Mixed up with Machine Translation: Multi-word Units Disambiguation Challenge

Anabela BARREIRO[1], Annibale ELIA[2], Johanna MONTI[2] and Mario MONTELEONE[2]
barreiro_anabela@hotmail.com, elianni@tin.it, jmont@tin.it, mmonteleone@unisa.it,

[1] Centro de Linguística da Universidade do Porto, Via Panorâmica, s/n, 4150 – 564 Porto, Portugal
[2] Dip. di Scienze della Comunicazione, Università degli Studi di Salerno, Via Ponte Don Melillo 84084, Fisciano, Italia

Abstract

With the rapid evolution of the Internet, translation has become part of the daily life of ordinary users, not only of professional translators. Machine translation has evolved along with different types of computer-assisted translation tools. Qualitative progress has been made in the field of machine translation, but not all problems have been solved. One problem in particular, namely the poor analysis and translation of multi-word units, is an arena where investment in linguistic knowledge systems with the goal of improving machine translation would be beneficial. This paper addresses the difficulties multi-word units present to machine translation, by comparing translations performed by systems adopting different approaches to machine translation. It proposes a solution for improving the quality of the translation of multi-word units by adopting a methodology that combines Lexicon Grammar resources with OpenLogos lexical resources and semantico-syntactic rules. Finally, it discusses the need to create new evaluation metrics and a machine translation evaluation tool to correctly evaluate the performance of machine translation engines with regards to multi-word units and thus to contribute to the improvement of translation quality.

Introduction

The Internet has helped machine translation to become increasingly popular within the general public. Today millions of Internet users take advantage of machine translation to quickly obtain information on the contents of a text or a web page written in a foreign language, to exchange information in real-time, to retrieve information in unknown languages, or even to produce publishable translations. Most recently, machine translation is used for dissemination purposes in online collaborative translation environments (Monti, forthcoming). This unpredictably quick turn on machine translation usability complements traditional uses as the ones described in Hutchins (2005), where the challenge of producing high quality translations was big, but more controllable. The world of machine translation has changed forever: the spectrum of language to be translated by machines is now broadening and more complex, less controlled and more idiomatic. Considerable progress has also been made in qualitative terms because of the availability and use of large parallel

---

[1] Anabela Barreiro is author of abstract, introduction and section 4, Annibale Elia is author of section 1, Mario Monteleone is author of section 2, Johanna Monti is author of sections 3 and 5 and conclusions.

corpora, the development of knowledge bases, the adoption of statistical models, and the integration with various computer assisted translation tools, particularly with translation memories. However, despite recent significant progress, lexical problems still represent a critical area in machine translation, and among lexical problems, multi-word units, are particularly difficult to be processed by machine translation systems.

The aim of this paper is to provide evidence of the shortcomings of existing machine translation systems with reference to the processing of multi-word units, and in the line of thought of evaluation proposed by (Barreiro, 2008) suggest a systematic qualitative evaluation of different linguistic phenomena, starting with multi-word units with different degrees of variability. The paper points out benefits, strengths and weaknesses of distinct machine translation approaches and discusses the usage of combined Lexicon-Grammar lexical resources and OpenLogos lexical resources together with semantico-syntactic rules (SEMTAB rules) as a possible solution to overcome machine translation limitations with regard to the automated processing and translation of multi-word units. We propose that, for a fair machine translation evaluation activity, there is the need for a serious joint qualitative evaluation of the systems to balance with the numerous quantitative evaluations that have taken place in the latest years by automated evaluation tasks, including BLEU, NIST and METEOR, which we consider insufficient to measure translation accuracy and linguistic quality. We propose that qualitative evaluation will be made with the aid of a new machine translation evaluation tool. This paper presents the results of a research based on the translation of sentences containing multi-word units from English into Italian of a non-specialised text corpus. Section 1 presents the notion of multi-word unit in the framework of the Lexicon Grammar theory. Section 2 analyses how multi-word units are processed using the state-of-the-art machine translation technology. Section 3 discusses several examples of lexical ambiguities concerning multi-word units in the translations performed by a statistical machine translation system and a rule-based machine translation system, and highlights, analyzes and discusses how two machine translation systems of a different conceptual nature perform with regards to different types of multi-word unit. Section 4 discusses the possibility of using semantico-syntactic rules in order to obtain better translation quality results. Section 5 discusses the need for new evaluation metrics and for a machine translation evaluation tool

to correctly evaluate the performance of machine translation engines with regards to multi-word units. Section 6 presents the conclusions.

1 – Multi-word Units in the Lexicon-Grammar

Multi-word units have been an increasingly important concern for natural language processing scholars, being considered as a ''pain in the neck for NLP'' (Sag et al., 2001), because of the many difficulties they raise. To begin with, there is no universally agreed definition or term for the concept of multi-word unit. Concurrent terms of *multi-word unit* are *multiword*, *multiword expression*, *fixed expression*, *idiom*, *compound word*, and *collocation* used by many authors of different theoretical schools or following distinct natural language processing approaches. In earlier Lexicon Grammar Theoretical Framework, established by Maurice Gross (cf. Gross, 1975, 1981), the most essential features of what we call multi-word unit, were the non-compositionality and semantic opaqueness. (Gross, 1986), uses the term *compound word* to refer to a string composed of several words whose meaning cannot be computed from its elements. (De Mauro, 2000) describes it as a group of words with a single meaning, which cannot be inferred from the meanings of the individual words that are part of it, both in the current usage of language and in specialized language. Recently the significance of compositionality has changed and the term multi-word unit has evolved in such a way that it can also be referred to non-idiomatic units, being now used to refer to various types of linguistic entities, including idioms, compounds, phrasal verbs, light or support verb constructions, lexical bundles, etc. At present, many applications that involve translation, such as machine translation systems, cross-language information retrieval, and computer-assisted language learning applications process texts and look for equivalent translations on a word level (or n-gram level), whereas in many cases equivalence has to be established on level of units larger than single words. Some combinations of words or n-grams have no linguistic significance (e.g., *the war*), while others are linguistically meaningful (e.g., *cold war*). With these two examples, only *cold war* is considered a multi-word unit. Machine translation needs to recognize and handle correctly multi-word units.

Lexicon-Grammar scholars have been studying multi-word units for years now and the practical analytical formalization has been done for several languages (cf. Ranchhod 1990,

Baptista 2005, and Chacoto 2005, for Portuguese; Silberztein, 1993, Laporte & Voyatzi, 2008, for French; and Elia & D'Agostino 1998, Vietri 2004, and Monteleone et al. 2008, for Italian, among others). Lexicon-Grammar scholar Morris Salkoff has dedicated attention to multi-word units in his contrastive studies, particularly between English and French (Salkoff, 1990; 1999).

Elia & D'Agostino (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. They identify four different types of combinations of phrases or sentences, namely (i) with a high degree of variability of co-occurrence among words, and therefore are combinations with free internal distribution, compositional and denotative meaning, such as in *dirty water,* or *clean water*; (ii) with a limited degree of variability of co-occurrence among words, and therefore are combinations with restricted internal distribution, such as in *natural water,* or *mineral water*; (iii) with no or almost no variability of co-occurrence among words, and therefore are combinations with fixed internal distribution, such as in *heavy water*; and (iv) without any variability of co-occurrence among words, and therefore are proverbs, such as *all good things come to he who waits*. The several degrees of variability or invariability can manifest in compounds, as the illustrated *water* compounds, but also in other types of multi-word units. As demonstrated in (Barreiro, 2008), multi-word units have been classified into three main categories: lexical units (with all the compounds), frozen and semi-frozen expressions (including phrasal verbs (*show up*), support verb constructions (*give a (big) hug to*) and proverbs), and lexical bundles (*I think that*; *Would you mind if*). Descriptions and examples of all the different types of multi-word unit can be found in the same work. Some multiword units do not fit into any of these three major types.

Each type of multi-word unit may need to follow a different formalization method. There is the morphological aspect of multi-word unit (i.e., the morphology of composition) that weights considerable for morphologically-rich languages and remains a highly challenging task. From a lexicographical point of view, multi-word units with a specific grammatical function and an autonomous meaning need to be registered in dictionaries in a systematic way, i.e. as autonomous lemmata and not, as often is the case in traditional dictionaries, as examples of use of head nouns or adjectives. As far as lemmatization is concerned, a clear

distinction between multi-word units with a high degree of variability of co-occurrence among words and those with a limited or no variability of co-occurrence among words (compound words, idiomatic expressions, proverbs) should be made. This is one of the most critical issues in the description of natural languages. For example, there is a relevant difference in Italian between *colletto bianco* (with the meanings of **"**white collar" and "white collar worker") and *colletto rosso* ("red collar"). The first has to be lemmatized since it has the specific meaning of "employee" with distinctive morpho-grammatical and lexical properties, i.e. singular masculine "human being", with *colletti bianchi,* as its masculine inflected form. This is quite a simple example of the difference between opposite poles in the continuum. Sometimes, however, multi-word units are much more difficult to classify and describe. For example, the Italian multi-word unit *editto bulgaro* (*Bulgarian edict*), taken from the political language and referring to a speech of the Italian prime minister Berlusconi in 2002 against some journalists and their banishment from the Italian Broadcasting Service, and *elezione bulgara (Bulgarian elections)* are on the edge between the status of compound words and that of free nominal groups. This is a problem that occurs most frequently with compound words. (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, the nominal group must be lexicalized (=> it is therefore treated as a compound noun). This is the case in Italian of *berretto verde* (officers of the Guardia di Finanza, an Italian police force under the authority of the Minister of Economy and Finance), *teste di cuoio* (member of a special anti-terrorist police team), *casa chiusa* (brothel) *guerra fredda* (cold war), in which each element of the compound participate 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, then this distributional restriction must be acknowledged by classifying the series of nominal groups in a lexicon, which again, amounts to treating it as a compound noun. For example the above-mentioned example of *colletto bianco* and *colletto rosso* follow this criterion; (iii) **Institutionalization of the usage**: certain nominal groups, even those that are semantically and distributionally "free", 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 identifying a larger group of compound words than it is normally and traditionally assumed for a language.

Another important level of analysis of multi-word units concerns their morpho-syntactic classification, which can be performed inside simple sentences and on a distributional basis. For example, compound words can be identified and therefore lemmatized also on the basis of their morpho-syntactic properties. Lemmatization of multi-word units that belong to classes with limited or no variation of distribution (semi-frozen or frozen expressions), such as technical multi-word units, idioms, and proverbs, has important consequences in natural language processing, in text automatic analysis, in terminology, in the structure of the semantic web, and in computer aided translation. In particular, the correct identification of multi-word units has important effects on the quality of translation. As pointed out in (Barreiro, 2008: 38), **non-translatability**, i.e. the meaning cannot be translated literally across cultural and linguistic boundaries, is a property of some multi-word units with limited or no variation of distribution. Many idioms and proverbs cannot be translated word by word. 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*. The same property can be applied to other types of multi-word unit. For example, the English literal translation of the Italian verbal expression *compilare un modulo* (*compile a module*) does not convey the correct meaning. The correct translation is indeed *to fill in a form*.

In section 2, we will focus on multi-word units in machine translation, and in particular on the different solutions applied in multi-word unit processing.

2 – Multi-word Units in Machine Translation

The importance of a correct processing of multi-word units in machine translation and computer aided translation has been stressed by several authors. (Thurmair 2004) underlines

how translating multi-word units word-by-word destroys their original meanings. Villavicenzio et al. (2005) underline how the recognition of multi-word units is necessary for machine translation systems to preserve the meaning and produce accurate translations. (Diakonescu, 2004), stresses the difficulties of multi-word unit processing in machine translation and proposes a method based on Generative Dependency Grammars with features. Váradi (2006) underlines how multi-word units significantly contribute to the robustness of the machine translation systems since they reduce the ambiguity in word to word machine translation matching and proposes the usage of local grammars to capture the productive regularity of multi-word units. Lambert and Banchs (2006) suggest a strategy for identifying and using multi-word units in statistical machine translation, based on grouping bilingual multi-word units before performing statistical alignment. (Barreiro, 2008) describes where and why machine translation engines are unsuccessful at handling the translation of support verb constructions, and finds a method based on paraphrases to overcome the machine's inability in translating them. The proposed solution helps improve their translation in 40% with the help of a tool called ReWriter that transforms support verb constructions into single verbs. (Hurskainen, 2008), states that the main translation problems in machine translation are connected to multi-word units. (Rayson et al., 2009), underline the need for a deeper understanding of the structural and semantic properties of multi-word units in order to develop more efficient algorithms. Moszczyński (2010) explored the potential benefits of creating specialized multiword lexica for translation and localization applications.

The most critical problems in multi-word unit processing is that they often have unpredictable, non-literal translations; they are numerous[2], and not all included in dictionaries; they may have different degrees of compositionality (from free combinations to frozen multi-word units, as for the English noun phrase *round table)*; 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

---

[2] Barreiro, 2008 performed an experiment that showed that in a corpus of fiction, 89% of the occurrences of the verb *dar* (*give*), 88% of the verb *tomar* (*take*), 77% of the verb *por* (*put*), 47% of the verb *fazer* (*do/make*), and 20% of the verb *ter* (*have*) (i.e., 5 of the most common verbs in Portuguese) were in a support verb construction. This means that, globally, 5 of the most frequent verbs occur in support verb constructions 64.2% of the time in fiction texts.

meaning, i.e., the meaning of the unit cannot be achieved by the meaning of the individual constituents that make up the unit, so a literal translation is often not understandable, and incorrect.

The problem of multi-word unit recognition in machine translation has 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, which evolved from the IBM word-based models (Brown et al., 1988, 1990) to phrase-based models (Zens et al., 2002; Koehn et al., 2003; Tillmann and Xia, 2003), 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. The statistical phrased-based approach, based on n-grams, i.e. words which are adjacent, are quite able to identify multi-word units with no or almost no variability in co-occurrence among words (i.e. idiomatic expressions and proverbs), whereas it shows some shortcomings in identifying multi-word units with a high and limited degree of variability of co-occurrence, as it is the case of phrasal verbs and support verb constructions, 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 approach. In the lexical approach, multi-word units are considered as single lemmata and lemmatized as such in the system dictionaries. In the compositional approach, multi-word unit processing is obtained by means of tagging and syntactic analysis of the different components of a multi-word unit.

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.

In section 3, we will analyze examples of lexical ambiguities concerning multi-word units in the translations performed by a statistical machine translation system and a rule-based machine translation system, and compare the shortcomings of existing machine translation

systems when translating multi-word units with different degrees of variability. With this exercise, we point out benefits, strengths and weaknesses of the two distinct approaches.

3 - Lexical Ambiguities Handled by Different Systems

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 (approximatively10,000 words) containing multi-word units extracted from the Web using two different tools: *Webcorp LSE*[3], developed by the Research and Development Unit for English Studies (RDUES), based in the School of English at Birmingham City University, and *Web as a Corpus*[4], developed by Bill Fletcher. We used this small corpus with the purpose of analyzing how multi-word units are translated by two machine translation systems with different architectures. The two systems are: Google Translate, a statistical data-driven statistical machine translation system and Openlogos, a rule-based machine translation system.

In order to give an idea of the dimension of the problem we have in hands, let us consider the word *up*. This word is listed in the dictionary as a verb, adverb, noun, preposition and adjective, and 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. 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-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*.

---

[3] http://www.webcorp.org.uk/webcorp_linguistic_search_engine.html
[4] http://178.63.122.132/wac/

(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*?

Sentence (3) 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.

(3) … 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*

In sentence (4), 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.

(4) 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 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*.

> (5) and travels to some of the *world's trouble spots*
>   Google Translate: e *viaggia ad* alcuni dei *problemi del mondo spot*
>   OpenLogos: bolletta e *viaggia a* alcuni dei *punti caldi del mondo*

Finally, the phrasal verb "*stand up for*" in sentence (6) 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*.

> (6) ... 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.

Translation problems due to the presence of multi-word units in a sentence, as those discussed in (1)-(6), have highlighted how an inadequate multi-word unit processing may heavily affect the understandability and the correctness of translations. In the cases illustrated, statistical machine translation performs worse than rule-based machine translation. Sometimes, mainly in presence of idiomatic expressions, both approaches are not able to produce an acceptable translation. The generally better performance of the OpenLogos system in translating sentences (1)-(6) is a result of the integration of linguistic knowledge into the system. This linguistic knowledge is a set of semantico-syntactic rules called SEMTAB rules, which we will describe in section 4.

4 – Integration of Semantico-Syntactic Knowledge

The translation problems discussed in Section 3 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, like compound nouns, or without any variability of co-occurrence among words, like idioms, have to be processed as a single unit and therefore need to be lexicalized, even if one element of the multi-word unit inflects, as for example the verb in the

expression *make a storm in a teacup*. 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. SEMTAB rules analyze, formalize, and translate words in context. SEMTAB disambiguates 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 (TRAN 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 (7), 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 (8), it means to prepare something by combining two or more different substances. In (9), it means to think wrongly that somebody/something is somebody/something else. In (10), it means to be into a state of confusion.

(7)   try not to **mix up** all the different problems together

(8)   **mix up** the ingredients in the cookie mix

(9)   Tom **mixes** John **up** with Bill

(10)   I'm all **mixed up**

All these different meanings of *mix up* represented in (7)-(10) correspond, obviously, 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
1. MIX UP(VT) IN	MESCOLARE IN
2. MIX UP(VT) N IN	MESCOLARE N IN
3. MIX UP(VT) N WITH	CONFONDERE N CON
4. MIX UP(VT) N(HUMAN) IN	CONFONDERE N IN
5. MIX UP(VT) N(INGREDIENT)	MESCOLARE N
6. MIX UP(VT) N(MEDICINE)	PREPARARE N
7. MIX UP(VT) WITH	CONFONDERE CON
8. MIX UP(VT) N(HUMAN,INFO) WITH	CONFONDERE N CON
9. MIX(VT) UP(PART)	CONFONDERE

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

For example, the SEMTAB rule number 8 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 (11).

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

In order to properly disambiguate multi-word units, it is necessary to take into consideration 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 example (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., approaching Chomsky theoretical assumptions about universality of language. 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 hand-drafted semantico-syntactic rules can influence the performance of a machine translation system with reference to different language pairs. Linguists can create rules that are more or less general, or they can create very specific rules, depending on the type of multi-word unit. SEMTAB comment lines are written by a linguist, but the rules are built automatically by using an appropriate tool (SEMANTHA or SEMTAB rule editor).

5- Qualitative Machine Translation Evaluation Metrics

In order to verify the validity of our approach to multi-word unit processing on a large-scale, a joint evaluation tool should be created and used. Evaluation is a crucial issue in machine translation development and the design of objective and fast evaluation techniques to measure machine translation quality systems represents one of its main goals. 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. Among the most well-known automatic evaluation metrics, Bleu Metric, the best-adopted IBM-developed metric (Papineni 2002) and NIST, the National Institute of Standards and Technology developed metric (Doddington 2002), are both based on the comparison between reference translations and candidate translations performed by a machine translation system in order to determine how close are the latter to professional human translations. Their algorithms use n-grams and count the number of matches. The higher the number of matches the higher are the scores attributed to machine translations. 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: Lavie and Agarwal, (2007, 2008), Gimenez and Marquez (2010), Denkowski and Lavie (2010), among others.

All these metrics have been designed with a view to the reduction of time and cost of assessment in comparison to human judgment, but they only partially give reliable results concerning machine translation quality. 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, whether human or machine translation, the target text has to be assessed from two different viewpoints: (i) the

target text as a text derived from a source text, to which it has to be compared in terms of accuracy, i.e. how accurately the contents of the source text have been reformulated in the target text; (ii) the target text as an autonomous text in the target language and culture, so that it has to be as fluent as the source text in the source language and culture.

Translation quality criteria have been discussed by several authors such as Scarpa (2001). Accuracy together with fluency are the two main quality criteria that have to be taken into account when evaluating translations, both human and machine translations. Accuracy and fluency cannot be considered separately, since a translation can be accurate but not fluent at all, or on the contrary not accurate but very fluent in the target language. Besides these two criteria, there are other two criteria which are more dynamic and oriented towards situationality, i.e. adequacy and acceptability. Adequacy has sometimes been used in the sense of accuracy, having been selected for instance by Defense Advanced Research Projects Agency (DARPA) researchers as one of the quality assessment criteria together with informativeness, and fluency. Adequacy, from their viewpoint, measures the degree to which the translation communicates information present in the original. But from the viewpoint of many translation scholars (Hatim and Mason 1990, for example) adequacy can be judged in terms of adherence of translations to the specifications of a particular translation task and in terms of user's needs satisfaction. Adequacy is a dynamic concept and can vary according to the different actors involved in a translation process: the customer, the translator, the final user, etc. From the user's perspective the translation is adequate when it meets the relevance principle (Grice 1975; Sperber and Wilson, 1986), i.e. when the translation reaches the best communicative effect with the least cognitive effort by the user. Of course, 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).

From this perspective, quantitative evaluation seems not to be adequate to our purposes because, first of all, 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, and secondly, 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 variation which can occur between the candidate and its references.

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 the target text. Some experiments in this direction were made recently by authors such as Kroon (2007) and Jonker (2008) to balance with the numerous quantitative evaluations that have been taking place in recent years. In order to perform fair qualitative machine translation evaluation, human assessment with the aid of specific machine translation evaluation tools, is required.

The current times are auspicious for the development of more sophisticated evaluation tools that measure the performance of specific linguistic phenomena, including multi-word units. An "ideal" evaluation tool should allow professional translators to submit a translation simultaneously to various machine translation systems available on the Web, and rank the accuracy of the translation results. Such tool would have the advantage of permitting to compare results and evaluate the quality obtained by the various engines at a glance. Development of this evaluation tool in the wished direction would be the possibility by the professional translators to choose the best of those translations and correct them, if needed, on the spot. The data collected could be used as a basis for the selection and implementation of new semantico-syntactic rules to improve the translation of multi-word units.

6- Conclusions

This paper analyses the problem of multi-word unit processing in machine translation, proposes a new approach to improve the quality of machine translation by using semantic-syntactic rules, and envisages the development of a new joint qualitative evaluation tool. Multi-word units were defined according to the Lexicon-Grammar theoretical framework. We have mentioned the state-of-the-art in machine translation, and highlighted the differences between the statistical and the rule-based approach concerning this problematic linguistic phenomenon. We also presented the results of the translation of a small non-specialized corpus by Google Translate and OpenLogos, and highlighted the different problems, which emerged from this analysis. We focused then on the different possible

solutions for an effective processing of the multi-word units and suggested to adopt semantico-syntactic rules such as the SEMTAB rules, which were only superficially discussed. The paper ended up by discussing the need of a joint qualitative machine translation evaluation metric. We proposed the creation of an evaluation tool that allows comparison of different systems with regards to distinct linguistic phenomena. This envisioned tool retrieves several machine translation outputs simultaneously and allows professional translators to compare results of the different systems and suggest translations that would correct bad machine translations. This tool would gather linguistic knowledge that would help build sets of rules to be used in improving machine translation. We believe that current machine translation would benefit significantly from the proposed resources and tool, especially in what concerns multi-word units.

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*, [http://www.cs.cmu.edu/~alavie/METEOR/old/Lavie-Agarwal-2007-METEOR.pdf]

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*.
[http://www.cs.cmu.edu/~alavie/METEOR/old/Agarwal-Lavie-2008-METEOR.pdf]

Baptista, J. (2005*). Sintaxe dos Nomes Predicativos com verbo-suporte SER DE*. Fundação para a Ciencia e a Tecnologia/Fundacao Calouste Gulbenkian, Lisboa.

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. [http://www.linguateca.pt/Repositorio/AB-Thesis_030409.pdf]

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.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V., Jelinek, F., Mercer, R. and Roossin, P. (1988). A statistical approach to language translation. *Coling Budapest*, 71-76.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V., Jelinek, F., Lafferty, D., Mercer, R.L. and Roossin, P.S.(1990). A statistical approach to Machine Translation. *Computational Linguistics*, 16, 79-85.

Chacoto, L. (2005). *O Verbo Fazer em Construcoes Nominais Predicativas*. PhD dissertation. Universidade do Algarve, Portugal.

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

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

De Mauro, T. (2000). *Dizionario della lingua italiana.* Paravia, Milano.

Diakonescu, S. (2004). Multiword Expression Translation Using Generative Dependency Grammar. *Advances in Natural Language Processing 4th International Conference, EsTAL 2004*, Alicante, Spain, October 20-22: 243-254.

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.

Elia, A., De Bueriis, G. (a cura di) (2008). *Lessici elettronici e descrizioni lessicali, sintattiche morfologiche ed ortografiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici – ALTI*, Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno, Plectica, Salerno.

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

Grice, P. (1975). *Syntax and Semantis.* Academic Press, New York.

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

Gross, M. (1981). Les bases empiriques de la notion de prédicat semantique. *Formes Syntaxiques et Predicat Semantiques, 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.

Hatim, B. and Mason, I. (1990). *Discourse and the translator*. Longman, London.

Hurskainen, A. (2008). Multiword Expressions and Machine Translation. *Technical Reports in Language Technology* Report No 1, 2008 [http://www.njas.helsinki.fi/salama/multiword-expressions-and-machine-translation.pdf]

Hutchins, J. (2005). Current commercial machine translation systems and computer-based translation tools: system types and their uses. *International Journal of Translation* vol.17, no.1-2, Jan-Dec 2005: 5-38. [http://www.hutchinsweb.me.uk/IJT-2005.pdf]

Jonker, J. A. (2008). *Qualitative Analysis of Moses.* [http://hmi.ewi.utwente.nl/verslagen/capita-selecta/CS-Jonker-Jorik-2.pdf]

Koehn, P., Och, F.J. and Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 41*[th] *Annual Meeting of the Association for Computational Linguistics.*

Kroon, R.W. (2007). *Improving phrase-based translation*. Master's thesis, University of Twente.

Lambert, P. and Banchs, R. (2006). Grouping multi-word expressions according to part-of-speech in statistical machine translation. *Proceedings of the Workshop on Multi-word-expressions in a multilingual context*:9-16.

Laporte, E. and Voyatzi, S. (2008). An electronic dictionary of French multiword adverbs. *Proceedings of the LRECWorkshop Towards a Shared Task for Multiword Expressions (MWE 2008),* LREC: 31-34.

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.

Monteleone, M., Elia, A., de Bueriis, G. and Di Maio, F. (2008). Le polirematiche dell'italiano. Elia, A., de Bueriis, G. (a cura di), *Lessici elettronici e descrizioni semantiche, sintattiche e morfologiche. Risultati del Progetto PRIN 2005 Atlanti Tematici Informatici - ALTI,* Collana "Lessici & Combinatorie", n. 2, Dipartimento di Scienze della Comunicazione dell'Università degli Studi di Salerno, Plectica, Salerno: 11-65.

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.

Monti, J. (forthcoming). La E-translation da Google a Second Life: le più recenti applicazioni di Traduzione Automatica online. *Atti del XIII Congresso della Società Linguistica Italiana*, Bulzoni editore, Roma.

Moszczyński, R. (2010). Towards a bilingual lexicon of information technology multiword units. *Proceedings of the XIV Euralex International Cong*ress. Leeuwarden, the Netherlands.

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.

Ranchhod, E. and Carvalho, P. (2006). *Expressoes Multipalavra - Questoes lexicais e sintacticas*. Material de ensino na Primeira Escola de Verao da Linguateca (UP, Porto, Portugal, 10-14 July 2006).
[http://www.linguateca.pt/escolaverao2006/Sintaxe/EDV2006SintaxeRanchhodCarvalho.pdf]

Rayson, P. Piao, S. Sharoff, S. Evert, S. and Villada Moirón B.. (2010). Multiword expressions: hard going or plain sailing? *Journal of Language Resources and Evaluation*. Lang Resources & Evaluation (2010) 44:1–5.

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.

Sag A., Baldwin T., Bond, F., Copestake, A., Flickinger, D. et al. (2002). Multiword Expressions: A Pain in the Neck for NLP? *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CI-CLING 2002).* Mexico City, Mexico: 1-15. [http://lingo.stanford.edu/pubs/WP-2001-03.pdf]

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.
[http://www.linguateca.pt/Repositorio/ScottBarreiro09FreeBMT_OL.pdf]

Silberztein, M. (1993). "Les groupes nominaux productifs et les noms composés lexicalisés", *Lingvisticae Investigationes* 17:2, Amsterdam/Philadelphia: Benjamins, pp. 405-426.

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

Sperber, D. and Wilson D. (1986). *Relevance: Communication and cognition*, Oxford University Press, Oxford.

Thurmair, G. (2004). Multilingual content processing. *Proceedings of 4thInternational Conference on Language Resources and Evaluation (LREC) 2004*, Lisbon, Portugal.

Tillmann, C.and Xia, F. (2003). A Phrase-Based Unigram Model for Statistical Machine Translation. *The 3rd Human Language Technology Conference (HLT/NAACL 2003),* Edmonton, Canada.

Váradi, T. (2006). Multiword Units in an MT *Lexicon. EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop on Multi-word expressions in a Multilingual Context,* Trento, Italy: 73-78 [http://www.aclweb.org/anthology-new/N/N03/N03-2036.pdf]

Vietri, S. (2004). *Lessico-grammatica dell'italiano. Metodi, descrizioni e applicazioni.* UTET, Torino.

Villavicencio, A., Bond, F., Korhonen, A. and McCarthy, D. (2005). Introduction to the special issue on multiword expressions: having a crack at a hard nut. *Journal of Computer Speech and Language Processing*, 19(4):365–377.

Zens, R., Och, F.J. and Ney, H. (2002). *Phrase*-based statistical machine translation. *Proceedings of KI '02,* Springer-Verlag, London, UK: 18-32.

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.