Spanish Multiword Expressions: typology and structural analysis

Carla Parra Escartín[1], Almudena Nevado Llopis[2], Eoghan Sánchez Martínez[2], María Pilar Cardos Murillo[2]

[1]University of Bergen, Bergen (Norway), [2]San Jorge University, Zaragoza (Spain)

In this paper, we will analyze Spanish multiword expressions (MWEs) and describe their linguistic properties. As a starting point of our study, we take the MWE taxonomy proposed by Ramisch (2012). He distinguishes three morphosyntactic classes and three additional so-called difficulty classes. The three morphosyntactic classes are nominal expressions, verbal expressions and adverbial and adjectival expressions. Nominal expressions are further subdivided in noun compounds, proper names and multiword terms. Verbal expressions are further subdivided in phrasal verbs and light verb constructions. The difficulty classes are divided in fixed expressions, idiomatic expressions, and "true" collocations.

Table 1 below exemplifies all of these categories with Spanish examples and their translations into English.

MWE type			Spanish
Morphosyntactic classes	Nominal expressions	Noun compounds	*Sacacorchos* (bottle opener); *ruleta rusa* (Russian roulette)
		Proper names	*Nueva York* (New York), *Unión Europea* (European Union), Barack Obama
		Multiword terms	*Cuenta de resultados* (profit and loss account); *infarto de miocardio* (myocardial infarction)
	Verbal expressions	Phrasal verbs	*Acordarse de X* (remember X); *atenerse a X* (stick to); creer en (believe in)
		Light verb constructions	*Tener fe* (have faith); *hacer una foto* (take a picture); *dar un paseo* (go for a walk)
	Adverbial and adjectival expressions		*Más o menos* (more or less); *en líneas generales* (by and large)
Difficulty classes	Fixed expressions		*Ad hoc*; *en lo que respecta a* (with regard to)
	Idiomatic expressions		*Estirar la pata* (kick the

		bucket); *poner la antena* (listen without being invited to); *ponerse las pilas* (to get one's act together); *cargar las pilas* (to recharge one's batteries)
	"True" collocations	*Escribir una carta* (write a letter); *firmar un acuerdo* (sign an agreement)

Table 1: Typology of MWEs and examples in Spanish

The classes distinguished in the typology will be further analyzed and classified by their degree of difficulty for Natural Language Processing (NLP) purposes. To this aim, we will study whether the strategies proposed in the seminal papers by Sag et al. (2002) and Baldwin and Kim (2010) hold for Spanish.

The starting point of our study will be a list of Spanish MWEs classified using the typology proposed by Ramisch (2012) and further analyzed and classified as fixed, semi-fixed and fixed expressions, as proposed by Baldwin and Kim (2010). The linguistic analysis of the data will allow us to describe the specific linguistic features of Spanish MWEs and discuss the challenging cases.

As Spanish is a rich morphological language, we foresee the need for a deeper linguistic analysis than in other languages such as English. A similar observation was made in Savary (2008) and Graliński et al. (2010). These authors already studied the complexity of encoding MWEs in morphologically rich languages such as Polish and French. Testing the formalisms they propose is beyond the scope of this article. However, we will take into consideration their findings, as Spanish MWEs may require a similar treatment to that of French. We expect that our analysis can also give light to the MWE treatment in other Romance languages such as French and Italian, which also have a rich morphology.

Bibliography

Baldwin, Timothy and Su Nam Kim (2010) *Multiword Expressions*, in Nitin Indurkhya and Fred J. Damerau (eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267-292.

Graliński, Filip and Savary, Agata and Czerepowicka, Monika and Makowiecki, Filip. 2010. *Computational Lexicography of Multi-Word Units: How Efficient Can It Be?* In: Proceedings of the Multiword Expressions: From Theory to Applications (MWE 2010), pages 2–10, Beijing, August 2010.

Ramisch, Carlos. 2012. *A generic and open framework for multiword expressions treatment: from acquisition to applications.* University of Grenoble (France) and Federal University of Rio Grande do Sul (Brazil)

Sag, Ivan A. and Baldwin, Timothy and Bond, Francis and Copestake, Ann and Flickinger, Dan. 2002. *Multiword Expressions: A Pain in the Neck for NLP*. In Proc.

of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002).

Savary, Agata. 2008. *Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches*. Linguistic Issues in Language Technology, 1(2):1–53.