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.