10.12871/CLICIT2014151

An English-Italian MWE dictionary

Johanna Monti
Dipartimento di Scienze
Umanistiche e Sociali
Università degli Studi di Sassari
Via Roma 151 - Sassari
[jmonti@uniss.it]

Abstract

English. The translation of Multiword Expressions (MWEs) requires the knowledge of the correct equivalent in the target language which is hardly ever the result of a literal translation. This paper is based on the assumption that the proper treatment of MWEs in Natural Language Processing (NLP) applications and in particular in Machine Translation and Translation technologies calls for a computational approach which must be, at least partially, knowledge-based, and in particular should be grounded on an explicit linguistic description of MWEs, both using an electronic dictionary and a set of rules. The hypothesis is that a linguistic approach can complement probabilistic methodologies to help identify and translate MWEs correctly since hand-crafted and linguistically-motivated resources, in the form of electronic dictionaries and local grammars, obtain accurate and reliable results for NLP purposes. The methodology adopted for this research work is based on (i) Nooj, an NLP environment which allows the development and testing of the linguistic resources, (ii) an electronic English-Italian MWE dictionary, (iii) a set of local grammars. The dictionary mainly consists of English phrasal verbs, support verb constructions, idiomatic expressions and collocations together with their translation in Italian and contains different types of MWE POS patterns.

Italiano. *La traduzione delle polirematiche richiede la conoscenza del corretto equivalente nella lingua di arrivo che raramente è il risultato di una traduzione letterale. Questo contributo si basa sul presupposto che il corretto trattamento delle polirematiche in applicazioni di Trattamento Automatico del Linguaggio (TAL) ed in particolare di Traduzione Automatica e nelle tecnologie per la traduzione, più in generale, richiede un approccio computazionale che deve essere, almeno in parte, basato su dati linguistici, ed in particolare su una descrizione linguistica esplicita delle polirematiche, mediante l'uso di un dizionario macchina ed un insieme di regole. L'ipotesi è che un approccio linguistico può integrare le metodologie statistico-probabilistiche per una corretta identificazione e traduzione delle polirematiche, poiché risorse linguistiche quali dizionari macchina e grammatiche locali ottengono risultati accurati per gli scopi del TAL. La metodologia adottata per questa ricerca si basa su (i) Nooj, un ambiente TAL che permette lo sviluppo e la sperimentazione di risorse linguistiche, (ii) un dizionario macchina Inglese-Italiano di polirematiche, (iii) un insieme di grammatiche locali. Il dizionario è costituito principalmente da verbi frasali, verbi supporto, espressioni idiomatiche e collocazioni inglesi e contiene diversi tipi di modelli di polirematiche nonché la loro traduzione in lingua italiana.*

1 Introduction

This paper presents a bilingual dictionary of MWEs from English to Italian. MWEs are a complex linguistic phenomenon, ranging from lexical units with a relatively high degree of internal variability to expressions that are frozen or semi-frozen. They are very frequent and productive word groups both in everyday languages and in languages for special purposes and are the result of human creativity which is not ruled by algorithmic processes, but by very complex processes which are not fully representable in a machine code since they are driven by flexibility and intuition. Their interpretation and translation sometimes present unex-

pected obstacles mainly because of inherent ambiguities, structural and lexical asymmetries between languages and, finally, cultural differences.

The identification, interpretation and translation of MWEs still represent open challenges, both from a theoretical and a practical point of view, in the field of Machine Translation and Translation technologies.

Empirical approaches bring interesting complementary robustness-oriented solutions but taken alone, they can hardly cope with this complex linguistic phenomenon for various reasons. For instance, statistical approaches fail to identify and process non high-frequent MWEs in texts or, on the contrary, they are not able to recognise strings of words as single meaning units, even if they are very frequent.

Furthermore, MWEs change continuously both in number and in internal structure with idiosyncratic morphological, syntactic, semantic, pragmatic and translational behaviours.

The main assumption of this paper is that the proper treatment of MWEs in NLP applications calls for a computational approach which must be, at least partially, knowledge-based, and in particular should be grounded on an explicit linguistic description of MWEs, both using a dictionary and a set of rules.

The methodology adopted for this research work is based on: (I) Nooj an NLP environment which allows the development and testing of the linguistic resources, (ii) an electronic English-Italian (E-I) MWE dictionary, based on an accurate linguistic description that accounts for different types of MWEs and their semantic properties by means of well-defined steps: identification, interpretation, disambiguation and finally application, (iii) a set of local grammars.

2 Related work

The current theoretical work on this topic deals with different formalisms and techniques relevant for MWE processing in MT as well as other translation applications such as automatic recognition of MWEs in a monolingual or bilingual setting, alignment and paraphrasing methodologies, development, features and usefulness of handcrafted monolingual and bilingual linguistic resources and grammars and the use of MWEs in Statistical Machine Translation (SMT) domain adaptation, as well as empirical work concerning their modelling accuracy and descriptive adequacy across various language pairs.

The importance of the correct processing of MWEs in MT and Computer-aided translation (CAT) tools has been stressed by several authors. Thurmair (2004) underlines how translating MWEs word-by-word destroys their original meanings. Villavicenzio et al. (2005) underline how MT systems must recognise MWEs in order to preserve meaning and produce accurate translations. Váradi (2006) highlights how MWEs significantly contribute to the robustness of MT systems since they reduce ambiguity in word-for-word MT matching and proposes the use of local grammars to capture the productive regularity of MWEs. Hurskainen (2008) states that the main translation problems in MT are linked to MWEs. Rayson et al. (2010) underline the need for a deeper understanding of the structural and semantic properties of MWEs in order to develop more efficient algorithms.

Different solutions have been proposed in order to guarantee proper handling of MWEs in an MT process. Diaconescu (2004) stresses the difficulties of MWE processing in MT and proposes a method based on Generative Dependency Grammars with features. Lambert & Banchs (2006) suggest a strategy for identifying and using MWEs in SMT, based on grouping bilingual MWEs before performing statistical alignment. Moszczyński (2010) explores the potential benefits of creating specialised MWE lexica for translation and localisation applications.

Recently, increasing attention has been paid to MWE processing in MT and translation technologies and one of the latest initiatives in this research area is the MUMTTT workshop series specifically devoted to“ Multiword Units in Machine Translation and Translation Technology” (Monti & al. 2013). Finally, experiments in incorporating MWEs information in SMT have been carried out by Parra et al. (2014), who add compound lists to training sets in SMT, Kordoni & Simova (2014), who integrate phrasal verb information in a phrase-based SMT system, and finally Cholakov & Kordoni (2014), who use a linguistically informed method for integrating phrasal verbs into SMT systems. Automatic and manual evaluations of the results of these experiments show improvements in MT quality.

3 NooJ: an NLP environment for the development and testing of MWE linguistic resources

NooJ is a freeware linguistic-engineering development platform used to develop large-coverage formalised descriptions of natural languages and apply them to large corpora, in real time (Silberztein, 2002).
The knowledge bases used by this tool are: electronic dictionaries (simple words, MWEs and frozen expressions) and grammars represented by organised sets of graphs to formalise various linguistic aspects such as semi-frozen phenomena (local grammars), syntax (grammars for phrases and full sentences) and semantics (named entity recognition, transformational analysis). NooJ's linguistic engine includes several computational devices used both to formalise linguistic phenomena and parse texts such as: (i) Recursive Transition Networks (RTNs), (ii) Enhanced Recursive Transition Networks (ERTNs), (iii) Regular Expressions (RegExs) and finally (IV) Context-Free Grammars (CFGs in general).
NooJ is a tool that is particularly suitable for processing different types of MWEs and several experiments have already been carried out in this area: for instance, Machonis (2007 and 2008), Anastasiadis, Papadopoulou & Gavriilidou (2011), Aoughlis (2011). These are only a few examples of the various analysis performed in the last few years on MWE using NooJ as an NLP development and testing environment.

4 The Dictionary of English-Italian MWEs

The translation of MWEs requires the knowledge of the correct equivalent in the target language which is hardly ever the result of a literal translation. Given their arbitrariness, MT and Translation technologies have to rely on the availability of ready solutions in the source and target language in order to perform an accurate translation process.
The English-Italian MWE dictionary is the result of a contrastive English-Italian analysis of continuous and discontinuous MWEs with different degrees of variability of co-occurrence among words and different syntactic structures, carried out during the development and testing of the English-Italian language pair for Logos, a rule-based MT system, and subsequently further developed in the framework of the Lexicon-Grammar (LG) formalism (Monti, 2012).
The dictionary is based on the LG approach to MWEs (Gross, 1986), where these complex and varied linguistic phenomena are described according to a flat structure composed of the POS tags of the MWE elements and their sequence. Furthermore, according to this approach it is possible to distinguish fixed MWEs and MWEs that allow syntactic variations, such as the insertion of other elements or the variation of one or more elements. Green et al. (2011) adopt a similar approach for the MWE description and show the usefulness of this model for several NLP tasks in which MWE pre-grouping has improved accuracy.
The E-I MWE dictionary contains over 10,000 entries and is used to represent and recognise various types of MWEs. Each entry of the dictionary is given a coherent linguistic description consisting of: (i) the grammatical category for each constituent of the MWE: noun (N), Verb (V), adjective (A), preposition (PREP), determiner (DET), adverb (ADV), conjunction (CONJ); (ii) one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to nominalise them), preceded by the tag +FLX; (iii) one or more syntactic properties (e.g. "+transitive" or +N0VN1PREPN2); (iv) one or more semantic properties (e.g. distributional classes such as "+Human", domain classes such as "+Politics"); (v) the translation into Italian.
The dictionary contains different types of MWE POS patterns. The main part of the dictionary consists of English phrasal verbs, support verb constructions, idiomatic expressions and collocations together with their Italian translations.

Intransitive Verbs:

[VIntrans+ADJ]
lie,V+FLX=LIE+JM+FXC+Intrans+ADJ="flat"+IT ="sdraiarsi"

[VIntrans+PART]
bear,V+FLX=BEAR+JM+FXC+Intrans+PART= "down"+IT="avanzare"

[VIntrans+PART+PREP+N2]
break,V+FLX=SPEAK+JM+FXC+Intrans+ PART="off"+PREP="from"+N2="work"+IT "interrompere il lavoro"

[VIntrans+PART+PREP+Ving]
break,V+FLX=SPEAK+JM+FXC+Intrans+ PART= "off"+PREP="from"+VG+IT="smettere di Vinf"

[VIntrans+PREP+N2]
account,V+FLX=ASK+JM+FXC+Intrans+PREP

=“for”+N2+IT=“spiegare N2”

Transitive Verbs:

[VTrans+N1]
advance,V+FLX=LIVE+JM+FXC+Trans+N1= “reason”+IT=“esporre N1”

[VTrans+ADJ+N1]
break,V+FLX=SPEAK+JM+FXC+Trans+N1+ ADJ= "free”+IT=“liberare N1”

[VTrans+PART+N1]
bring,V+FLX=BRING+JM+FXC+Trans+PART= “up”+N1=“question”+IT=“sollevare N1(problema)”

[VTrans+PART+N1+PREP+N2]
bring,V+FLX=BRING+JM+FXC+Trans+PART=“back”+N1+PREP=“from”+N2=“memory”+IT= “richiamare a N2(mente)“

[VTrans +N1+PREP+N2]
break,V+FLX=SPEAK+JM+FXC+Trans+N1= news”+PREP=“to”+N2Hum+IT=“comunicare N1 a N2”

[VTrans+N1+PREP+Ving]
bar,V+FLX=ADMIT+JM+FXC+Trans+N1Hum+PREP=“from”+VG+IT=“impedire a N1 di Vinf”

5 Grammars

Syntactic or semantic grammars (.nog files) are used to recognise and annotate expressions in texts, e.g. to tag noun phrases, certain syntactic constructs or idiomatic expressions, extract certain expressions (name of companies, expressions of dates, addresses, etc.), or disambiguate words by filtering out some lexical or syntactic annotations in the text.
These grammars recognise different types of MWEs, such as frozen and semi-frozen units, and are particularly useful with discontinuous MWEs (Machonis, 2008 and Silberztein, 2008).
It is possible: (i) to identify MWEs of different types in texts by means of specific local grammars, (ii) annotate texts with the corresponding translations of the identified MWEs, (iii) export the annotated texts in XML.
Annotated texts can be used in this way for instance for SMT training purposes.
Once texts are annotated, they can be exported as XML files, like in the following example:

He <EXPV TYPE="JM" IT="rinunciare a"> abandons</EXPV> the <EXPN IT="appello"> appeal</EXPN>.

He <EXPV TYPE="JM" IT="rinunciare a">abandons</EXPV> the <EXPN IT="speranza"> hope</EXPN>.

He <EXPV TYPE="JM" IT="acquisire ">acquires</EXPV> a <EXPN IT="conoscenza"> knowledge</EXPN> of the specific domain.

6 Future work

For future work, we plan to further investigate MWEs in particular with respect to cross-linguistic asymmetries and translational equivalences.
Our final goal is to integrate MWE treatment in either data-driven or hybrid approaches to MT in order to achieve high quality translation by combining probabilistic and linguistic information.
However, to achieve this goal, we must devise efficient strategies for representing deep attributes and semantic properties for MWEs in a cross-linguistic perspective.

7 Conclusion

In conclusion, the focus of this research for the coming years will be to improve the results obtained so far and to extend the research work to provide a more comprehensive methodology for MWE processing in MT and translation technologies, taking into account not only the analysis phase but also the generation one.
This experiment provides, on the one hand, an investigation of a broad variety of combinations of MWE types and an exemplification of their behaviour in texts extracted from different corpora and, on the other hand, a representation method that foresees the interaction of an electronic dictionary and a set of local grammars to efficiently handle different types of MWEs and their properties in MT as well as in other types of NLP applications.
This research work has therefore produced two main results in the field of MWE processing so far:

- the development of a first version of an English-Italian electronic dictionary, specifically devoted to different MWEs types,
- the analysis of a first set of specific MWE structures from a semanto-syntactic point of view and the development of local grammars for the identification of continuous and discontinuous MWEs in the form of FST/FSA.

References

Anastasiadis, M., Papadopoulou, L., & Gavrilidou, Z. 2011. Processing Greek frozen expressions with Nooj. K. Vučković, B. Bekavac, & M. Silberztein (eds). *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik)*. Newcastle: Cambridge Scholars Publishing.

Aoughlis, F. 2011. A French-English MT system for Computer Science Compound Words. K. Vučković, B. Bekavac, & M. Silberztein (eds). *Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 International Conference (Dubrovnik)*. Newcastle: Cambridge Scholars Publishing.

Cholakov, K. & Kordoni, V. 2014. Better Statistical Machine Translation through Linguistic Treatment of Phrasal Verbs. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) 2014*, Association for Computational Linguistics: 196-201. http://aclweb.org/anthology/D14-1024

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

Green, S., de Marneffe, M. C., Bauer, J., & Manning, C. D. 2011. Multiword expression identification with tree substitution grammars: A parsing tour de force with French. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: 725-735.

Gross, M. 1986. Lexicon-Grammar: the representation of compound words. *Proceedings of COLING '86*. Bonn: University of Bonn, http://acl.ldc.upenn.edu/C/C86/C86-1001.pdf.

Hurskainen, A. 2008. *Multiword Expressions and Machine Translation.* Technical Reports. Language Technology Report No 1.

Kordoni, V. & Simova I. 2014. Multiword expressions in Machine Translation. *LREC 2014 Proceedings*.

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

Machonis, P. A. 2007. Look this up and try it out: an original approach to parsing phrasal verbs. *Actes du 26 Colloque international Lexique Grammaire, Bonifacio 2-6 octobre 2007.*

Machonis, P. A. 2008. NooJ: a practical method for Parsing Phrasal Verbs. *Proceedings of the 2007 International NooJ Conference*. Newcastle: Cambridge Scholars Publishing: 149-161.

Monti, J. 2012. *Multi-word Unit Processing in Machine Translation. Developing and using language resources for multi-word unit processing in Machine Translation* – PhD dissertation in Computational Linguistics- Università degli Studi di Salerno.

Monti J, Mitkov R, Corpas Pastor G, Seretan V (eds). 2013. *MT Summit workshop proceedings for: Multiword Units in Machine Translation and Translation Technologies (Organised at the 14th Machine Translation Summit)*. CH-4123 Allschwil: The European Association for Machine Translation.

Moszczyński, R. 2007. A Practical Classification of Multiword Expressions. *ACL '07 Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*. Association for Computational Linguistics.

Parra Escartín, C., Peitz, S., and Ney, H. 2014. German Compounds and Statistical Machine Translation. Can they get along? *EACL 2014, Tenth Workshop on Multiword Expressions (MWE 2014)*, Gothenburg, Sweden, April 2014: 48-56.

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

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

Silberztein, M. 2008. Complex Annotations with NooJ. X. *Proceedings of the 2007 International NooJ Conference*, Jun 2007, Barcelone, Spain. Cambridge Scholars Publishing. <hal-00498042>

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

Váradi, T. 2006. Multiword Units in an MT Lexicon. *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts* Trento, Italy: Association for Computational Linguistics: 73-78.

Villavicencio, A. B. 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.