Using Named Entity Recognition to Identify Personification Constructions in an English <> Spanish Intermodal Corpus of the EP Committee on Petitions

Fernando Sánchez Rodas [1]

[1] *University of Málaga, Campus de Teatinos s/n, 29071-Málaga, Spain*

Abstract

This paper presents a PhD project which tests the effectiveness of NLP-based methods to extract and analyze large amounts of data from translation and interpreting corpora. More specifically, Named Entity Recognition (NER) applications are used in combination with an intermodal corpus of EU texts (that is, a multilingual corpus which contain all possible variations of mediated and non-mediated discourse) in order to identify personification constructions, especially those related to organizations. From the point of view of Construction Grammar (CxG), personifications are argument-structure constructions loaded with relational meaning, which makes them valuable data to feed Machine Translation (MT) and Machine Interpreting (MI) systems or related electronic tools used by translation and interpreting professionals in the briefing preparation phase. In the future, we expect that the compiled intermodal corpus (named PETIMOD) and the NLP techniques can be used to study further types of constructions in institutional discourse, which would be an important contribution not only to corpus-based translation and interpreting studies, but also to CxG itself.

Keywords

Corpus-based translation and interpreting studies, intermodal corpora, NER, Construction Grammar, personification, translation and interpreting technologies

1. Introduction

The present PhD project, funded by the Spanish Ministry of Education and Professional Training (ref. FPU18/05803), focuses on specialised phraseology in mediated and non-mediated discourse in the European Parliament, both from the point of view of the process (shifts) and the product (translationese and interpretese). With the goal of analysing samples by means of state-of-the-art corpora and NLP techniques, an intermodal corpus of EU texts, PETIMOD, was built in the framework of the project. PETIMOD is an English <> Spanish intermodal corpus of translated, non-translated, interpreted, and non-interpreted texts and speeches from the Committee on Petitions of the European Parliament. It is designed to be a flagship for a new multifactorial, interdisciplinary approach in the research agenda for

[1] *Doctoral Symposium on Natural Language Processing from the PLN.net network 2021 (RED2018-102418-T), October 19-20, 2021, Baeza (Jaén), Spain*
EMAIL: fersanchez@uma.es
ORCID: 0000-0003-4244-1835

corpus-based translation studies [1], [2], meant to mark a difference in the upcoming research on translated and interpreted discourse. When a collection of multifactorial linguistic material (different languages, mediation modes, channels, and subtopics) is extracted from a same source, these individual factors can be related to an overarching context, which in turn help researchers determine whether the corpus «coordinates» or metadata are contingent or not in the prevalence of different types of translation and interpreting universals.

Among the different constructions which are prototypical of institutional discourse and empirically observable in the corpus, special attention is devoted to personification metaphors. A basic assumption is that personifications are highly frequent constructions in institutional discourse and common for all the collected texts, no matter their coordinates. Extracting and analysing the key relational meaning posited by personifications through automated and semi-automated NLP techniques such as NER would help knowledge advance in several manners, especially in the development of electronic resources for institutional translation and interpreting. For example, multifactorial translationese and interpretese analysis of these constructions could be applied to train highly specialised machine translation systems on how to behave and relate concepts in specific institutional translation tasks (e.g. when dealing with texts from the Committee on Petitions in contrast to other EP committees). When the translation is performed by a human, NER-based personification extraction can be used to create glossaries in a similar way to already existing tools (e.g. VIP[2] or EU-Bridge[3]), or even to feed multilingual knowledge-based systems which the human translator can interact with in the preparation phase, such as query-answer systems, or Automatic Content Enrichment (ACE). From a more general, interdisciplinary point of view, the collection of intermodal corpora and the application of NER-based techniques could also be used to improve the functioning of the Committee on Petitions itself. Any linguistic study with a statistical basis could shed light on the trends in relation to petitions and assist MEPs in basing their decisions on in-depth and independent expertise, data, and information [3, p. 45]

2. Background

Since its first definition [4], intermodal corpora have attracted growing attention. In recent years, a number of projects in this venue have been developed, mostly centred on institutional translation. The majority have used the European Parliament plenary meetings as source, such as EPIC [5], EPICG [6], and EPTIC [7], while more recent projects [8] have explored the possibilities of the United Nations repositories. Among papers derived from these intermodal resources, one starring topic is the study of formulaicity in non-translated and translated discourse, frequently questioning the validity of the simplification universal in phraseological units e.g. [9]. Our project is at line with the above-cited research in two different manners. First, it continues the tradition of using EP material for the creation of intermodal corpora, creating new synergies by introducing collections from a specific committee rather than from the plenary meetings. Second, it employs such a type of material to explore the contingent features of translationese and interpretese in the light of different textual factors, an exploration which so far has yielded significant results [10], [11].

Despite this continuing line with previous research, one important innovation must be noted. To the best of our knowledge, our project is the first to exploit intermodal corpora from the point of view of Construction Grammar (CxG). Construction Grammar recovers the concept of “construction” first introduced in traditional grammar and reduced to formal terms by Chomskyan grammar. According to CxG, constructions are symbolic units in which form and function are unified, and language is formed

[2] http://www.lexytrad.es/VIP/index_en.php.

[3] https://www.eu-bridge.eu/.

by a constructed continuum [12, pp. 467–468], [13, p. 1]. Among the existing wide range of constructional approaches [14], our project falls under the label of Contrastive Construction Grammar, or CCxG [15]. Contrasting constructions between languages and modes will expectably ease generalisation, an important process in understanding and identifying constructions [16]. This is especially relevant in the case of personification, a linguistic phenomenon which has received little attention cf. [17].

3. Main hypotheses and description of the research

The main hypothesis of the project is that personification metaphors are a common linguistic mechanism in all texts derived from the Committee on Petitions of the European Parliament, regardless of their channel (oral or written), language (English or Spanish), mediation mode (translated/non-translated, interpreted/non-interpreted), or topic. We also hypothesize that the best theoretical approach to study such personification phenomena in EU discourse is Contrastive Construction Grammar, and the best practical method is the exploitation of intermodal corpora through NER-based techniques and quantitative-qualitative analyses.

The research process followed a top-down approach, commencing with a state of the art which took account of the newest updates in translation and interpreting technology, including the pros and cons of intermodal corpora compilation and the most recent advances in corpus-based machine-learning studies of translationese [18]. In a second paper, the first version of the PETIMOD corpus was introduced, and a contrastive English-Spanish/Spanish-English analysis was performed on the basis of the translation and interpreting subcorpora respectively [10]. Although this first analysis was process-oriented (translation and interpreting shifts), it already presented some data which supported the hypothesis of the relatively factor-dependent nature of translation and interpreting universals (ibid.). The validity of these findings was further explored in a third paper [11] which not only presented an enlargement of the intermodal corpus, but also centered on translationese and interpretese and approached the core object of the project (phraseological verbal patterns of entity-as-subject and entity-as-complement type). Findings in this paper revealed the existence of personification metaphors in both translated and interpreted discourse (ibid.). Current research is trying to schematize the construal nature of such personifications, deeply analyzing its metaphorical nature by comparison with real-person entity constructions, and proving possible differences in construction between mediated and non-mediated discourse, among others.

4. Methodology and proposed experiments

As previously suggested in the literature [8], [19], our project heavily relies on the application of NLP techniques as an effective method for the detection, description, and contrast of phraseological units in large translation and interpreting corpora. The specialized nature of personification metaphors in institutional discourse makes it necessary to draw on applications able to work with the syntax-terminology interface, that is, Named Entity Recognition (NER) applications. Named entities, especially those referring to persons, locations and organizations, behave exactly like terms and present similar challenges for their systematic study [20].

The project employs the NER functionality of VIP (Voice-text integrated system for InterPreters). VIP is an online platform developed by the research group Lexytrad that explores the impact and feasibility of using Human Language Technology (HLT) and Natural Language Processing (NLP) for interpreting

training, practice and research [21]. In Corpas Pastor and Sánchez Rodas [10], chunking, detection and extraction of NEs were carried out in two phases. First, entities were retrieved automatically with the VIP NER module and exported to an Excel file. The NER module integrated SpaCy and pre-trained models for English and Spanish. In order to assess the system performance (and, therefore, the accuracy of results), precision and recall were calculated. Low precision/recall results were associated with various issues related to transcription conventions, lack of finer-grained named entity categories or different pre-trained language models. In order to overcome those limitations, transcription conventions were revised and simplified in the next experiment [11]. In addition, a new, finer-grained classification of NEs was adopted by replacing the initial four categories with a longer list of pre-trained categories. In this second study we used DeepPavlov, an open-source framework for deep learning tasks in Python. Precision results with SpaCy improved slightly. By contrast, NER based on DeepPavlov showed slightly lower precision results for English, but it exhibited better performance for Spanish, with a higher mean precision rate for both languages and a smaller difference rate.

A second methodological difference between both studies can be found in the degree of automatization. In the first experimental paper [10], recall calculation made it necessary to add an extra phase of manual NE extraction after the first automatic list obtained with VIP/SpaCy. The combined lists (automatic and manual) obtained from each corpus were then contrasted and used for NE shift identification and extraction, a task which was also manual. The second experiment with VIP, however, employed both the NER module and the Query Corpus (QC) module to study phraseological verbal patterns [11]. The QC Pattern functionality was used to combine part-of-speech tags (PoS) with the "named entity" tag for automatic pattern extraction. Thus, the study included up to three pattern retrieval modes: automatic, semi-automatic, and manual.

The analysis of the extracted units was always qualitative-quantitative. In the first experiment [10], a descriptive transfer operations typology [22] was chosen to categorize the shifts. Instances in each category were then counted, distributed in charts attending to confluent factors (e.g. NE type and language direction), and results compared. The second chapter [11] used the phraseological classification of Biel [23] for qualification. It included more refined statistical tools for quantification, such as absolute and normalized frequencies for NEs. Future publications are expected to follow this fine-grained sort of analysis. They will stick to descriptive methods for constructions [24] and explore the quantification and representation tools which could best account for multi factoriality and the need of establishing generalizations for the personification patterns extracted across corpora. At the end of the process, we would ideally have developed a corpus-based method akin to collostructional analysis [25] for the study of institutional personification metaphors, as it happened in recent papers on a similar matter [26].

5. Discussion

The most relevant matters of discussion so far in both experiments were related to translation and interpreting shifts and/or universals. In Corpas Pastor and Sánchez Rodas [10], the layers of innovation mentioned in previous sections (intermodal corpora, automatic and manual NER) helped add new aspects to the analysis of translation and interpreting shifts (a new category called "normalization" and the possibility of correlating shifts to the semantic labels of the NEs involved). In turn, this originated interesting findings in relation with previous studies (normalization as a language-dependent feature of translation, transformation and simplification as contextual, topic-dependent features of interpreting). In the next paper [11], some data were in line with the previous findings. Simplification appeared to be a contingent feature which depends on the mediation mode and the source languages involved (and also on the topic of the source text). Other relevant findings were that there were clear differences in the nature of shifts between EN-ES translations and ES-EN interpretations of NEs, as well as in the normalization of specialized phraseology. Last but not least, a special case of transference (termed

"transposition") has been found to operate from mediated Spanish to mediated English through original Spanish.

So far, the NER-enhanced analysis of intermodal corpora has proved to be a powerful methodology to discover shifts and translationese traits through phraseological patterns. However, it also comes with a series of limitations and challenges. The suitability of the selected transcription conventions had to be revised in the second experiment. Even though we introduced certain modifications to the system, some of the proposed features seemed more adequate for multimodal corpora and were counterproductive when recognizing NEs. Other suggested improvements include alignment and a tailor-made NE labelling system [10]. Despite the limitations generally faced in the compilation of interpreting and intermodal corpora cf. [18], it must be also noted that they are a conscious, if not necessary, choice for taking new directions and applications of the studies on translation and interpreting norms. From the point of view of Construction Grammar, the NER-based systematic extraction of real instances from a corpus combining all kinds of communication modes (in terms of language, channel, mediation mode, and topic) forms an extremely powerful alliance which is fully in line (and indeed, enhances) the study and validation of CxG principles such as complexity, schematicity, slot identification, network establishment, meaning, compositionality, and generalization. If these data are handled by a human researcher which can apply mathematical and statistical methods accountable for all the textual, cross-convergent factors in the most precise way, the results would be empirical-based models surprisingly close to the reality of translated and interpreted products and the decisions made by the human brain throughout all the translation and interpreting process.

As it can be seen, our project opens new and exciting avenues of research in terms of how linguistic constructions can be analyzed at a high-specific level in order to train brain-like machine translation and interpreting systems (e.g. Neural Machine Translation) designed for specialized institutional settings. The time and resource limitations of our project allow for focusing only on one type of construction, that is, personifications, but these have a good amount of potential in other technology-assisted translation and interpreting tasks as well, such as documentation. The remaining questions are mostly of a methodological nature. It must be clarified which are the best methods for extracting large amounts of data from intermodal corpora in an upmost automated fashion. Different models of quantitative-qualitative analysis must still be tested in order to determine the best to comprise this complex reality and to provide a fruitful communication with IT researchers and/or the machine itself. It is still to be seen whether statistically-refined proposals akin to collostructional analysis answer these needs, or on the contrary, more graphic-oriented models based on metadata and/or network-like representations are more successful in representing the extracted results. Nevertheless, one aspect is clear: from our point of view, the researcher plays the role of a mediator in this task, working from technology to technology, enhancing its work through NLP in order to create more sophisticated systems in the near future.

Acknowledgements

The research reported in this paper has been funded by the Spanish Ministry of Education and Professional Training (ref. FPU18/05803). The paper has also been carried out in the framework of the projects VIPII (PID2020-112818GB-I00), TRIAGE (UMA18-FEDERJA-067) and MI4ALL (CEI-RIS3).

References

[1] G. Corpas Pastor, *Investigar con corpus en traducción: los retos de un nuevo paradigma*.

Frankfurt: Peter Lang, 2008.
[2] G. De Sutter and M.-A. Lefer, “On the need for a new research agenda for corpus-based translation studies: a multi-methodological, multifactorial and interdisciplinary approach,” *Perspectives (Montclair).*, vol. 28, no. 1, pp. 1–23, Jan. 2020, doi: 10.1080/0907676X.2019.1611891.
[3] Policy Department for Citizens’ Rights and Constitutional Affairs, “Achievements Of The Committee On Petitions During The 2014-2019 Parliamentary Term And Challenges For The Future,” 2019.
[4] M. Shlesinger, “Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies,” *Meta*, vol. 43, no. 4, pp. 486–493, 1998, doi: 10.7202/004136ar.
[5] M. Russo, C. Bendazzoli, A. Sandrelli, and N. Spinolo, “The European Parliament Interpreting Corpus (EPIC): Implementation and developments,” in *Breaking Ground in Corpus-based Interpreting Studies*, F. S. Sergio and C. Falbo, Eds. Peter Lang, 2012, pp. 53–90.
[6] Ghent University, “EPICG. The European Parliament interpreting corpus Ghent. A step further,” 2021. [Online]. Available: https://research.flw.ugent.be/en/projects/epicg-european-parliament-interpreting-corpus-ghent-step-further.
[7] S. Bernardini, A. Ferraresi, and M. Miličević, “From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective,” *Target. Int. J. Transl. Stud.*, vol. 28, no. 1, pp. 61–86, 2016, doi: 10.1075/target.28.1.03ber.
[8] M. Tonkopeeva, “Investigating Interpreting and Translation Strategies: A Corpus-Based Approach,” in *TC42*, 2020.
[9] A. Ferraresi, S. Bernardini, M. M. Petrović, and M.-A. Lefer, “Simplified or not Simplified? The Different Guises of Mediated English at the European Parliament,” *Meta*, vol. 63, no. 3, pp. 717–738, 2018, doi: 10.7202/1060170ar.
[10] G. Corpas Pastor and F. Sánchez Rodas, “NLP-enhanced Shift Analysis of Named Entities in an English<>Spanish Intermodal Corpus of European Petitions,” in *Empirical investigations into the forms of mediated discourse at the European Parliament*, M. Kajzer-Wietrzny, S. Bernardini, A. Ferraresi, and I. Ivaska, Eds. Language Science Press, 2021/In press.
[11] G. Corpas Pastor and F. Sánchez Rodas, “EU phraseological verbal patterns in the PETIMOD corpus: a NER-enhanced approach,” in *Handbook of Legal Terminology*, Ł. Biel, Ed. 2021/In press.
[12] G. Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press, 1987.
[13] P. Kay and C. J. Fillmore, “Grammatical constructions and linguistic generalizations: the what’s X doing Y? Construction,” *Language (Baltim).*, vol. 75, pp. 1–33, 1999.
[14] T. Hoffmann and G. Trousdale, Eds., *The Oxford Handbook of Construction Grammar*. New York: Oxford University Press, 2013.
[15] H. C. Boas, Ed., *Contrastive Studies in Construction Grammar*. Amsterdam: John Benjamins, 2010.
[16] P. Wasserscheidt, “Construction Grammar: Basic Principles and Concepts,” *Ukr. Linguist.*, no. 49, pp. 94–116, 2019, doi: 10.17721/um/49(2019).94-116.
[17] A. Dorst, “Personification in discourse: Linguistic forms, conceptual structures and communicative functions,” *Lang. Lit.*, vol. 20, pp. 113–135, 2011.
[18] G. Corpas Pastor and F. Sánchez Rodas, “Now what? A fresh look at language technologies and resources for translators and interpreters,” in *Corpora in translation and Contrastive research in the digital age: recent advances and explorations*, J. Lavid, C. Maíz, and J. R. Zamorano, Eds. John Benjamins, 2020.
[19] G. Corpas Pastor, “Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas,” in *Fraseopragmática* , 2013, pp. 335–373.
[20] G. Jacquet, M. Ehrmann, J. Piskorski, H. Tanev, and R. Steinberger, “Cross-lingual linking of

multi-word entities and language-dependent learning of multi-word entity patterns," in *Representation and parsing of multiword expressions: Current trends*, 2019.

[21] G. Corpas Pastor, "Language Technology for Interpreters: the VIP project," in *Proceedings of 42th Conference Translating and the Computer (TC42)*, 2020.

[22] S. Bernardini, "Intermodal corpora: A novel resource for descriptive and applied translation studies," in *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications*, G. Corpas Pastor and M. Seghiri, Eds. Frankfurt: Peter Lang, 2016, pp. 129–148.

[23] Ł. Biel, "Phraseology in legal translation: A corpus-based analysis of textual mapping in EU law," in *The Ashgate Handbook of Legal Translation*, 2014, pp. 177–192.

[24] T. Hoffmann, "From constructions to construction grammars," in *The Cambridge Handbook of Cognitive Linguistics*, B. Dancygier, Ed. Cambridge: Cambridge University Press, 2017, pp. 284–309.

[25] A. Stefanowitsch and S. T. Gries, "Collostructions: Investigating the interaction of words and constructions," *Int. J. Corpus Linguist.*, vol. 8, no. 2, pp. 209–243, 2003, doi: 10.1075/ijcl.8.2.03ste.

[26] J. Viimaranta and A. Mustajoki, "What Can Science, Religion, Politics, Culture and the Economy Do? A Corpus Study of Metonymical Conceptualization Combined with Personification," *Scando-Slavica*, vol. 66, no. 1, pp. 71–85, Jan. 2020, doi: 10.1080/00806765.2020.1741025.