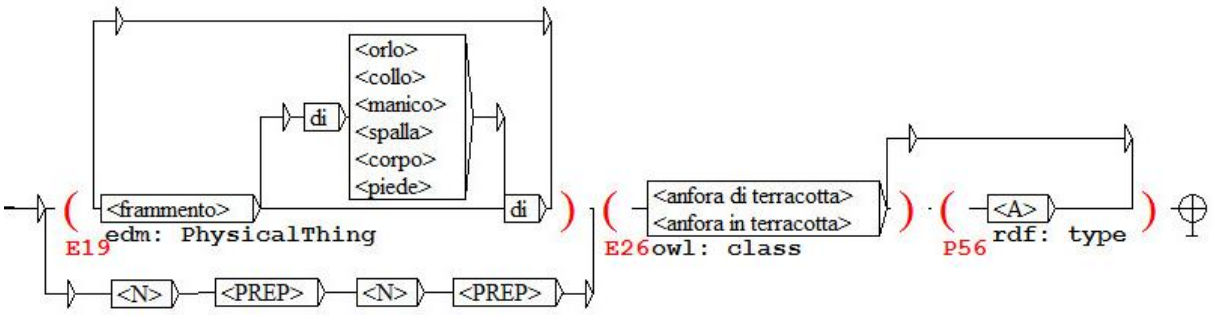

Figure 4: Use of FSA variables for identifying classes for subject, predicate and object

cottura ad alte temperature, N + NPAN + FLX =C611 + DOM=RA1 + EN=high fired, N+AN+FLX=EC4
fregio dorico, N + NA + FLX = C523 + DOM=RA1 + EN=doric frieze, N+AN+FLX=EC3
fusto a spirale, N + NPN + FLX = C7 + DOM=RA1 + EN=spiral stem, N+AN+FLX=EC3
fossile marino, N + NA + FLX = C501 + DOM=RA1 + EN=fossilised marine organism, N+AN+FLX=EC3
smalto verde rame, N + NAN + FLX=C04 + DOM=RA1 + EN=copper green glaze, N+AN+FLX=EC4

The compound words belong to the «Archaeological Artifacts» domain, marked with the domain tag «DOM=RA1» in the dictionary.

For each entry, a formal and morphological description is also given with (i) the internal structure of each compound, such as in the compound word *fregio dorico,* where the tag «NA» specifies that it is formed by a Noun, followed by an Adjective. (ii) the inflectional class, such as the tag «+FLX=C523», which indicates the gender and the number of the compound *fregio dorico,* together with its plural form, i.e. that *fregio dorico* is masculine singular, does not have any feminine corresponding form, and its plural form is *fregi dorici*. Each inflection class is associated to a local grammar which produces all the inflected forms of the compound words according to the inflection class associated to them.

Together with electronic dictionaries, local grammars are used in NLP routines to parse texts. Local grammars are useful to cope with specific characteristics of natural language; more appropriately, local grammars design is based on syntactic descriptions, which encompasses both transformational rules and distributional behaviours (Harris, 1957). Local grammars are developed in the form of FSA/FST (Silberztein, 1993 and 2002)[9].

5.2 Semantic annotation

As for ontologies, the formal definition we rely upon is the one given by the International Council of Museums - Conseil International des Musèes (ICOM – CIDOC) Conceptual Reference Model (CRM), which states that "a formal ontology (is) intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information" (Crofts et al., 2008).

CIDOC CRM is a core ontology composed of 90 classes (which includes subclasses and superclasses) and 148 unique properties (and subproperties). The object-oriented semantic model and its terminology are compatible with the Resource Description Framework (RDF). This ontology is constantly developed and updated.

We use FSA variables for identifying ontological classes and properties for subject, object and predicate within RDF graphs, as presented in Figure 4. FSA are based on LRs, which are used during the analysis of corpora to retrieve recursive phrase structures, in which combinatorial behaviours and co-occurrence between words identify properties, also denoting a relationship. Furthermore, electronic dictionaries include all inflected verb forms allowing to process queries

9 To develop and test electronic dictionaries and local grammars we use the NooJ software, an NLP environment, based on the DELA system of electronic dictionaries, on LG syntactic tables and on FSA/FST, developed in the form of graphs and used in LG to parse texts.

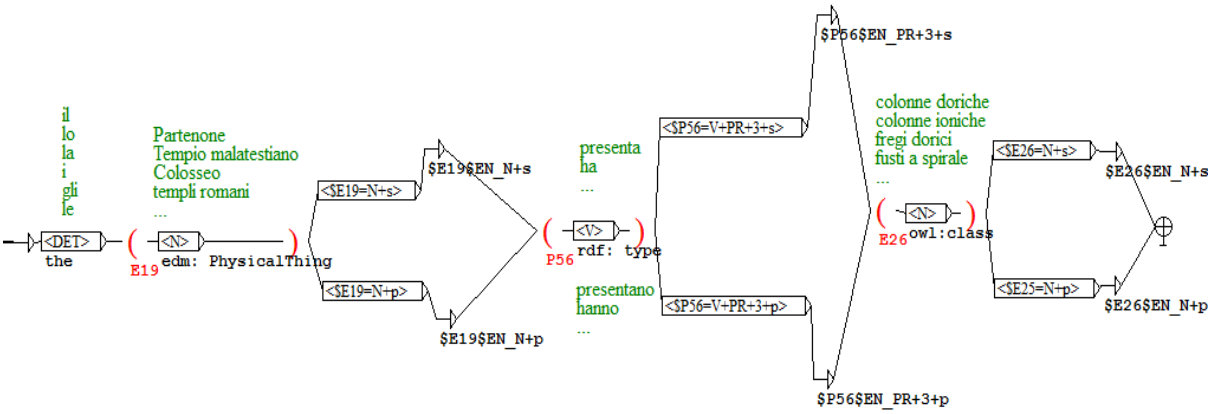

Figure 5: Example of a translation FST

expressed also with passive and more generally non-declarative sentences.

This matching of linguistic data to RDF triples and their translation into SPARQL/SERQL path expressions allows the use of specific meaning units to process natural language queries.

Figure 4 is a sample of an automaton which recognizes the following MWU:

> *frammento di (*Empty + *orlo* + *collo* + *manico* + *spalla* + *corpo* + *piede) (*Empty + *di) (anfora di terracotta* + *anfora in terracotta)* (Empty + any adjective)

According to our approach, electronic dictionaries entries (simple words and MWUs) are the subject and the object of the RDF triple.

In Figure 4 we also use FSA variables which apply to the sentence the following CIDOC-CRM classes and property: (i) E19 indicates "Physical Object" class; (ii) P56 stands for "Bears Feature" property; (iii) E26 indicates "Physical Feature" class.

Together with FSA variables we also associate POS to the Europeana Semantic Elements (ESE) metadata format[10], currently used in Europeana, i.e. edm: PhysicalThing, owl: class, rdf: type.

Furthermore, the automaton, built using lexical classes (Fig. 4), recognizes all instances included in E19 and and E26 classes, the property of which is P56, and not only the original MWUs.

5.3 Query translation

In our model, the Translation Routines are applied independently of the mapping process of the pivot language. This allows us to preserve the semantic representation in both languages.

Indeed, identifying semantics through FSA guarantees the detection of all data and metadata expressed in any different language.

Figure 5 shows an FST in which a translation process from Italian to English is performed on the basis of a dictionary look-up, a morpho-syntactic and semantic analysis. This translation FST, in fact, identifies and annotates the different linguistic elements of declarative sentences such as "Il Partenone presenta fregi dorici", "I templi romani hanno fusti a spirale", etc., with their morpho-syntactic and semantic information and performs automatic translations on the basis of an LG bilingual dictionary.

For instance, if a grammar variable, say $E26, holds the value "fusti a spirale", the output $E26$EN will produce the correct translation "spiral stems", on the basis of the value associated to the +EN feature in the bilingual entry " fusto a spirale, N+NPN+FLX=C7+DOM = RA1EDEAES+EN= spiral stem,N+AN+FLX= EC3" and the morpho-syntactic analysis performed by the graph in Figure 5, which identifies and produces the plural form of the compound noun "fusto a spirale".

6 Conclusions and future work

The proposed architecture ensures not only the coverage of a large knowledge portion but preserves deep semantic relations among different languages.

Future work aims at implementing our Linguistic Resources to test the accuracy of cross-

[10] http://pro.europeana.eu/edm-documentation

language information retrieval, extraction and semantic search.

Note

Maria Pia di Buono is author of sections 4, 5 and 5.2, Johanna Monti is author of sections 1, 2 and 5.3, Mario Monteleone is author of sections 5.1 and 6 and Federica Marano is author of section 3 and 3.1.

References

Ballesteros L. and Croft B. 1996. *Dictionary Methods for Cross-Lingual Information Retrieva*l. Proc. of the 7th DEXA Conference on Database and Expert Systems Applications, Zurich, Switzerland, September 1996: 791-801.

Ballesteros L. and Croft B. 1997. *Phrasal translation and query expansion techniques for crosslanguage information retrieval*. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval.

Ballesteros L. and Croft B. 1998. *Resolving Ambiguity for Cross-language Retrieval*. SIGIR'98, Melbourne, Australia, August 1998: 64-71.

Bouamor D., Semmar N., and Zweigenbaum, P. 2011. *Improved statistical machine translation using multi-word expressions.* Proceedings of MT-LIHMT. Barcelona, Spain.

Carpuat M. and Diab M. 2010. *Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation*. HLT-NAACL 2010.

Carpuat M. and Wu D. 2007. *Improving statistical machine translation using word sense disambiguation.* Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL): 61-72.

Chiang, D. 2005. *A hierarchical phrase-based model for statistical machine translation.* Proceedings of Association of Computational Linguistics (ACL).

Crofts N., Doerr M., Gill T., Stead S., Stiff M. (eds.). 2008. *Definition of the CIDOC Conceptual Reference Model, Version 5.0*.

Davis M. W., and Ogden W. C. 1997. *Free resources and advanced alignment for cross-language text retrieval.* The Sixth Text Retrieval Conference (TREC-6). NIST, Gaithersbury, MD.

De Bueris G., Elia, A. (eds.). 2008. *Lessici elettronici e descrizioni lessicali, sintattiche, morfologiche ed ortografiche*. Plectica, Salerno.

Gao J., Nie J., Xun E., Zhang J., Zhou M., Huang C. 2001. *Improving Query Translation for Cross-Language Information Retrieval using Statistical Models.* Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM.

Gross M. 1968. *Grammaire transformationnelle du français. – I – Syntaxe du verbe*, Larousse, Paris.

Gross M. 1975. *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris.

Gross M. 1989. *La construction de dictionnaires électroniques*. Annales des Télécommunications, vol. 44, n° 1-2: 4-19, CENT, Issy-les-Moulineaux/Lannion.

Harris Z.S. 1957. *Co-occurrence and transformation in linguistic structure*. Language 33: 293-340.

Harris Z.S. 1964. *Transformations in Linguistic Structure.* Proceedings of the American Philosophical Society 108:5:418-122.

Harris Z.S. 1982. *A Grammar of English on Mathematical Principles*. John Wiley and Sons, New York, USA.

Hull D. A. and Grefenstette G. 1996. *Querying across languages: a dictionary-based approach to multilingual information retrieval,* Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval: 49-57.

Lambert P. and Banchs R. 2006. *Grouping multi-word expressions according to Part-Of-Speech in statistical machine translation*. Proceedings of the EACL Workshop on Multi-word expressions in a multilingual context. Trento, Italy.

Maeda, A., Sadat, F., et al. 2000. *Query Term Disambiguation for Web Cross-Language Information Retrieval using a Search Engine*. Proc. of the Fifth Int'l Workshop on Info. Retrieval with Asian Languages, Hong Kong, China: 173-179.

Marano F. 2012. *Exploring Formal Models of Linguistic Data Structuring. Enhanced Solutions for Knowledge Management Systems Based on NLP Applications.* PhD Dissertation, University of Salerno, Italy.

Marcu D., Wei W., Echihabi A., and Knight K. 2006. *SPMT: Statistical Machine Translation with Syntactified Target Language Phrases*. Proceedings of Empirical Methods in Natural Language Processing (EMNLP).

Monti, J. 2013. *Multi-word unit processing in Machine Translation: developing and using language resources for multi-word unit processing in Ma-*

chine Translation. PhD dissertation. University of Salerno, Italy.

Oard D. W. 2009. *Multilingual Information Access*. Encyclopedia of Library and Information Sciences, 3rd Ed., edited by Marcia J. Bates, Editor, and Mary Niles Maack, Associate Editor, Taylor & Francis.

Oard, D. W. and Dorr, B. J. 1996. *A survey of multilingual text retrieval*. Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies.

Pirkola A. 1998. *The Effects of Query Structure and Dictionary Setups in Dictionary-Based Cross-language Information Retrieval*. In Croft, W.., et al., 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008), Melbourne, Australia, August 24-28:55-63.

Ren, Z. Lü, Y., Cao J., Liu Q., and Zhixiang Y. 2009. *Improving statistical machine translation using domain bilingual multiword expressions*. Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, Singapore : 47-54.

Sadat F., Maeda A., Yoshikawa M. and Uemura S. 2001. *Query expansion techniques for the CLEF bilingual track*. Working Notes for the CLEF 2001 Workshop: 99-104.

Saralegi X. and de Lacalle M. L. 2010. *Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR*. Proceedings of the 7th International Conference on Language Resources and Evaluations (LREC). Malta.

Silberztein M. 1993. *Dictionnaires électroniques et analyse automatique de textes*, Masson, Paris.

Silberztein M. 2002. *NooJ Manual*. Available for download at: www.nooj4nlp.net.

Szpektor I., Dagan I., Lavie A., Shacham D., Wintner S. 2007. *Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation*. Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data, Prague, Czech Republic.

Vietri S., Elia A. and D'Agostino E. 2004. *Lexicon-grammar, Electronic Dictionaries and Local Grammars in Italian*, Laporte, E., Leclère, C., Piot, M., Silberztein M. (eds.), Syntaxe, Lexique et Lexique-Grammaire. Volume dédié à Maurice Gross, Lingvisticae Investigationes Supplementa 24, John Benjamins, Amsterdam/Philadelphia.

Volk M., Vintar S., and Buitelaar P. 2003. *Ontologies in cross-language information retrieval.* Proceedings of WOW2003 (Workshop Ontologie-basiertes Wissensmanagement), Luzern, Switzerland.

Vossen P., Soroa A., Zapirain B. and Rigau G. 2012. *Cross-lingual event-mining using wordnet as a shared knowledge interface.* Proceedings of the 6th Global Wordnet Conference, C. Fellbaum, P. Vossen (Eds.), Publ. Tribun EU, Brno, Matsue, Japan, January 9-13:382-390.

Wu, H., Wang, H., & Zong, C. 2008. *Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora.* Proceedings of Conference on Computational Linguistics (COLING): 993-100.

Yang M. and Kirchhoff K. 2012.*Unsupervised Translation Disambiguation for Cross-Domain Statistical Machine Translation*. Proceedings of AMTA.

Yapomo M., Corpas G. and Mitkov R. 2012. *CLIR- and ontology-based approach for bilingual extraction of comparable documents*. The 5th Workshop on Building and Using Comparable Corpora.

Zollmann A., and Venugopal A. 2006. *Syntax augmented machine translation via chart parsing*. In Proceedings of the Workshop on Statistical Machine Translation, HLT/NAACL.

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*).

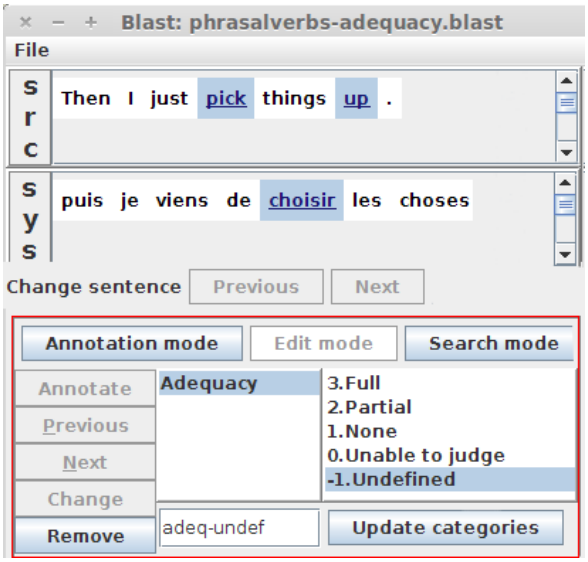

Figure 1: Annotation interface using BLAST.

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]

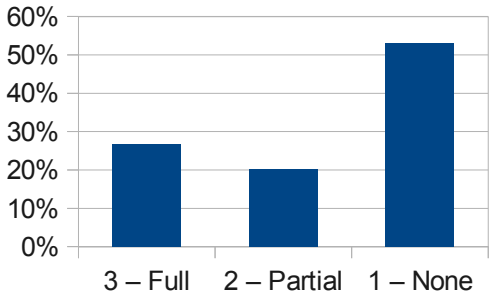

Figure 2: Proportion of translations judged as FULL, PARTIAL and NONE for adequacy.

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.

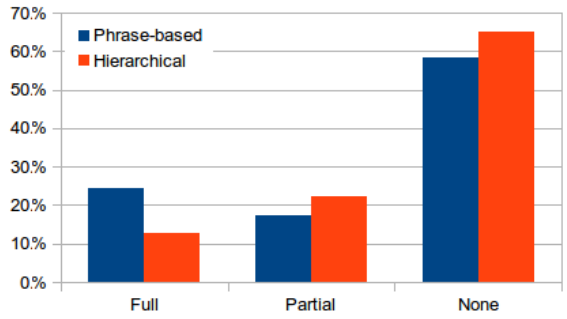

Figure 3: Proportion of different translations judged as FULL, PARTIAL and NONE for adequacy.

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.

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/

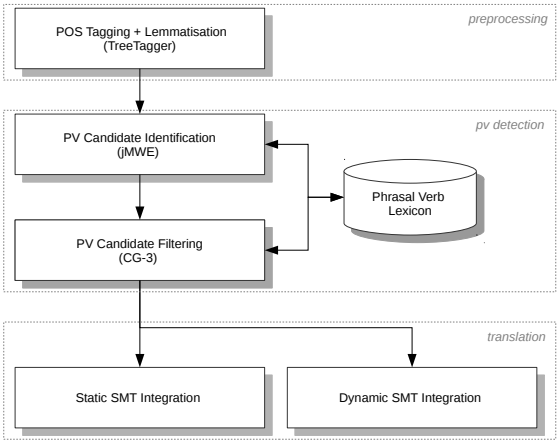

Figure 1: Pipeline of the experiment including phrasal verb detection and integration into the English part of the parallel corpora.

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.