Taking on new challenges in multi-word unit processing for machine translation

Johanna Monti[1] jmont@tin.it

Anabela Barreiro[2] barreiro_anabela@hotmail.com

Annibale Elia[3] elia@unisa.it

Federica Marano[4] federicamarano@gmail.com

Antonella Napoli[5] antonella.napoli@ymail.com

[1, 3, 4, 5] Università degli Studi di Salerno, Via Ponte Don Melillo 84084, Fisciano, Italy
[2] Universidade do Porto, Via Panorâmica, s/n, 4150 – 564 Porto, Portugal

Abstract

This paper discusses the qualitative comparative evaluation performed on the results of two machine translation systems with different approaches to the processing of multi-word units. It proposes a solution for overcoming the difficulties multi-word units present to machine translation by adopting a methodology that combines the lexicon grammar approach with OpenLogos ontology and semantico-syntactic rules. The paper also discusses the importance of a qualitative evaluation metrics to correctly evaluate the performance of machine translation engines with regards to multi-word units.

1 Introduction

Recently, the availability and use of large parallel corpora, the development of knowledge bases, the adoption of statistical models, and the integration with various computer assisted translation tools has contributed to a significant progress in the machine translation field. However, lexical problems still represent a critical area in machine translation, and among these, multi-word units are particularly difficult to be processed by the different systems.

The aim of this paper is to analyze the differences among existing machine translation systems with reference to the processing of multi-word units. The paper compares the results of distinct machine translation approaches and discusses the usage of combined lexicon-grammar lexical approach and OpenLogos[1] ontology together with semantico-syntactic rules (SEMTAB rules) as a promising solution to overcome machine translation current limitations.

In the line of thought of evaluation proposed by Barreiro et al. (2010), this paper suggests a systematic qualitative evaluation of different linguistic phenomena, starting with multi-word units with different degrees of variability. We propose that, for a fair machine translation evaluation activity, there is the need for a joint qualitative evaluation of the systems to balance with the numerous quantitative evaluations that have taken place in the latest years, which we consider insufficient to measure translation accuracy and linguistic quality.

2 The notion of multi-word unit

A multi-word unit is a group of two or more words or terms in a language lexicon that generally conveys a single meaning. In NLP, many scholars have recently paid special attention to multi-word units, since they represent a thorny issue for most applications, from information retrieval to computer aided translation, or from text mining to semantic web. The usage of concurrent terms of *multi-word unit* (*multi-word*, *multi-word expression*, *fixed expression*, *idiom*, *compound word*, *collocation,* among others) by different theoretical schools, denotes the difficulties in determining the object of study with scientific precision.

These sequences or combinations of words often co-occur with high frequency, recurrently and in a predictable way. They can be contiguous or discontinuous, i.e., with other words in between, but they 'go together' regularly with a precise or conventional meaning. The recombination of these words with their synonyms is usually unacceptable or unusual.

[1] OpenLogos is an open-source rule-based machine translation system available at http://logos-os.dfki.de/.

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 11–19
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5646

2.1. The lexicon-grammar approach

Multi-word units of different types have been extensively and systematically studied within the lexicon-grammar theory, from both theoretical and practical perspectives over a considerable period, by many authors. Researchers of the Laboratoire d'Automatique Documentaire et Linguistique (LADL) have worked on multi-word units since the seventies, inspired by the work done on French by M. Gross (cf. Gross 1975, 1981, 1986). Practical analytical formalization of multi-word units exists for several languages and multi-word units have also been taken into account in contrastive studies, such as those performed by Salkoff (1990; 1999) for English and French.

D'Agostino and Elia (1998) consider multi-word unit as part of a continuum, in which combinations can vary from a high degree of variability of co-occurrence of words (combinations with free distribution), to the absence of variability of co-occurrence. Different processing solutions should be adopted for the different types of multi-word unit combinations.

On one side, multi-words units with a specific grammatical function, an autonomous meaning and with no or almost no variability of co-occurrence among words, such as compound words, need to be lemmatized. Silberztein (2004:117) adopts the following criteria for identifying in a correct way compound words: (i) **Semantic atomicity**: if the exact meaning of a nominal group cannot be deduced from the meaning of the components, as in the case of the Italian *guerra fredda* (*cold war*), in which each element of the compound participates in the construction of a complete and non-literal meaning; (ii) **Distributional restriction**: if certain constituents of the nominal group, which by the way, belong to certain natural distributional classes, cannot be freely replaced, as in the example of the Italian *colletto bianco* (*white collar worker*); (iii) **Institutionalization of the usage**: certain nominal groups are used in a quasi-obligatory manner, to the detriment of other potential syntactic constructions that are just as valid, but are never used. The Italian expression *in tempo reale* (a loan translation of the English *in real time*) is an example for this criterion, which use in Italian seems to be unmotivated if we take into consideration that the antonym **in tempo irreale* (**in unreal time*) is not used at all. These criteria allow the identification of a larger group of compound words than it is normally and traditionally assumed for a language.

The correct identification of multi-word units has also important effects on the quality of translation. As pointed out in Barreiro (2008: 38), **non-translatability**, i.e. the impossibility of a literal translation across cultural and linguistic boundaries, is a property of some multi-word units with limited or no variation of distribution. For example, the famous English idiom: *It's raining cats and dogs*, cannot be literally translated into Italian as *Sta piovendo cani e gatti*. Adaptation of the concept to the Italian language is required, so that the expression *Sta piovendo a catinelle* (literally: *It's raining from jars*) is understood as an *extremely heavy rain*.

On the other side, multi-word units with a high degree of variability of co-occurrence among words have to be handled in a different way and in particular by means of rules, because of their specific morpho-syntactic properties, as we will thoroughly motivate in the next sections.

2.2. The corpus linguistics approach

Within the area of corpus linguistics (Sinclair, 1991, Biber et al., 2000 and Stubbs, 2002, among others) multi-word units, referred as *word clusters* or, most commonly, as *collocations* are extracted by means of concordance tools. Some word co-occurrences are random and purely arbitrary; others are statistically relevant. Statistically relevant co-occurrences represent what corpus linguists designate as *collocation pairs*. Research into collocations has resulted in reference material, such as the Collins COBUILD Collocations (1995) and the BBI Dictionary of English Word Combinations (1997). Other studies on collocations consist in identifying collocates within a corpus, with the goal of including them in extended dictionaries. Biber et al. (2000) discussed the importance of *lexical bundles*, i.e., combinations of words that occur frequently and act as units even though the combination includes not only different parts-of-speech but also words that represent syntactic

functions (e.g. *I don't think that*…). Hoey's *lexical priming* (Hoey, 2005) represents a step forward in the analysis of co-occurring lexical items across larger areas of text and as features of certain types of text.

3 Multi-word units in machine translation

The most critical problems in multi-word unit processing is that they often have unpredictable, non-literal translations; they are numerous and not all included in dictionaries; they may have different degrees of compositionality and their morpho-syntactic properties allow, in some cases, a certain number of formal variations with the possibility of dependencies of elements even when distant of each other in the sentence. These problems result in mistranslations by machine translation systems since not all approaches are capable of processing them correctly. In addition, they can have an opaque meaning, i.e., the meaning of the unit cannot be achieved by the meaning of the individual constituents that make up the unit, so that a literal translation is often not understandable, and incorrect.

The difficulties of multi-word unit recognition in machine translation have been discussed from different viewpoints according to the machine translation modeling approach, i.e. statistical machine translation or rule-based machine translation.

In statistical machine translation, multi-word unit recognition has been handled as a problem of automatically learning and integrating translations of very specific multi-word unit categories, such as domain specific multi-word units (Ren et al., 2009) or as a problem of word alignment (Brown et al. 1993, Och and Ney 2000a, 2000b, 2003 among others). In rule-based machine translation, the identification of multi-word units is mainly based on two different approaches: the lexical approach and the compositional one. In the lexical approach, multi-word units are considered as single lemmata whereas in the compositional approach, multi-word unit processing is obtained by means of tagging and syntactic analysis of its different components.

Current approaches to multi-word unit processing move towards the integration of phrase-based models with linguistic knowledge, in particular syntactic and semantic structures (Chiang, 2005; Marcu et al., 2006; Zollmann and Venugopal, 2006), in order to obtain better translation results, but the solutions undoubtedly vary according to the different degrees of compositionality of the multi-word unit.

4 Comparative evaluation of RBMT and SMT concerning multi-word unit processing

Since multi-word units are processed differently according to the type of approach, we set up a small corpus of non-specialized texts of about 300 sentences (approximately 10,000 words) containing multi-word units extracted from the Web[2]. We used this small corpus with the purpose of analyzing how multi-word units are translated from English into Italian by two machine translation systems with different architectures: Google Translate, a data-driven statistical machine translation system and OpenLogos, a rule-based machine translation system.

All the occurrences of multi-words in the corpus have been compared with the corresponding translations into Italian by the two systems chosen.

In this section we illustrate typical mistranslations concerning (i) multi-word units in which the word *up*[3] occurs, such as in the phrasal verbs *to come up*, *to catch up*, *to stand up for, to mix up* or in expressions like *up and running* and finally (ii) multi-word units which are either complex phrases or idiomatic expressions.

If we analyze the translations into Italian of the sentences (1) and (2) performed by Google Translate, it clearly emerges from the corresponding machine outputs that there is lack of adequate analysis of the source multi-

2 The corpus was extracted from the Web by means of Webcorp LSE (http://www.webcorp.org.uk/webcorp_linguistic_search_engine.html) and Web as a Corpus (http://178.63.122.132/wac).

3 We haven chosen the word *up* which is listed in the dictionary as a verb, adverb, noun, preposition and adjective, since it occurs in many different multi-word units, such as in the phrasal verbs to *mix up, to come up, to call up* or in expressions such as to *be up to something/someone, up and down*, and so on.

word unit *come up* in both contexts, with the generation of the wrong Italian translations. The Italian translation for *come up* in (1) is *venire*, while in (2) is *salire*, both grammatically incorrect and semantically inappropriate. OpenLogos performance in (1) is not very good either, but in (2) the translation of *come up* is correct. The OpenLogos system takes into consideration a wider context than the word level in (2), and analyzes the verb *come up* in connection with the noun questions.

(1) Why does this topic always *come up* at meetings?
Google Translate: Perché questo tema sempre *venire* alle riunioni?
OpenLogos: Perché questo argomento *sale* sempre alle riunioni?

(2) Why did these questions never *come up*?
Google Translate: Perché mai queste domande salire?
OpenLogos: Perché queste domande non *si sono* mai *poste*?

In sentence (3), the phrasal verb *catch up with* occurs with an animate human noun, *philanthropists*. When occurring with a noun of this kind (or a pronoun), the Italian translation is *raggiungere*. Google Translate translated the preposition *with* (as *con*), because it did not recognize it as an element of the multi-word unit. In Italian, the phrasal verb is translated as a single verb, which is immediately followed by the complement noun. OpenLogos linguistic knowledge database permits a correct analysis and translation of this English phrasal verb into the Italian single verb.

(3) Scott Pelley *catches up with* the world's most generous philanthropists
Google translate: Scott Pelley *raggiunge con* più generosi filantropi del mondo
OpenLogos: Scott Pelley *raggiunge* il philanthropists più generoso del mondo

The phrasal verb *stand up for* in sentence (4) is translated literally by Google Translate as *alzare in piedi*. The OpenLogos system produces an acceptable translation for Italian. The correct translation for the multi-word unit (stand up for N/PRON) where N/PRON is a non-animate noun or pronoun, is *difendere* or *lottare per*.

(4) ... this year the Europeans *stood up for* freedom of speech.
Google Translate: quest'anno gli europei *si alzò in piedi per* la libertà di parola.
OpenLogos: questo anno gli Europei *hanno sostenuto* la libertà del discorso.

In sentence (5), the phrasal verb *mix up* occurs with the noun *problems.* In this case it means "to change the order or arrangement of a group of things, especially by mistake or in a way that you do not want". The corresponding Italian translation is *confondere*, as correctly identified by OpenLogos and not *mescolare,* This latter translation is used when the verb *mix up* is in connection with nouns which refer to substances and means "to prepare something by combining two or more different substances".

(5) First of all, IMHO, try not to *mix up* all the different problems *together*.
Google Translate: Prima di tutto, secondo me, cercare di non *mescolare* i vari problemi *insieme*.
OpenLogos: Prima di tutti, IMHO, il tentativo di non *confondere* tutti i problemi diversi *insieme*."

The multi-word unit *world's trouble spots*, in example (5), is also not recognized as an expression by Google Translate, but it is translated correctly by the OpenLogos system as *punti caldi del mondo.*

(6) and travels to some of the *world's trouble spots*
Google Translate: e *viaggia ad* alcuni dei *problemi del mondo spot*
OpenLogos: e *viaggia a* alcuni dei *punti caldi del mondo*

Sentence (6) contains a complex noun phrase containing two compound nouns: *oil rig platform* and *crew survivors*. The correct Italian translation for the noun phrase is *superstisti dell'equipaggio della piattaforma petrolifera*. None of the systems was able to translate the noun phrase correctly, yet, as a grammar-based machine translation system, OpenLogos was capable of inserting the

correct prepositions and determiners (definite articles) that are proper of the Italian morpho-syntactic system for noun phrases. Google Translate translated *oil rig* correctly by using the adjective *petrolifera*, but the internal structure of the noun phrase does not respect the grammar of the Italian language.

(7) … and speaks to one of the *oil rig platform crew survivors*
Google Translate: e parla di uno dei *superstiti piattaforma piattaforma petrolifera equipaggio*
OpenLogos: e parla a uno dei *superstiti dell'equipaggio della piattaforma dell'attrezzatura dell'olio*

Concerning the idiomatic expression *up and running* in sentence (8), neither Google Translate nor OpenLogos are able to produce an acceptable translation.

(8) In Northern Ireland we were *up and running* with Internment
Google Translate: In Irlanda del Nord *abbiamo installato e funzionante* con internamento
OpenLogos: In Irlanda del Nord, siamo stati *alzati e trattare gestire* con l'internamento

Translation problems due to the presence of multi-word units in a sentence, as those discussed in (1)-(8), have highlighted how an inadequate multi-word unit processing may heavily affect the accuracy and the fluency of translations. Mainly in presence of idiomatic expressions, statistical machine translation and rule-based machine translation are not able to produce acceptable translations. In some other examples OpenLogos performs better than Google Translate as a result of the integration of linguistic knowledge into the system by means of a set of semantico-syntactic rules called SEMTAB rules.

5 Integration of semantico-syntactic knowledge

The translation problems discussed in Section 4 can be solved differently, according to the different types of multi-word units. Multi-word units with almost no variability of co-occurrence among words, such as compound nouns, or without any variability of co-occurrence among words, such as idioms, have to be processed as a single unit. Multi-word units with a limited degree of variability of co-occurrence among words can be formalized in semantico-syntactic rules, such as the SEMTAB rules of the OpenLogos system (Scott, 2003; Scott and Barreiro, 2009; and Barreiro et al., forthcoming) and be used to correct mistranslation. These rules analyze, formalize, and translate words in context. They disambiguate the meaning of words of the source text by identifying the semantic and syntactic structures underlying each meaning and provide the correct equivalent translation in the target language. In OpenLogos, they are invoked after dictionary look-up and during the execution of source and/or target syntactic rules at any point in the transfer phase in order to solve various ambiguity problems: (i) homographs, such as *bank*, which can be a transitive and intransitive verb or a noun; (ii) verb dependencies, such as the different argument structures, *speak to*, *speak about*, *speak against*, *speak of*, *speak on*, *speak on N* (radio, TV, television,etc.), *speak over N1*(air) *about N2*, for the verb *speak*; (iii) multi-word units of different nature.

In order to explain the nature and the operation of this type of rule, we discuss it on the basis of the English phrasal verb *mix up.* This verb assumes different meanings according to the words and the nature of the words it occurs with. In (9), it means to change the order or arrangement of a group of things, especially by mistake or in a way that you do not want. In (10), it means to prepare something by combining two or more different substances. In (11), it means to think wrongly that somebody/something is somebody/something else. In (12), it means to be into a state of confusion.

(9) try not to **mix up** all the different problems together
(10) **mix up** the ingredients in the cookie mix
(11) Tom **mixes** John **up** with Bill
(12) I'm all **mixed up**

The different meanings of *mix up* represented in (9)-(12) correspond to different translations in Italian or any other language. Table 1 illustrates the SEMTAB rules comment lines written for the English-Italian language pair. These rules comprehend the

different semantico-syntactic properties of each verb (also called linguistic constraints).

Semantic table (SEMTAB) rule	Italian Transfer
MIX UP (VT) IN	MESCOLARE IN
MIX UP(VT) N IN	MESCOLARE N IN
MIX UP(VT) N WITH	CONFONDERE N CON
MIX UP(VT) N(HUMAN) IN	CONFONDERE N IN
MIX UP (VT) N (INGREDIENT)	MESCOLARE N
MIX UP(VT) N(MEDICINE)	PREPARARE N
MIX UP (VT) WITH	CONFONDERE CON
MIX UP(VT) N(HUMAN,INFO) WITH	CONFONDERE N CON
MIX (VT) UP (PART)	CONFONDERE

Table 1: SemTab rules comment lines for the verb *mix up*

For example, the SEMTAB rule [MIX UP(VT) N(HUMAN,INFO) WITH], describes the meaning (iii) of the verb *mix up*, generalizing to an abstract level of representation the nature of its direct object, classifying it under the Information or Human noun superset of the Semantico-syntactic Abstract Language (SAL) ontology. SAL is the OpenLogos representation language, containing over 1,000 concepts (expandable), organized in a hierarchical taxonomy consisting of supersets, sets, and subsets, distributed over all parts-of-speech. In SAL, both meaning (semantics), and structure (syntax) are merged. This type of abstraction allows coverage of a number of different sentences in which different types of Human nouns occur, as illustrated in (13).

(13) Tom **mixed** *John/him/the brother/the man/the buyer/the Professor*, **up with** Bill.

In order to properly disambiguate multi-word units, it is necessary to take into account a much wider context than the simple word level and apply context-sensitive semantico-syntactic rules, which in the case of the different meanings of *come up*, in (1) and (2), distinguish between (*N (topic, question) Vprep (come) Prep (up) → N (domanda) V (porsi)*) and (*V (come) Prep (up) → V (salire)*).

An unusually powerful aspect of SEMTAB is that the rules are conceptual, deep structure, meaning that each rule can apply to a variety of surface structures, regardless of word order, passive/active voice construction, etc., The same rule can apply to different surface structures, e.g., *the mixing up of languages*, *mix up the languages*, *languages mix up*, etc. These very simple examples show how an adequate identification and analysis of multi-word units in the source language by means of semantico-syntactic rules can influence the performance of a machine translation system with reference to different language pairs.

SEMTAB rules are integral part of the OpenLogos system[4] and represent one of its most important and powerful processing components: as an example, the English-Italian language pair has over 14,000 semantico-syntactic rules to identify the meanings of words in context and to assign the correct translation to each of the detected meanings. In our opinion, it is possible to reuse and integrate the linguistic knowledge provided by the semantico-syntactic rules also in other machine translation systems as long as the SAL ontology is adopted for the description of the semantico-syntactic features of the lexical items and deep parsing is performed in order to properly disambiguate the source language texts.

6 Qualitative machine translation evaluation metrics

In order to verify the validity of our approach to multi-word unit processing on a large-scale, a qualitative evaluation metrics should be adopted. Evaluation is a crucial issue in machine translation development and in this respect, automatic machine translation evaluation, which assesses the results of a machine translation process by ranking the quality of translations on the basis of statistical, language-independent algorithms, has been considered the best method in recent years, such as Bleu (Papineni, 2002) and NIST (Doddington, 2002) metrics. Recent studies move towards the usage of linguistic knowledge, either to integrate or to substitute pure statistical methods in order to obtain metrics which are closer to human evaluation of translations. Interesting proposals have been presented by Agarwal and Lavie (2007, 2008), Giménez, J. and Màrquez (2010), Lavie and Denkowski (2010), among others.

[4] Typical SEMTAB rules in the OpenLogos system can be viewed at http://logossystemarchives.homestead.com/SEMTAB/SEMTABscan10662.html

All these metrics only partially give reliable results concerning machine translation quality, since the judgement is based not on whether a machine translation system translates accurately the meaning and the message of an original text, but only how well it scores against references. When evaluating translations, the target text has to be assessed from two different viewpoints: (i) as a text derived from a source text, to which it has to be compared in terms of accuracy, (ii) as an autonomous text in the target language and culture, which has to be judged in terms of fluency.

Accuracy together with fluency are the two main quality criteria that have to be taken into account and cannot be considered separately. Besides these two criteria, there are other two criteria which are more dynamic and oriented towards situationality, i.e. adequacy and acceptability[5]. Machine translation results have been often judged according to these two criteria, as the expressions "good enough" or "fit-to-the-purpose" highlight. Users' expectations may vary according to their final communicative goals and in this case also acceptability standards may change so that fairly inaccurate translations, as machine translation results sometimes are, can nevertheless perfectly meet user's requirements (Monti, 2005), but only if translations are used for assimilation purposes.

When we consider machine translation for dissemination purposes, quantitative evaluation seems not to be adequate because, it is not possible to assess the accuracy of the transmission of the contents of the source text in the target text and consequently the presence of translation errors. This is due to the fact that the comparison is performed between the candidate translation (the machine translation output) and the reference translations and not, as it should be, between the source text and the machine translation output. Furthermore, the automatic comparison measures only the similarity of the candidate translation to one or more reference texts and inevitably penalizes any motivated lexical, syntactical, stylistic variations which can occur between the candidate and its references.

[5] Translation quality criteria have been discussed by several authors such as Scarpa (2001).

There is, in our opinion, the need for a qualitative evaluation metrics of machine translation which, besides fluency, takes into account the accuracy of machine translation outputs, by means of a comparison between the source text and one or more target texts in order to balance with the numerous quantitative evaluations that have been taking place in recent years.

As pointed out in Ming Zhou et al. (2008), who propose to evaluate the capability of a machine translation system in handling various important linguistic test cases called Check-Points, i.e. linguistically motivated units, which are pre-defined in a linguistic taxonomy for diagnostic evaluation, we suggest the development of more sophisticated evaluation tools that measure the performance of specific linguistic phenomena from a qualitative point of view.

In our opinion, different types of multi-word unit combinations represent an important linguistic critical area to be investigated and evaluated with respect to different machine translation approaches.

In order to perform fair qualitative machine translation evaluation, human assessment of the outputs of different machine translation systems with the aid of specific machine translation evaluation tools, is required. An "ideal" evaluation tool should allow users to submit a translation simultaneously to various machine translation systems and rank the accuracy of the translation results, with regards to specific linguistic test cases, such as multi-word units.

7 Conclusions

This paper analysed the problem of multi-word unit processing in machine translation systems with different approaches, i.e. Google Translate and OpenLogos. The paper focused on the different possible solutions for an effective processing of the multi-word units and suggested to adopt the lexicon-grammar approach and OpenLogos semantico-syntactic rules for multi-word unit processing. In this way, it is possible to handle different types of multi-word units with different representation levels (dictionaries and/or semantico-syntactic rules). The integration of the lexicon-grammar and OpenLogos approaches, based on the analysis of millions of phrasal contexts

by distributional constraints and contextual semantico-syntactic rules, leads to an accurate treatment of multi-word units.

The paper also highlighted the need of a joint qualitative machine translation evaluation metrics that allows the comparison of machine translation systems based on different approaches, with regard to specific linguistic phenomena.

Note

Johanna Monti is author of sections 3, 4, 6 and conclusions, Anabela Barreiro is author of abstract and sections 1 and 5, Annibale Elia is author of section 2.1. Federica Marano is author of section 2.2. and Antonella Napoli is author of section 2. The English-Italian parallel corpus used for this paper was analyzed by Antonella Napoli and Federica Marano with reference to the detection of the multi-word units in the English corpus and by Johanna Monti with reference to the comparative evaluation of the Italian machine translations of the multi-words detected.

References

Agarwal, A., and Lavie, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments, *Proceedings of the ACL 2007 Workshop on Statistical Machine Translation*,

Agarwal, A., and Lavie, A. (2008). METEOR, M-BLEU and M-TER: Evaluation Metrics for High-Correlation with Human Rankings of Machine Translation Output, *Proceedings of the ACL 2008 Workshop on Statistical Machine Translation*.

Barreiro, A. (2008). *Make it simple with paraphrases. Automated paraphrasing for authoring aids and machine translation*. PhD dissertation. Faculdade de Letras da Universidade do Porto, Oporto, Portugal.

Barreiro, A., Elia, A., Monti, J. and Monteleone M. (2010). Mixed up with machine Translation: Multi-word Units Disambiguation Challenge. *Proceedings of Translating and the Computer 32 ASLIB Conference* – London, 18-19 November.

Barreiro, A., Scott, B., Kasper, W. and Kiefer, B. (forthcoming - 2011). OpenLogos Rule-Based Machine Translation: Philosophy, Model, Resources and Customization. Machine Translation. *Special Issue on Free/Open Source Machine Translation.* Editors Mikel L. Forcada and Felipe Sánchez-Martínez.

Biber, D., Johannson, S., Leech, G., Conrad, S. and E. Finegan (2000). *Longman Grammar of Spoken and Written English*. Harlow: London.

Brown, P., Della-Pietra, V., Della-Pietra, S., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2), 263–312.

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

D'Agostino, E. and Elia, A. (1998). Il significato delle frasi: un continuum dalle frasi semplici alle forme poliremathiche, Leoni, F.A., Gambarara, D., Gensini S., Lo Piparo, F., and Simone, R. *Ai limiti del linguaggio*, Bari, Roma: 287-310.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-Gram Co- Occurrence Statistics. *Proceeding of the Second International Conference on Human Language Technology*, San Diego, CA: 138-145.

Giménez, J. and Màrquez, L. (2010). *Linguistic Features for Automatic MT Evaluation*. To Appear in Machine Translation, Springer, Netherlands.

Gross, M. (1975). *Méthodes en Syntaxe - Regime des constructions completives*. Hermann, Paris.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Formes Syntaxiques et Prédicat Sémantiques, Langages 63,* Larousse, Paris, France: 7-52.

Gross, M. (1986). Lexicon-Grammar. The representation of compound words. *Proceedings of the 11th International Conference on Computational Linguistics, COLING'86*, Bonn, West German: 1-6.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language.* London: Routledge.

Lavie, A. and Denkowski, M. (2010). The METEOR Metric for Automatic Evaluation of Machine Translation. *Machine Translation*.

Marcu, D., Wang, W., Echihabi, A. and Knight, K. (2006). SPMT: statistical machine translation with syntactified target language phrases. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, July 22-23, Sydney, Australia.

Monti, J. (2005). La traduzione automatica deve sempre essere trasparente? Spunti di riflessione su trasparenza e qualità nella traduzione. Guarino, A., Montella, C., Silvestri, D., and Vitale, M., *La Traduzione: Il Paradosso Della Trasparenza.* Liguori Editore, Napoli: 299-318

Och, F. J., and Ney, H. (2000a). A comparison of alignment models for statistical machine translation. *Proceedings of the 18th international conference on computational linguistics (COLING-2000),* Saarbrücken, Germany: 1086–1090.

Och, F. J., and Ney, H. (2000b). Improved statistical alignment models. *Proceedings of the 38th annual meeting of the ACL*, Hong Kong, China: 440–447.

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

Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002)*: 311-318.

Ren, Z., Lü, Y., Cao, J., Liu, Q. and Huang, Y. (2009). Improving statistical machine translation using domain bilingual multiword expressions. *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, Suntec, Singapore: 47-54.

Salkoff, M. (1990). Automatic translation of support verb constructions. *13th International Conference on Computational Linguistics (COLING 1990)* (University of Helsinki, Finland, 20-25 August 1990): 243-246.

Salkoff, M. (1999). *A French-English Grammar: A Contrastive Grammar On Translational Principles.* John Benjamins Publishing Company, Amsterdam, Philadelphia.

Scarpa, F. (2001). *La traduzione specializzata*, Milano, Hoepli Editore.

Scott, B. (2003). The Logos Model: An Historical Perspective. *Machine Translation* 18: 1-72.

Scott, B. and Barreiro, A. (2009). OpenLogos MT and the SAL Representation Language. Pérez-Ortiz, J.A. , Sánchez-Martínez, F. and Tyers, F.M., (eds.), *First International Workshop on Free/Open-Source Rule-Based Machine Translation* (Alicante, Spain, 2-3 November 2009): 19-26.

Silberztein, M. (2004). *Intex Manual*. (http://*intex*.univ-fcomachine translatione.fr/downloads/*Manual*.pdf.)

Sinclair, J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.

Stubbs, (2002). Two quantitative methods of studying phraseology. *English. International Journal of Corpus Linguistics.* 7, 2: 215-44.

Zhou, M., Wang, B., Liu, S., Li, M., Zhang D. and Zhao T. (2008). Diagnostic evaluation of machine translation systems using automatically constructed linguistic check-points. *Proceedings Coling 2008:* 1121–1128.

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