Procesamiento del Lenguaje Natural, Revista nº 64, marzo de 2020, pp. 123-126 recibido 04-11-2019 revisado 11-01-2020 aceptado 12-01-2020

Identification and translation of verb+noun Multiword Expressions: A Spanish-Basque study

Identificación y traducción de Expresiones Multipalabra de tipo verbo+sustantivo: análisis de castellano-euskera

Uxoa Inurrieta
Ixa NLP group, University of the Basque Country (UPV/EHU)
usoa.inurrieta@ehu.eus

Abstract: This is a summary of the PhD thesis written by Uxoa Iñurrieta under the supervision of Dr. Gorka Labaka and Dr. Itziar Aduriz. Full title of the PhD thesis in Basque: *Izena+aditza Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala.* The defense was held in San Sebastian on November 29, 2019. The doctoral committee was integrated by Ricardo Etxepare (Centre National de la Recherche Scientifique), Margarita Alonso (Universidad de Coruña) and Miren Azkarate (University of the Basque Country).
Keywords: Multiword Expressions, Phraseology, Identification, Machine Translation, Basque

Resumen: Este es un resumen de la tesis doctoral escrita por Uxoa Iñurrieta bajo la supervisión del Dr. Gorka Labaka y la Dra. Itziar Aduriz. Título completo de la tesis en euskera: *Izena+aditza Unitate Fraseologikoak gaztelaniatik euskarara: azterketa eta tratamendu konputazionala.* La defensa de la tesis se celebró en Donostia-San Sebastián el 29 de Noviembre de 2019, ante el tribunal formado por Ricardo Etxepare (Centre National de la Recherche Scientifique), Margarita Alonso (Universidad de Coruña) y Miren Azkarate (UPV/EHU).
Palabras clave: Expresiones Multipalabra, Fraseología, Identificación, Traducción Automática, Euskera

1 Motivation

Multiword Expressions (MWEs) are combinations of words which exhibit some kind of lexical, morphosyntactic, semantic, pragmatic or statistical idiosyncrasy (Baldwin and Kim, 2010). Due to their idiosyncratic nature, MWEs pose important challenges to Natural Language Processing (NLP), and sophisticated strategies are needed in order to process them correctly (Constant et al., 2017). Two main types of word combinations are comprised in the category of MWEs: idioms (example 1) and collocations (example 2), the latter including light verb constructions (example 3). All of them are considered in this work.

(1) *She always ends up* ***spilling the beans****.* (lit. revealing the secret)

(2) *All students* ***passed*** *the* ***exam****.*

(3) *She is* ***giving*** *a* ***lecture*** *this afternoon.*

In this PhD, two tasks concerning MWE processing are addressed: on the one hand, their identification in corpora, and on the other hand, their processing within Machine Translation (MT).

Automatic identification of MWEs involves finding occurrences of previously known MWEs (Ramisch et al., 2018). For instance, considering the MWE *make conclusions*, occurrences would need to be identified in examples 4a–c, but not in 4d. Likewise, for the MWE *pull somebody's leg*, an identification system would need to distinguish MWE occurrences like the one in example 5a from non-MWEs like the one in example 5b. In order to do so, it must be taken into account that many MWEs, especially verbal ones, tend to be morphosyntactically variable, discontiguous, and ambiguous depending on the context.

(4) a. *They* ***made*** *a* ***conclusion****.*
b. *They* ***made*** *some simple but*

ISSN 1135-5948. DOI 10.26342/2020-64-15

interesting ***conclusions****.*

c. *The* ***conclusions*** *they* ***make*** *are always interesting.*

d. *They will make progress and will come to a conclusion.*

(5) a. *She is not serious. She is just* ***pulling your leg****.*

b. *Grab your knee, pulling your leg toward your chest.*

Concerning Machine Translation, apart from identifying MWEs in the source text, an appropriate equivalent must be given to them in the target language. This poses additional challenges, since word co-occurrence is often arbitrary (example 6) and some MWEs are non-compositional (example 7), meaning that word-for-word translations are incorrect in many cases.

(6) EN: ***pay attention***
ES: ***prestar atención***
(lit. lend attention)
EU: ***arreta jarri***
(lit. put attention)

(7) EN: ***pull*** *somebody's* ***leg***
ES: ***tomar el pelo*** *a alguien*
(lit. take sb's hair)
EU: *norbaiti* ***adarra jo***
(lit. play the horn to sb)

The main assumption behind the work in this PhD is that specific linguistic information (mainly lexical and morphosyntactic data) is helpful for MWE identification and translation of MWEs within MT. An in-depth linguistic analysis and several experiments were undertaken to verify this, which are outlined in Section 2. Then, in Section 3, the hypotheses covered in the PhD are listed and the main contributions are briefly described.

2 General outline of the dissertation

The contents of the dissertation can be divided into five parts. A brief overview will now be given, and the papers linked to each of the parts will be specified when available.

- Preanalysis (Iñurrieta et al., 2018a). A study based on a Spanish-Basque bilingual dictionary was undertaken, where verb+noun entries and their translations were analysed along lexical and morphosyntactic dimensions. Phraseological differences and similarities of both languages were made evident, as well as the prevalence of non-word-for-word translations concerning verb+noun MWEs.
- Manual analysis (Iñurrieta et al., 2016; Iñurrieta et al., 2017). An in-depth manual analysis of a set of Spanish MWEs and their Basque translations was carried out. Lexical and morphosyntactic data were considered, which were then employed in two experiments to see how they affected identification and MT. Results were positive in both cases.
- (Semi)automatic analysis. Considering the good results of the experiments using manual data, but taking into account that fully manual studies are limited and expensive, an improved analysis method was proposed to extract linguistic data about MWEs and their translations from corpora. Part of the information was then manually corrected to evaluate the different steps of the method. Finally, two more experiments were undertaken, where the positive impact of the gathered data was confirmed for identification and, to a lesser extent, for MT.
- The *Konbitzul* database (Iñurrieta et al., 2018b). All the linguistic data gathered from the previous studies were collected in a publicly accessible database: *Konbitzul*. Information can be either queried online or downloaded to be used for NLP purposes.
- PARSEME-related work (Iñurrieta et al., 2018; Savary et al., 2019). Verbal MWEs were annotated in a Basque corpus, which was integrated into the PARSEME multilingual corpus. A study of literal occurrences of MWEs was then undertaken based on the previously annotated corpus.

3 Hypotheses and contributions

Several hypotheses were proposed and validated through the work outlined in the previous section. These hypotheses are listed below (Section 3.1), and the main contributions of the research undertaken are then summarised (Section 3.2).

3.1 Hypotheses

Six hypotheses were tested and confirmed in this PhD, four of which were mostly related to linguistic aspects of MWEs, and the remaining two, to MWE processing. All six hypotheses are listed below. A summary of the main evidence found to support them can be found in the English version of the PhD thesis (Inurrieta, 2019, pp. 12–14).

[**H1**] Many MWEs are not translated word-for-word from one language to another.

[**H2**] Verbal MWEs tend to be rather flexible concerning morphosyntax, although not completely, since they also have some restrictions.

[**H3**] Compared to many other languages which have been analysed from a phraseological perspective, light verb constructions are especially frequent in Basque.

[**H4**] Although many word combinations can be idiomatic or literal depending on the context, very few of them are actually used literally in real texts.

[**H5**] Detailed morphosyntactic information is helpful for MWE identification.

[**H6**] MWE-specific linguistic information is beneficial for MT.

3.2 Contributions

Apart from confirming the hypotheses listed above, a number of contributions were made through the work in this PhD. The main ones are listed below.

[**C1**] **Comprehensive NLP-applicable study of verb+noun MWEs in Spanish and Basque.** Although there exist other studies on verb+noun MWEs in both languages, the one in this PhD differs from them in two main aspects. On the one hand, because it is NLP-oriented, unlike most of the phraseological analyses carried out for Spanish and Basque. On the other hand, because most of the data obtained from it is quantified, which is helpful to see the extent of the MWE-specific features under study. Furthermore, as will be shown in contribution 6, all data were made publicly available.

[**C2**] **Analysis of the translation of verb+noun MWEs between Spanish and Basque.** Almost no research has been undertaken about phraseology in Spanish-Basque translation, and this work brings a contribution into the field. The verb+noun entries and translations in the *Elhuyar* dictionary were firstly analysed, and translations were automatically extracted from parallel corpora for further MWEs from other sources. In both analyses, lexical and morphosyntactic features were examined, to see how these change when MWEs are translated from one language to the other.

[**C3**] **Proposal or application of methodologies which are adaptable to other languages.** The idea of replicability and reusability of our methods was present throughout the whole work. Firstly, in order to test whether the proposed analysis was applicable to other languages, the manual study of Spanish verb+noun MWEs was undertaken also in English. The output data was then used for an MWE identification experiment, where results were even better than the Spanish ones. Secondly, part of the method proposed to automatise the analysis of MWEs was reused on Basque corpora to gather translation-oriented information. Only a few modifications needed to be done, which means that it is easily adaptable. Thirdly, the PARSEME universal guidelines were followed to annotate verbal MWEs in a Basque corpus, just like in 19 other languages. And finally, the study of literal occurrences of MWEs was done in five languages of different phylogenetic families.

[**C4**] **Improvement of the identification of verb+noun MWEs.** The identification method proposed in this PhD outperforms all results in the Spanish part of the PARSEME shared task edition 1.1, with an F score of 0.51, which is 13 points higher than the best result in the Spanish task. Besides, it was made evident precisely what morphosyntactic features are helpful for identification.

[**C5**] **Integration of MWE-specific linguistic data into MT.** Lexical and morphosyntactic information specific to a set of verb+noun MWEs was added to the *Matxin* rule-based MT system, and results were better than the basic system both according to a manual evaluation (improvement of 62-65%) and according to statistical measures (increase of 2.25% in BLEU).

[**C6**] **Creation of a database collecting all MWEs and translations covered in this work, along with NLP-applicable**

linguistic data. The *Konbitzul* database is publicly accessible online[1]. Its interface enables users to make queries according to several criteria and filters, and all NLP-applicable information can be fully downloaded. In all, 1,927 Spanish MWEs (along with 4,043 translations) and 2,074 Basque MWEs (along with 3,022 translations) are collected in it, out of which 894 Spanish MWEs and their corresponding translations contain NLP-applicable information.

[C7] Annotation of Spanish and (especially) Basque corpora from a phraseological perspective. The PARSEME multilingual corpus comprises texts of 20 different languages, including Spanish and Basque. Its annotation was carried out in two phases, and we contributed to both of them: in the first edition, as part of the Spanish annotation team; in the second one, by creating the Basque corpus, which consists of 11,158 sentences (157,807 words) and 3,823 MWE annotations. Then, literal occurrences were also studied and annotated on five of the languages in the PARSEME corpus, including Basque. Both the original PARSEME corpus and the one including annotations about literal occurrences are available online[2].

Acknowledgements

The Spanish Ministry of Economy and Competitiveness, who awarded Uxoa Iñurrieta a predoctoral fellowship (BES-2013-066372) to conduct research within the SKATeR project (TIN2012-38584-C06-02).

References

Baldwin, T. and S. N. Kim. 2010. Multiword Expressions. *Handbook of Natural Language Processing*, 2:267–292.

Constant, M., G. Eryiğit, J. Monti, L. Van Der Plas, C. Ramisch, M. Rosner, and A. Todirascu. 2017. Multiword Expression processing: a survey. *Computational Linguistics*, 43(4):837–892.

Inurrieta, U. 2019. *Verb+Noun Multiword Expressions: A linguistic analysis for identification and translation.* Ph.D. thesis, University of the Basque Country (UPV/EHU).

Iñurrieta, U., I. Aduriz, A. Díaz de Ilarraza, G. Labaka, and K. Sarasola. 2017. Rule-based translation of Spanish verb-noun combinations into Basque. In *Proceedings of the 13th Workshop on Multiword Expressions*, pages 149–154. Valencia, Spain.

Iñurrieta, U., I. Aduriz, A. Díaz de Ilarraza, G. Labaka, and K. Sarasola. 2018a. Analysing linguistic information about word combinations for a Spanish-Basque rule-based Machine Translation system. In *Multiword Units in Machine Translation and Translation Technologies*. John Benjamins Publishing C., pages 41–60.

Iñurrieta, U., I. Aduriz, A. Díaz de Ilarraza, G. Labaka, and K. Sarasola. 2018b. Konbitzul: an MWE-specific database for Spanish-Basque. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC2018)*, pages 2500–2504. Miyazaki, Japan.

Iñurrieta, U., I. Aduriz, A. Díaz de Ilarraza, G. Labaka, K. Sarasola, and J. Carroll. 2016. Using linguistic data for English and Spanish verb-noun combination identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING2016): Technical Papers*, pages 857–867. Osaka, Japan.

Iñurrieta, U., I. Aduriz, A. Estarrona, I. Gonzalez-Dios, A. Gurrutxaga, R. Urizar, and I. Alegria. 2018. Verbal Multiword Expressions in Basque corpora. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pages 86–95. Santa Fe, New Mexico, USA.

Ramisch, C., S. Cordeiro, A. Savary, et al. 2018. Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions*, pages 222–240. Santa Fe, New Mexico, USA.

Savary, A., S. R. Cordeiro, T. Lichte, C. Ramisch, U. Iñurrieta, and V. Giouli. 2019. Literal occurrences of multiword expressions: rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.

[1] http://ixa.eus/node/4484

[2] PARSEME multilingual corpus of verbal MWEs: http://hdl.handle.net/11372/LRT-2842. Corpus of literal occurrences: http://hdl.handle.net/11372/LRT-2966.