Formalising Multilingual Verbal MultiWord Expressions to Support Neural Machine Translation

Maria Pia di Buono and **Johanna Monti**
UNIOR NLP Research Group
University of Naples L'Orientale
{mpdibuono,jmonti}@unior.it

Relevant UniDive working groups: WG2, WG3

Abstract

Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014) is producing better translations over previous Machine Translation (MT) approaches such as Rule-based (RBMT) and Phrase-based Translation (PBMT) in recent years.
However, multiword expressions (MWE) still represent a critical area for MT, even for current neural approaches (Isabelle et al., 2017): their correct identification and meaning-preserving translation remain challenging due to the idiosyncratic properties of this complex lexemes.
Furthermore, the need of reusable, interoperable and interlinked linguistic resources in Natural Language Processing (NLP) downstream tasks has been proved by the increasing efforts to develop standards for the representation of different layers of information (e.g., syntax, morphology) of various phenomena, particularly MWEs (Moreno et al., 2003; Copestake et al., 2002; Francopoulo et al., 2006, 2009).
Nevertheless, despite those efforts, the achievement of a full processing of complex phraseology is still far away to be reached, mainly due to the lack of compatibility for metadata in linguistic resource production (Calzolari et al., 2011).
Starting from the analysis of the translation issues in NMT concerning Italian clitic verbal MWEs, such as *prendersela* 'to be offended', *buttarsi giù* 'to get depressed', *lasciarsi andare* 'to relax', the main goal of this contribution is to propose the modeling of a resource for this language-specific category according to the OntoLex-Lemon model[1] as Semantic Web technologies might enhance the quality of machine translation outputs (Moussallem et al., 2018).
Clitic VMWEs are here understood as a varied class made up by a verb with at least one deficient pronoun (Cardinaletti and Starke, 1999) whose semantics is not derivable from its constituent parts. The class of Italian pronouns lexicalised as verbal affixes is made up of *-si* pronouns, traditionally referred to as "reflexive pronouns" (*mi, ti, si, ci, vi, si*), the third person feminine of clitic complement pronouns (*la, le*) and two fixed pronominal particles: *ci* and *ne*, generally having locative or demonstrative meaning. These forms can also combine in six different clitic clusters: *cela, cele, cene, sela, sele, sene*.
For the sake of the current analysis we take into account only clitic verbs which either (i.) do not exist without the clitic, (ii.) have a different meaning with respect to the non-clitic verb form or (iii.) or are constitutive elements of more complex idiomatic expressions.
In order to assess if and how different categories give rise to specific translation issues we selected some examples for each identified category and evaluated the translation into English by Google Translate (GT), considering three forms for each verb: out-of-context infinitive form (OCIF), out-of-context past form (OCPF) , and in the context forms (ICF) of example sentences.
Verbs belonging to pronominal intransitive verbs, such as *arrabbiarsi*, *suicidarsi*, *pentirsi*, *vergognarsi*, which do not exist without the clitic, do not present specific translation issues both in the out-of-context and the in-context settings. If we take the verb *pentirsi* 'to regret' as an example for this category, GT translates correctly into English the infinitive (*pentirsi* → to repent) and past forms (*Mi sono pentito* → I repented) and its meaning in the context of the following sentence: *Si è pentito di non aver proseguito gli studi.* → He regretted not continuing his studies. In addition, GT is able to understand the shades of meaning which the verb can take, namely "feel remorse and repentance for having transgressed a moral or religious law" (→ to repent) or "feel regret for having or not having done an action" (→ to regret).
Clitic VMWEs can also be part of idiomatic expressions and therefore besides the clitic pronoun or clitic clusters, they are accompanied by further

[1] https://www.w3.org/community/ontolex/

lexicalised components, such as nouns, adjectives, like in *darsela a gambe* 'cut and run', *farla finita* 'to end it', *darci dentro* 'knuckle down', *prenderci gusto* 'get a taste for'. They are generally non-compositional, i.e., none of their lexicalized components retains any of their original meanings, this is why neural approaches seem to be unable to process them properly. Tests show that GT does not grasp the idiomatic meaning of the VMWEs as in the sentences with *darci dentro* and *farla finita*.

Considering the aforementioned translation issues and in order to support NLP and MT tasks, we proceed with the modelisation of Italian clitic VMWEs, proposing a three-level representation of linguistic features, namely, morpho-syntactic, syntactic and semantic levels, and define different sets of rules for each of them.

In order to proceed with the modelisation, firstly we collect the data from De Mauro (1999) with reference to two VMWE types:

- procomplement verbs, namely verbs which are regularly used with with procomplementary clitic particles, as *svignarsela* (sneak away) or which, when used with such particles, assumes specific values, independent of the base verb, such as *sentirsela* 'feel up to i)' which differs from *sentire* 'feel'.

- verbal idioms, namely verbs which have at least two lexicalized components including a head verb and at least one of its dependents[2].

To these two typologies we add Inherently reflexive verbs (IRVs) extracted from the annotated PARSEME-IT corpus (Monti and di Buono, 2019).The reason behind this choice is due to the fact that De Mauro does not distinguish among different types of intransitive pronominal verbs, e.g., reflexive.

In total, our resource is composed of 663 entries of which 411 entries have been extracted from De Mauro (1999) and 252 IRVs from PARSEME-IT corpus and manually checked (Table 1).

After collecting the data, we proceed with a bottom-up approach to describe morpho-syntactic properties of clitic VMWEs. In order to define the properties useful to model clitic VMWEs, firstly we classify different types of clitics, namely: (i) direct object pronouns, e.g., *la*; (ii) -si pronouns

[2]https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=050_Cross_lingual_tests/030_Verbal_idioms_LBVIDRB

Type	Source	#VMWEs
Procomplement	De Mauro	172
Idiom	De Mauro	239
IRV	PARSEME-IT	252
Total		**663**

Table 1: VMWE Types, source, and number of occurrences

with different functions (Jezek, 2005), e.g., *si*, *mi*; (iii) fixed pronominal particle, e.g., *ci*, *ne*. At the morpho-syntactic level, we consider the inflection constraints of clitic elements co-occurring with the verb and contributing to produce different senses. For each of them, we describe their morpho-syntactic features to be encoded in a set of morpho-syntactic properties (MSPs). MSPs generalise the behaviour of clitic elements, accounting for both inflection constraints of clitic elements (i.e., fixed number and gender) and order in clitic clusters. With reference to syntactic features, syntactic properties (SyP) are described on the basis of VMWE patterns to formalise other co-occurring elements, which may be mandatory in some cases, e.g., *darsela a gambe* 'to beat it'. We develop 50 SyP descriptions in total. It is worth stressing that SyP descriptions refer to syntactic functions of VMWE elements, thus, for instance, we describe the pronominal particle *la* as a direct object even though it does not represent a regular anaphoric/cataphoric reference.

Finally, as far as semantic description (SD) is concerned, we describe different types of VMWEs, according to the morpho-syntactic and syntactic features previously identified. In total, 55 SDs have been created. Such features contribute to deriving semantic information useful to link lexical entries to specific OntoLex-Lemon senses.

To model all the linguistic information related to clitic VMWEs and to develop a bilingual (IT-EN) OntoLex-Lemon resource useful for supporting MT applications, we apply different elements from the OntoLex-Lemon modules. We rely on the OntoLex-Lemon model and its modules, mainly the Lexicographic module (Bosque-Gil et al., 2016), to address the representation of different senses through the description of different entries, encoded as lexical entries and supplied with form restrictions and usage examples.

We apply different OntoLex-Lemon elements to model clitic verbs and their morpho-syntactic

behaviours together with their semantics so that the morphological description is directly linked to the lexical entry modelled in the lexicographic module. We also use OntoLex-Lemon core elements to describe syntactic patterns of VMWEs in which clitics occur and semantic aspects related to each type of clitic verbs.
Finally, terminological variants and translations are specified to support the development of the lexicographic resource.

Acknowledgment

This work is based on Monti, J., di Buono, M.P., Caruso, V. (*forthcoming*) Ontology-based formalisation of Italian Clitic Verbal MWEs: an Approach for Supporting Machine Translation in Mitkov, R., Corpas Pastor, G., and Monti, J. (eds.) John Benjamin Publishing Company

References

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda, and Guadalupe Aguado-de Cea. 2016. Modelling multilingual lexicographic resources for the web of data: The k dictionaries case. In *GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme*, page 65.

Nicoletta Calzolari, Nuria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi, and Claudia Soria. 2011. Final flarenet deliverable: Language resources for the future-the future of language resources. *The Strategic Language Resource Agenda. FLaReNet project*.

Anna Cardinaletti and Michal Starke. 1999. The typology of structural deficiency. *Clitics in the languages of Europe*, pages 145–233.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Ann A Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A Sag, and Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *LREC*.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for nlp in the lexical markup framework (lmf). *Language Resources and Evaluation*, 43:57–70.

Gil Francopoulo, Thierry Declerck, Monica Monachini, and Laurent Romary. 2006. The relevance of standards for research infrastructures. In *International Conference on Language Resources and Evaluation-LREC 2006*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Elisabetta Jezek. 2005. Interazioni tra aspetto e diatesi nei verbi pronominali italiani. *Studi di grammatica italiana*, (23):239–281.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Johanna Monti and Maria Pia di Buono. 2019. Parsemeit: an italian corpus annotated with verbal multiword expressions. *IJCoL. Italian Journal of Computational Linguistics*, 5(5-2):61–93.

Antonio Moreno, Susana López, Fernando Sánchez, and Ralph Grishman. 2003. Developing a Spanish treebank. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, pages 149–163. Kluwer.

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.