A KNOWLEDGE-BASED APPROACH TO MULTIWORDS PROCESSING IN MACHINE TRANSLATION: THE ENGLISH-ITALIAN DICTIONARY OF MULTIWORDS

COST Action IC1207 PARSEME meeting, 10–11 March 2014, Athens
WG1: Lexicon/Grammar Interface or WG3: Hybrid Parsing of MWEs
POSTER PROPOSAL

Johanna Monti – Department of Human and Social Sciences University of Sassari - Italy

This poster presents a knowledge-based approach to the identification and translation of multiword expressions (MWEs) from English to Italian. The main assumption of the methodology proposed is that the proper treatment of MWEs in MT 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.

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 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 mainly based on the following elements:

an NLP environment which allows the development and testing of the linguistic resources.

an electronic 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.

a set of local grammars

We will provide details about the methodology that can be applied to the identification and translation of MWEs.

1. 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.

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 FSTs, FSAs, Recursive Transition Networks (RTNs),[1] Enhanced Recursive Transition Networks (ERTNs),[2] Regular Expressions (RegExs),[3] Context Free Grammars

1 **Recursive Transition Networks (RTNs)** are grammars that contain more than one graph; graphs can be FST or FSA, and also include references to other embedded graphs; these latter graphs may in turn contain other references to the same or to other graphs. Generally, RTNs are used in NooJ to build libraries of graphs from the bottom-up: simple graphs are designed; they are then re-used in more general graphs; these are in turn re-used, etc.

2 **Enhanced Recursive Transition Networks (ERTNs)** are RTNs that contain variables; these variables typically store parts of the matching sequences and are then used to perform operations with them (e.g. put their content in the plural, etc.), and then produce the resulting output. Because variables can be duplicated, inserted and/or displaced in the output, ERTNs give NooJ the power to perform linguistic transformations on texts. Examples of transformations include negation, passivisation, nominalisation, etc.

3 **Regular Expressions (RegExs)** represent a way to perform simple queries without having to build specific grammars. When the sequence to be located consists of a few words, it is much quicker to enter these words directly into a regular expression.

(CFGs).[4]
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) and finally Vietri (2008). 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.

2. The Dictionary of English-Italian MWEs

The EIMWE.dic is a dictionary used to represent and recognise various types of MWEs.

This dictionary is based on a contrastive English-Italian analysis of continuous and discontinuous MWEs with different degrees of variability of co-occurrence among word compositionality and different syntactic structures.

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 has to rely on the availability of ready solutions in both languages in order to perform an accurate translation process.

Each entry of the dictionary is given a coherent linguistic description consisting of:

the grammatical category for each constituent of the MWE: noun (N), Verb (V), adjective (A), preposition (PREP), determiner (DET), adverb (ADV), conjunction (CONJ);

one or more inflectional and/or derivational paradigms (e.g. how to conjugate verbs, how to nominalise them), preceded by the tag +FLX;

one or more syntactic properties (e.g. "+transitive" or +N0VN1PREPN2);

one or more semantic properties (e.g. distributional classes such as "+Human", domain classes such as "+Politics");

the translation into Italian.

The EIMWE.dic contains different types of MWE POS patterns. The main part of the dictionary consists of phrasal verbs, support verb constructions, idiomatic expressions and collocations. In the poster, the main verb structures are explained with examples extracted from the *British National Corpus,* from the Internet by means of the *WebCorp LSE* application or with our own examples together with the Italian translations. Finally, the corresponding dictionary entry for each example of MWE POS pattern is provided.

3. References

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

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.* (p. 149-161). Newcastle: Cambridge Scholars Publishing.

Vietri, S. (2008). *Dizionari elettronici e grammatiche a stati finiti. Metodi di analisi formale della lingua italiana.* Cava dei Tirreni : Plectica.

4 **Context-Free Grammars (CFGs in general)** constitute an alternative means to entering morphological or syntactic grammars. For instance, NooJ includes an inflectional/derivational module that is associated with its dictionaries so that it can automatically link dictionary entries with their corresponding forms in the corpora