Chapter 8

NLP-enhanced shift analysis of named entities in an English<>Spanish intermodal corpus of European petitions

Gloria Corpas Pastor[a] & Fernando Sánchez Rodas[a]

[a]University of Malaga

This chapter aims at presenting an NLP-enhanced corpus-based analysis of the translation and interpreting shifts observed in the named entities (NEs) of PETIMOD, an English<>Spanish intermodal corpus of written and oral mediated texts from the Committee on Petitions of the European Parliament. Our main assumption is that shifts in institutional genres mostly occur in the transfer of NEs, and that NLP techniques such as automatic Named Entity Recognition (NER) can be applied to systematically extract and compare examples of these shifts, leading to the (possible) verification of translational and/or interpretational constraints. Results show that traits like normalisation, transformation and simplification depend not only on the language direction or the mediation mode, but also on the semantic category (person, organisation, etc.) of the NE involved. Further studies are needed in order to correlate observed shifts with different NE taxonomies.

1 Introduction

To the present day, a considerable amount of corpus-based research in translation and interpreting has relied on the European Parliament (EP) as a main or only source. Among the European Union (EU) institutions, the Parliament provides an open access repository of both official documents and speeches in a wide range of languages and topics. Before the appearance of intermodal corpora such as EPTIC (Bernardini et al. 2016), the EP had already been used as a source for building translation corpora, e.g., Europarl (Koehn 2005), the European Parliamentary Comparable and Parallel Corpora, or ECPC (Martínez & Serrat 2012),

Gloria Corpas Pastor & Fernando Sánchez Rodas. 2022. NLP-enhanced shift analysis of named entities in an English<>Spanish intermodal corpus of European petitions. In Marta Kajzer-Wietrzny, Adriano Ferraresi, Ilmari Ivaska & Silvia Bernardini (eds.), *Mediated discourse at the European Parliament: Empirical investigations*, 219–251. Berlin: Language Science Press. DOI: 10.5281/zenodo.6977052

and the EU resources at Sketch Engine (Baisa et al. 2016). In the field of corpus-based interpreting studies, it was early pointed out that EP linguistic material could provide researchers with numerous advantages (Bendazzoli 2010). The European Parliament Interpreting Corpus (EPIC) is an example of this (Russo et al. 2012). However, researchers have not yet attended the call. In spite of their unquestionable relevance and high-level complexity, legislative chambers have not received that much attention from linguistics until very recently (Calzada-Pérez 2017).[1] Bibliometric analyses of Europarl (one of the largest multilingual corpora available) show that it has hardly been used in translation studies (Ustaszewski 2019).[2] Reasons for this little academic interest may include corpora distribution in a format that largely disregards the needs of translation research and practice (ibid.) and the need for unexplored, more down-to-earth studies which empirically look at the compared properties of source texts, translations and interpretations and offer a modern, technology-based twist on the methodologies involved.

Against this background, we hypothesise that texts and speeches which originated in the Committee on Petitions of the European Parliament provide an excellent source for the observation of shifts in institutional translation and interpreting, and that shifts in these genres are mostly given in the transfer of Named Entities (NEs). We also assume that recent techniques based on Natural Language Processing (NLP) can be applied to the recognition, extraction and comparison of segments with NEs in two languages and/or modes, as a systematic way of observing shifts between them and proving (or not) the existence of translation and interpreting universals in the analysed texts. To this end, our main research objectives are as follows:

- compile an intermodal, bidirectional corpus (English<>Spanish) of translations and interpretations (plus their different, corresponding source texts) of suitable genres from the EP Committee on Petitions;

- apply NLP-based techniques (Named Entity Recognition) on the said corpus in order to extract relevant units for the study of shifts in both languages and modes;

[1] See Veroz González (2014a,b, 2017) and Prieto Ramos (2019) for examples of corpus-based discursive and/or linguistic analysis in this field.

[2] In order to make the wealth of linguistic data easily and readily available to the translation studies community, a toolkit named EuroparlExtract has been recently developed (Ustaszewski 2019).

- compare qualitatively and quantitatively the observed shifts in the English-Spanish translations and Spanish-English interpretations of the Committee;
- draw conclusions on the relation of three different parameters (language, mode, and semantic category of the NEs) with the presence of translation and interpreting universal features in the analysed documents, especially of simplification traits.

In connection with the objectives above, the chapter presents the following structure. After this introduction (§1), §2 covers basic notions related to communications in the Committee on Petitions. §3 describes the PETIMOD corpus, with a special focus on data collection and design criteria. The NLP-based methodology deployed in this study is spelled out in §4; the main findings are presented in §5 and then discussed in detail (§6). After considering some limitations of our study, §7 offers some concluding remarks on the implications of intermodal corpora for research in translation and interpreting, with special reference to shifts, mediation types and functions, among other relevant issues.

2 A brief overview of EU Petitions

The right to petition is set out in the European legislation. Article 44 of the Charter of Fundamental Rights of the European Union ensures the right to petition to the European Parliament. And Article 227 of the Treaty on the Functioning of the European Union states that "any citizen of the Union, and any natural or legal person residing or having its registered office in a Member State" shall have the right to address a petition to the European Parliament (European Union 2012). A petition may "take the form of a complaint, a request or an observation concerning problems related to the application of EU law or an appeal to the European Parliament to adopt a position on a specific matter" (European Parliament 2020b). After submission, original petitions are registered and given a number. Then, they are summarised (normally in English) and submitted to the members of the Committee on Petitions of the European Parliament for a decision on admissibility and follow-up (ibid.). This committee serves a core function within the governance of the Union, as it acts "as a bridge between Europeans and the EU institutions" (European Parliament 2020a).

As the Committee on Petitions plays an important, mediating role in the context of a multilingual institution and society such as the EU, translation and interpreting are especially relevant in assuring the transparency of its communications. Petition summaries are translated and published in all official EU languages on the Petitions Portal of the European Parliament right after a decision on admissibility has been taken (European Parliament 2020b).[3] The speeches of the committee meetings are also interpreted into each official language and published in the Webstreaming section of the European Parliament Committees website.[4]

As petitions are institutional texts, translators and interpreters have to deal with an important amount of terminology. As Goffin (1994: 637–638) states, the language used in the EU texts, or eurolect, is no different in origin, semantic organization or morpho-syntactic characteristics from any other specialized dialect. Depending on the concept they represent, EU terms are classified as *euronymes,* i.e. terms coined for new institutional realities, or *hétérolexies*, i.e. terms which convey notions and designations rooted in a given official EU language (Goffin 1994: 641).[5]

This classification indicates a prominence highly culture-bound of entities in this knowledge field. Entities are abstractions from external experience which are perceived as self-defined, that is, independent from each other in time and space (e.g. Dolors Montserrat, Bulgaria). Born out of quite specific worldly experiences, some entities pose a real challenge for translators and interpreters (Mayoral 1999). This is especially true for institutional references, like the Spanish *Civil Guard*, which are usually related to the political life of a society (Martin 1997; Ortega 2002). In the Committee on Petitions, where citizens and platforms strive to expose national problems and petitions are chosen by Members of the European Parliament (MEPs) on the basis of their political relevance, it is highly important to give these relevant entities a name (see §4).

3 The PETIMOD Corpus

The purpose of our compilation was to create an intermodal corpus of EU petitions suitable for the study of shifts in translated and interpreted NEs. The size of the corpus was initially limited to one month of institutional activity, and its medium written (see expanded size data in §3.2). The authorship of the documents was exclusively institutional and the topics were mostly agricultural and

[3] In fact, petitions are one of the most frequent briefings for the translation trainees of the EP Schuman Traineeships (https://ep-stages.gestmax.eu/website/homepage).

[4] https://www.europarl.europa.eu/committees/es/peti/meetings/webstreaming.

[5] Examples of the two categories extracted from our named-entity recognition would be "Eurobarometer" (*euronyme)* and "Boletín Oficial" (*hétérolexie*).

environmental, which was not determined by our sampling schema but given by the inherent frequency of the petitions. The publication date was a relevant criterion for the context of this research. As the elaboration of the paper ran parallel to the coronavirus crisis, a cancellation of the Committee activity and/or a change in the content of petitions was predicted. Therefore, the last Committee meeting before the health crisis (19th and 20th February 2020) was chosen as the main source of material. Finally, the languages of the corpus were Spanish and English in their institutional or EU varieties (for a fully-fledged study on eurolects, see Mori 2018).

3.1 Data collection

The retrieval, storing, and conversion of materials started with the oral transcriptions. First, the audiovisual material for the meeting was accessed via the Webstreaming section of the EP Committees site. Three sessions were available for this debate: two on 19 February 2020 (morning[6] and afternoon[7] sessions) and one on 20 February (morning[8] session). We downloaded the complete recordings for both Spanish and English, obtaining six video files in high quality (HQ) .mp4 format.[9] These were moved into a folder structure and coded with the date and time of each session plus the corresponding language abbreviation (e. g. "19feb1000_-EN.mp4"). The duration of each recording is indicated in Table 1.

For cost and ease-of-use reasons, YouTube was the selected application for further ASR (Automatic Speech Recognition) and ATT (Automatic Text Transcription).[10] The upload of the files was performed with a personal account in private visualisation mode to avoid copyright issues. The automatic transcription (without time marking) was generated, then copied and pasted in different TXT

[6] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200219-0900-COMMITTEE-PETI_vd.

[7] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200219-1430-COMMITTEE-PETI_vd.

[8] https://multimedia.europarl.europa.eu/es/peti-committee-meeting_20200220-0930-COMMITTEE-PETI_vd.

[9] Audio tracks are available for the original speeches and the interpretations into any official EU language, although only one version can be downloaded at once. Download is performed through a request system which allows for choosing between the complete session and a selected part, and also between different video qualities. After this, a download link is sent to the desired email account. Downloading high-quality videos was the less time-consuming option in the long term, since low and medium quality videos had to be re-downloaded because of visualization problems. This is a relevant point, as videos are quite helpful for identifying the speakers in each petition.

[10] See Gaber et al.'s (2020) assessment of ASR systems for corpus compilation in interpreting.

Table 1: Properties of the audiovisual files used for automatic transcription.

File(s) name(s)	Length (hour, minutes and seconds)
19feb1000_EN.mp4 19feb1000_ES.mp4	02:10:19
19feb1430_EN.mp4 19feb1430_ES.mp4	03:14:43
20feb900_EN.mp4 20feb900_ES.mp4	02:32:24

files, one for each intervention of the speakers. The naming pattern explained before was used, but three additional references were included for better localisation and connection with the petitions: intervention number, key word/expression related to the topic, and surname of the MEP/speaker (e.g. “19feb1430_17_-ES_oranges_Rego.txt”). In the case of interpretations, the speech’s original language was indicated between brackets with the mark “or-”, as in this example: “19feb1430_78_EN(or-ES)_radioactivewaste_Montserrat.txt”.

Finally, the transcriptions were double-checked manually. In a first round, the EPTIC conventions for transcribing interpretations (Bernardini et al. 2018: 26–27) were applied. In a second revision, the Spanish and English versions of the EU Interinstitutional Style Guide, or ISG (European Union 2021), were used for spelling and capitalisation, together with other resources, such as the English Style Guide from the European Commission’s Directorate-General of Translation[11] and the *Fowlers’ Dictionary of Modern English Usage* (Butterfield 2015). Although the complete six videos in Table 1 were uploaded to YouTube and their transcriptions extracted in different TXT files, the only material revised manually and included in the transcribed component of the corpus was the one from the second session (19th February 2020 14:30–17:30). This was decided because the manual revision of all data was considered too time-consuming for the scope of this chapter. Additional reasons were that it was the longest session, and it contained the largest number of original Spanish speeches, which was in line with our goal of building a bidirectional corpus. As a result of this revision, we obtained 80 transcripts (40 transcriptions of original Spanish interventions and

[11] https://ec.europa.eu/info/sites/info/files/styleguide_english_dgt_en.pdf.

their corresponding 40 interpretations into English, with 18,152 and 10,530 words respectively).

A similar procedure was followed in the case of written documents. The Notices to Members were accessed through the eMeeting portal[12] of the European Parliament. We did not only look for the petitions mentioned in the revised session (19th February 14:30), but for the ones debated in the other two sessions as well, as this was a much quicker way of building our corpus. We browsed and downloaded the petitions in English and Spanish in PDF format. When possible, we included all the other accessible PDF documents which were not petitions but were also handled in the debates, such as reports and opinions. This was done for the sake of coherence and terminological relevance. Similarly to the transcriptions, these files were organised in a folder structure and renamed using a coding system with date and time of the meeting, language abbreviation and key word/expression related to the topic (e.g. "19feb1430_EN_oranges.pdf", "20feb900_EN_insects.pdf"). In the case of translations, the document's original language was indicated between brackets with the mark "or-", as in this example: "19feb1000_ES(or-EN)_amendment.pdf". Finally, the documents were saved as plain text (TXT) files with UTF-8 encoding for correct character recognition by any corpus software.

3.2 Design criteria

PETIMOD is a parallel intermodal corpus which contains citizens' petitions and other documents related to the Committee on Petitions of the European Parliament, as well as transcribed speeches related to these documents. It comprises two subcorpora, allowing for various types of comparison to be carried out: PETIMOD_ORIG (original texts and speeches in English and Spanish) and PETIMOD_MEDIATED (their corresponding translations and interpretations from English into Spanish, and vice versa). At the same time, PETIMOD is a *bidirectional* corpus (Olohan 2004) because the mediating activity is not only represented in B-A direction (Spanish speeches interpreted into English), but also A-B (English documents translated into Spanish). Finally, it is important to recall that, in contrast to other intermodal corpora in the field (cf. the works on EPTIC), PETIMOD comprises translations and interpretations (texts and speeches) that belong to different genres, the first being mostly Notices to Members and the second being interventions of said MEPs and speakers invited to the Committee on Petitions' sessions held in Brussels monthly.

[12] https://emeeting.europarl.europa.eu/emeeting/committee/agenda/202002/PETI?meeting=PETI-2020-0219_1P&session=02-19-10-00.

Specifically, the corpus consists of all the petitions discussed during the three sessions of February 2020, whereas the original Spanish speeches and their English interpretations were extracted from a single session (19th February 2020 14:30–17:30), as explained in §3.1. In order to diversify our corpus and investigate further correspondences, some non-petitional public documents discussed in the sessions, such as reports or opinions, were also included.

According to classical typological parameters (Corpas Pastor 2001; Olohan 2004; Shlesinger 2008), the PETIMOD corpus can be classified as follows:

- it is *parallel,* as it is composed of original texts (and speeches) plus their translations (and interpretations).
- It is *intermodal,* as it encompasses original, translated, and interpreted components which can be compared to each other in a three-way fashion.
- It is *written*, as it contains official documents (PDF and TXT) as well as transcriptions of parliamentary speeches (TXT).
- It is *bidirectional*, as it comprises English documents translated into Spanish (A-B), and also of Spanish speeches interpreted into English (B-A).

The size of the PETIMOD corpus is provided in Table 2 and Table 3 (in total, per component and per language). The total number of documents, running words (tokens) and word types (types) were calculated using ReCor.[13]

Table 2: PETIMOD size per component.

Counts	**Petimod_orig**	**Petimod_mediated**	**Total**
Tokens	59,270	65,038	124,308
Types	6,523	6,622	13,145
Documents	59	59	118

Figure 1 provides a visual representation of the composition of our intermodal corpus, in which the double arrows represent the (ordered) envisaged comparisons for analysis (A). In this study, the selected comparisons are A_5 and A_6. As can be seen, cross-comparison of A_5 and A_6 presents differences not only in directions (EN<>ES), but also different language families in terms of origins (Anglosaxon and Romance), different modes (written and oral) and different types

[13] http://www.lexytrad.es/en/resources/recor-3/.

Table 3: PETIMOD size per language and component.

Counts	Petimod _orig_en	Petimod _orig_es	Petimod _ mediated_es	Petimod _ mediated_en
Tokens	46,625	12,645	54,295	10,743
Types	4,072	2,451	5,012	1,610
Documents	19	40	19	40

of linguistic mediation (translation and interpreting). This is a conscious choice, which aims at raising awareness of the multifactorial nature of translation and interpreting phenomena (cf. De Sutter & Lefer 2020), but also at trying to establish generalisations between the two communicative situations by looking at a possible core set of shared factors given by the function of the institution for which they are produced, that is, the Committee on Petitions.[14]

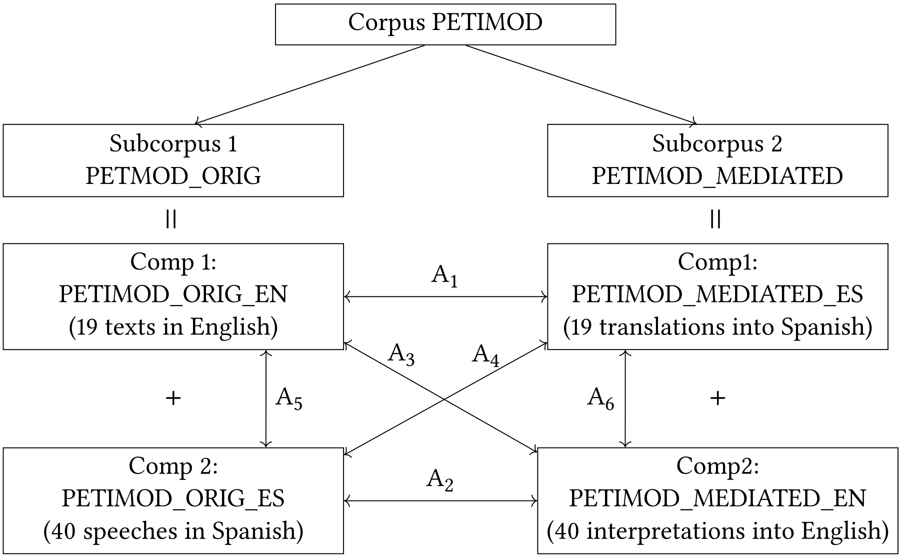


Figure 1: PETIMOD subcorpora and envisaged comparisons.

[14] Cf. Saldanha 2009 for discussion on the bridging role of "function" and "context" in linguistic approaches to translation and interpreting.

4 Methodology

In order to study shifts in translated speeches and interpretations, we have focussed on NEs and extraction techniques. Named entity recognition (NER) is the task of identifying and categorising key information or real-world objects (entities) in text. In NLP, a NE is a real-world "object" that is assigned a name (e.g., *Donald Trump*, *United States*, *The Foreign Office*, *World Health Organisation*, etc.).

For this study both automatic and manual extraction of NEs were performed. Both precision and recall were calculated in order to assess the system's performance. Then, a corpus-based study of NEs in the translated and interpreted components was carried out.

4.1 Automatic named entity recognition

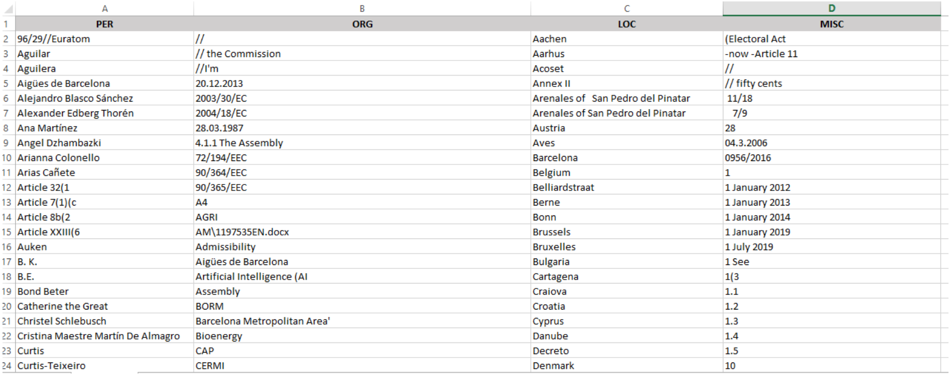
Similarly to other models trained on a Wikipedia-based corpus (Nothman et al. 2013), for this paper we have used the VIP[15] NER annotation scheme, that distinguishes four entity types: PER (named person or family), LOC (name of politically or geographically defined locations, e.g., cities, countries, regions, rivers, lakes, seas, mountains), ORG (named corporate, governmental or other organisational entities) and MISC (miscellaneous entities, e.g., laws, events, languages, products, work of art, etc.). In order to extract and identify NEs automatically, a script[16] has been programmed based on the VIP module for NE chunking, extraction, and identification. See Figure 2 for a screenshot of the Excel file generated by the script.

VIP integrates spaCy[17] (a free open-source library in Python). VIP provides a user-friendly interface and allows importing NEs into an Excel file. Pre-trained spaCy models rather than custom-made NER models were used. The two pre-trained spaCy models used – es_core_news_lg (Spanish) and en_core_web_lg (English) – differ in the degree of granularity of the NER annotation scheme. The Spanish model recognises four categories (PER, LOC, ORG and MISC), whereas the English model recognises twelve additional types of entities: ORDINAL (e.g., st, *second*), DATE (*13 October*, *2019*), GPE (countries, cities and states, e.g., *Madrid*), CARDINAL (*102*, *67.5*), NORP (nationalities, religious or political groups, e.g. *Democrats*),

[15] VIP (Voice-text integrated system for InterPreters) is a hub of online resources and computer-assisted tools for interpreters created by the research group Lexytrad of the University of Malaga. VIP includes a suit of interpreting-related tools with a NER module and its own annotation scheme. The platform can be accessed here: http://www.lexytrad.es/VIP/index_en.php.

[16] Authors would like to express their gratitude to Mr Francisco Javier Lima for writing the script used in this paper, which has been integrated in the VIP NER functionality.

[17] https://spacy.io/.



	A	B	C	D
1	PER	ORG	LOC	MISC
2	96/29//Euratom	//	Aachen	(Electoral Act
3	Aguilar	// the Commission	Aarhus	-now -Article 11
4	Aguilera	//I'm	Acoset	//
5	Aigües de Barcelona	20.12.2013	Annex II	// fifty cents
6	Alejandro Blasco Sánchez	2003/30/EC	Arenales of San Pedro del Pinatar	11/18
7	Alexander Edberg Thorén	2004/18/EC	Arenales of San Pedro del Pinatar	7/9
8	Ana Martínez	28.03.1987	Austria	28
9	Angel Dzhambazki	4.1.1 The Assembly	Aves	04.3.2006
10	Arianna Colonello	72/194/EEC	Barcelona	0956/2016
11	Arias Cañete	90/364/EEC	Belgium	1
12	Article 32(1	90/365/EEC	Belliardstraat	1 January 2012
13	Article 7(1)(c	A4	Berne	1 January 2013
14	Article 8b(2	AGRI	Bonn	1 January 2014
15	Article XXIII(6	AM\1197535EN.docx	Brussels	1 January 2019
16	Auken	Admissibility	Bruxelles	1 July 2019
17	B. K.	Aigües de Barcelona	Bulgaria	1 See
18	B.E.	Artificial Intelligence (AI	Cartagena	1(3
19	Bond Beter	Assembly	Craiova	1.1
20	Catherine the Great	BORM	Croatia	1.2
21	Christel Schlebusch	Barcelona Metropolitan Area'	Cyprus	1.3
22	Cristina Maestre Martín De Almagro	Bioenergy	Danube	1.4
23	Curtis	CAP	Decreto	1.5
24	Curtis-Teixeiro	CERMI	Denmark	10

Figure 2: English NEs file automatically retrieved by the VIP script.

FAC (buildings, airports, highways, bridges, etc., e.g. *Golden Gate*), PERCENT (percentage, including %), PRODUCT (objects, vehicle, foods, etc., e.g. *Toyota*), LAW (laws, directives, regulations, etc.), QUANTITY (measurements of weight, distance, etc., e.g., *hectare*), MONEY (e.g., *cents*, *dollars*), TIME (times smaller than a day), and LANGUAGE (e.g., *Spanish*). For this reason, English categories have been simplified. Thus, akin to the Spanish model, FAC and GPE have been subsumed under the category LOC and the rest have been grouped under MISC.

Precision has been calculated to measure how well our NER system performs. Precision is defined as the fraction of relevant instances among all retrieved instances, i.e. the total number of relevant NEs retrieved divided by the number of all NEs retrieved (correctly and incorrectly identified by the model).

$$Relevant\,NEs = Total\,number\,of\,correctly\,retrieved\,NEs - Errors$$

$$Precision = \frac{Relevant\,NEs}{Total\,number\,of\,extracted\,NEs}$$

For calculating the above formula, it was necessary to manually assign the retrieved NEs to three categories: (a) segments which were correctly identified as NEs (“Correct ID”), (b) segments wrongly identified as NEs (“Wrong ID”), (c) and segments correctly identified as NEs but wrongly labelled (“Wrong Class”).

NER performance has been calculated in terms of precision for both languages. Two levels of analysis have been established. The first level takes all NEs correctly identified as relevant, irrespective of their classification. For instance, the non-entity sequence [*Articles 20(2)(b*], retrieved as NE by the system, would be classified as an error, whereas the retrieved sequence [*2004/18/EC*] would be considered as relevant (correctly identified) whether it has been tagged correctly (MISC) or not (ORG). The mathematical formula for Level 1 is as follows:

$$Precision = \frac{(Correct\,ID + Wrong\,Class)}{(Correct\,ID + Wrong\,ID + Wrong\,Class)}$$

Table 4 presents results for this wider category of relevant NEs.

Table 4: NER performance in terms of precision (Correct ID + Wrong Class).

	English	**Spanish**
NEs retrieved	1,726	1,183
Correct identification	1042	456
Wrong identification	522	576
Wrong class	162	151
Errors	684	727
Relevant NEs retrieved	1204	607
Precision	0.697	0.513

A further level of analysis is achieved by discriminating between NEs correctly identified and correctly tagged (for instance, [*2004/18/EC*] correctly identified as NE and classified as MISC) and NEs correctly identified but wrongly tagged (for instance NE [*2004/18/EC*] classified as ORG). The formula below allows refining results by considering wrong-labelled NEs as errors (see Table 5).

$$Precision = \frac{(Correct\,ID)}{(Correct\,ID + Wrong\,ID + Wrong\,Class)}$$

4.2 Manual named entity extraction

In order to assess the performance of the system in terms of recall, it was necessary to identify and extract NEs manually for both languages. Recall is the fraction of retrieved instances among all relevant instances, i.e. it refers to the total number of relevant NEs retrieved versus the total number of relevant NEs found manually in our corpora. The idea was to delve into word lists generated by a corpus management tool, so we could identify NEs in the documents that had not been automatically recognised by our system. The sum of both types

Table 5: NER performance in terms of precision (Correct ID).

	English	**Spanish**
NEs retrieved	1,726	1,183
Correct identification	1042	456
Wrong identification	522	576
Wrong class	162	151
Errors	522	576
Relevant NEs retrieved	1,042	456
Precision	0.603	0.385

of NEs (automatically recognised and manually extracted) would bring the total number of relevant NEs in the corpus. The formula used to calculate recall is presented below:

$$Recall = \frac{Relevant\, NEs\, extracted}{Total\, number\, of\, relevant\, NEs\, in\, the\, corpus}$$

The selected corpus management platform was Sketch Engine, the same tool used for corpus statistics in §3.1. Sketch Engine was chosen for two reasons: it features European Parliament corpora (Ustaszewski 2019) and its interface allows for swift change when working with several subcorpora simultaneously. We uploaded the plain-text files for each of the four components of our intermodal corpus as four different monolingual comparable corpora, using the "New Corpus" functionality in the menu "Select Corpus → My Corpora".

Then, a starting point for manual NER was the wordlist generator of Sketch Engine, which was used in each component. We chose to compose a list of nouns filtered by two stopword lists (one for each language).[18] In Sketch Engine, this can be done in the "Advanced" tab of the wordlist menu, under the heading "Exclude these words"; the list has to be pasted manually, with one word per line. The PETIMOD_MEDIATED_EN subcorpus, for example, yielded a list of 643 nouns (e.g., *Commission*, *situation*, *petitioner*, *problem*, etc.).

Once the wordlist was generated (Wordlist 1), we had a basic frequency list which contained some nouns that could be used to refine the automatic NER, such as committee (22 occurrences), directive (16), agreement (13), group (11), plan

[18] Stopword lists were directly copied and pasted from http://members.unine.ch/jacques.savoy/clef/index.html. The interjection "ehm", used in the transcription conventions for representing hesitation in speech, was also added to the stopword list.

(10), fund (9), etc. Then, a second wordlist (Wordlist 2) was created by sorting the nouns alphabetically and filtering out those which were neither semantically nor frequency-wise relevant (e.g. *angle*, 1) or which had been correctly recognised by the automatic NER (e.g. *Aguilar*, 1). Although Sketch Engine did not allow for alphabetical sorting of the wordlist, nor for complete visualisation of the results in one column (the maximum is 500), it was possible to download the data in a CSV file and order the words by using the corresponding Excel function.

The next step was to search for the nouns in the wordlist manually. To this end, we opened a new window of concordances in Sketch Engine to directly search for the occurrences of each noun in the corpus. At this point, some basic functions of concordance search, such as alphabetical sort by context (left and right), file view, and wildcard search, were also used for easier and faster identification of new entities. Wildcard search proved especially useful in combination with the wordlist, as in some cases looking for lexical roots made it possible to inspect several instances of the list at once. For instance, a search for [**omission**] retrieved up to three instances of the wordlist simultaneously (*Commission*, *commission*, and *commissioner*).

Apart from wordlist frequency, institutionalisation was the second criterion for identifying relevant NEs. In this case, coverage in Eur-lex,[19] IATE[20] and/or TermCoord's Glossary Links[21] was taken as a reference (see Figure 3).

Following these criteria, new NEs were extracted from the concordances in each component and saved in an Excel file. Some examples of further relevant NEs manually extracted were *Directorate-General for the Mar Menor* (ORG, PETIMOD_MEDIATED_EN), *Acuerdo de Asociación Económica* (MISC, PETIMOD_MEDIATED_ES), *municipality of Real* (LOC, PETIMOD_ORIG_EN) and *Directiva de inundaciones* (MISC, PETIMOD_ORIG_ES), among others.

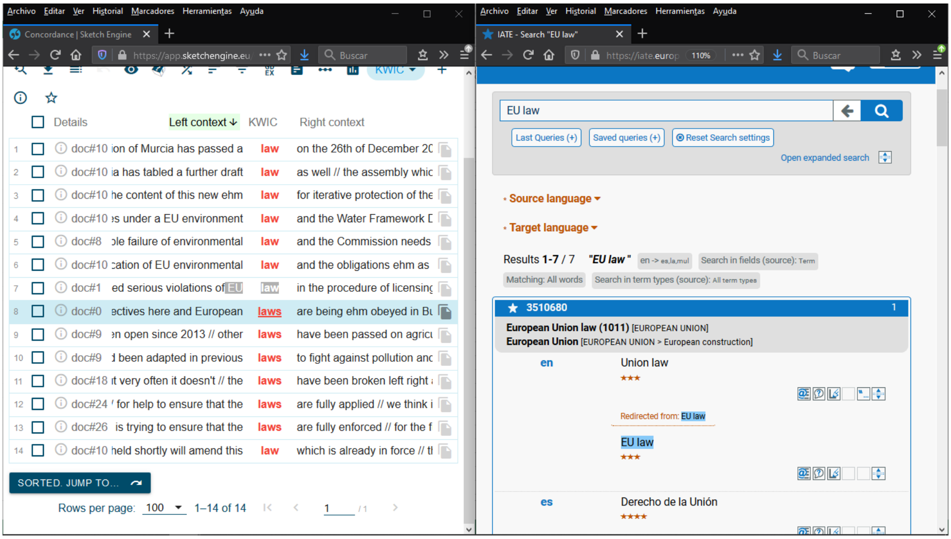
Finally, NER performance has been calculated in terms of recall for both languages. As in the case of precision, two granular levels of analysis have been used. The first level takes all NEs correctly identified by the script as relevant, irrespective of their classification (see §4.1). For these calculations, it was necessary to sum the manually retrieved NEs for each component, combining and sorting them by language.

A further level of recall analysis is achieved by discriminating between NEs correctly identified and correctly tagged by the automatic script (relevant) and

[19] https://eur-lex.europa.eu/.

[20] https://iate.europa.eu/home.

[21] A database of more than 8,000 glossaries managed by the Terminology Coordination Unit of the EP Directorate-General for Translation (https://termcoord.eu/glossarylinks/).



	Details	Left context ↓	KWIC	Right context
1	doc#10	ion of Murcia has passed a	law	on the 26th of December 20
2	doc#10	ia has tabled a further draft	law	as well // the assembly whic
3	doc#10	he content of this new ehm	law	for iterative protection of the
4	doc#10	es under a EU environment	law	and the Water Framework D
5	doc#8	ble failure of environmental	law	and the Commission needs
6	doc#10	cation of EU environmental	law	and the obligations ehm as
7	doc#1	ed serious violations of EU	law	in the procedure of licensing
8	doc#0	ectives here and European	laws	are being ehm obeyed in Bu
9	doc#9	en open since 2013 // other	laws	have been passed on agricu
10	doc#9	d been adapted in previous	laws	to fight against pollution and
11	doc#18	it very often it doesn't // the	laws	have been broken left right
12	doc#24	for help to ensure that the	laws	are fully applied // we think i
13	doc#26	is trying to ensure that the	laws	are fully enforced // for the f
14	doc#10	held shortly will amend this	law	which is already in force // th

EU law

Results 1-7 / 7 "EU law" en -> es,la,mul Search in fields (source): Term Matching: All words Search in term types (source): All term types

★ 3510680 1

European Union law (1011) [EUROPEAN UNION]
European Union [EUROPEAN UNION > European construction]

en Union law ★★★

Redirected from: EU law

EU law ★★★

es Derecho de la Unión ★★★★

Figure 3: Example of manual NER using institutional criteria. The consulted NE ("EU law") had not been automatically recognised.

Table 6: NER performance in terms of recall (Correct ID + Wrong Class).

	English	Spanish
Total no. of relevant NEs	1,557	896
Relevant NEs retrieved automatically	1,204	607
Relevant NEs retrieved manually	353	289
Recall	0.773281	0.677455

NEs correctly identified but wrongly tagged (not relevant). This allows refining recall results by excluding wrong-labelled NEs from calculation.

4.3 Corpus-based analysis

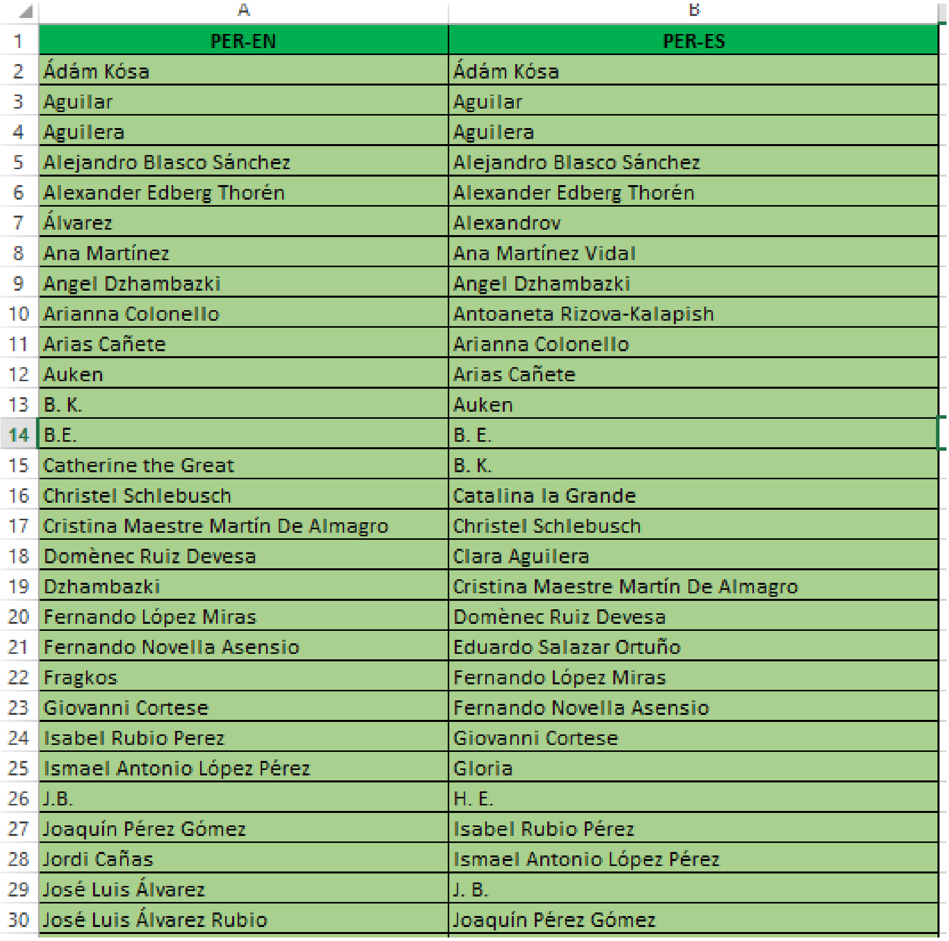
For the corpus-based analysis described in this section, all relevant NEs in the Excel files (correctly identified, correctly identified but mislabelled, and manually extracted) were prepared by listing them together in a new file, manually sorting them by category and language. Figure 4 below shows the two columns for the PER category (English and Spanish), the first one attending to the VIP annotation scheme order described above.

Table 7: NER performance in terms of recall (Correct ID).

	English	Spanish
Total no. of relevant NEs	1,075	745
Relevant NEs retrieved automatically	1,042	456
Relevant NEs retrieved manually	353	289
Recall	0.969302	0.612080

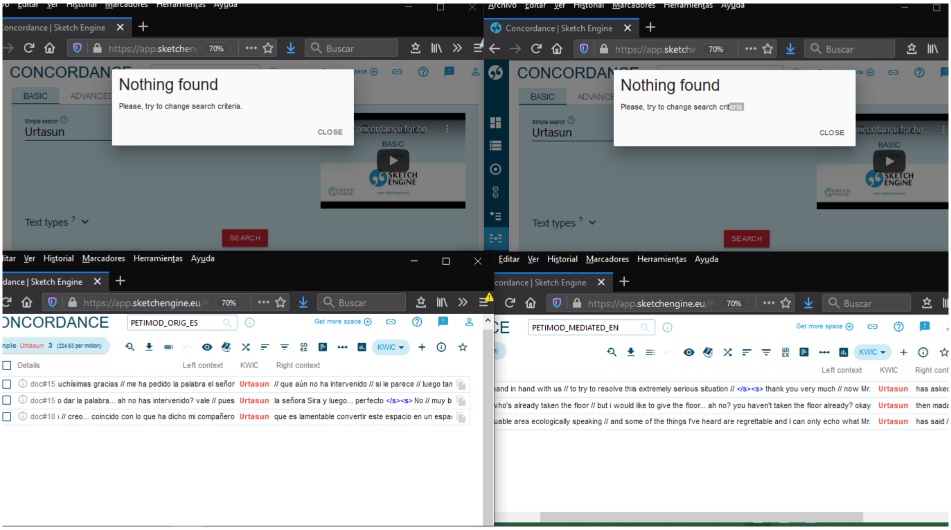
Once all NEs were prepared, the next step was analysing the observable shifts in their translation and/or interpretation. We decided to perform the shift analysis both in the EN>ES translation (components PETIMOD_ORIG_EN vs. PETIMOD_MEDIATED_ES, or direction A1 in Figure 1) and in the ES>EN interpretation (components PETIMOD_ORIG_ES vs. PETIMOD_MEDIATED_EN, or direction A2 in Figure 1). The reasons for this decision were two: it comprised all the different components in our corpora, and the cross-comparison of translation and interpreting analysis was expected to show interesting findings.

Provided that the raw material for analysis (i.e., the NEs) was already extracted, labelled, and sorted by language, the next three steps to be taken were: (1) contrasting them across languages to observe (possible) changes; (2) searching for them in the corpora in order to extract contextual exemplification of the shifts and identify their direction; (3) categorising the shifts. Step 1 could be done directly in the Excel file, underlining those units already analysed and/or not shifted. For Step 2, we prepared a mosaic-style panel of four windows, one for each uploaded component in Sketch Engine, in order to identify the directions and the exact alignment of the document in which the shift occurred (see Figure 5). A total of 142 shifts (69 for EN>ES translation, 73 for ES>EN interpretation) were identified and analysed. Regarding Step 3, the bottom-up transfer operations typology of Bernardini (2016) was chosen to categorise the shifts (see §5 for further description). The category was annotated next to the extracted concordances, in a table-like fashion. The Excel file for shift analysis included retrieved NEs and categorised shifts sorted by direction. As it can be inferred, the previous work with the entities in the automatic and manual NER phases was very helpful for this analysis, and allowed for quick identification and remembrance of the nature and direction of several shifts. Again, the institutional resources cited in §4.2 (Eur-lex, IATE, Glossary Links) were occasionally used in combination with generic searches in Google and/or Wikipedia in order to gain insights into the possible motivations behind some of the shifts encountered.



	A	B
1	PER-EN	PER-ES
2	Ádám Kósa	Ádám Kósa
3	Aguilar	Aguilar
4	Aguilera	Aguilera
5	Alejandro Blasco Sánchez	Alejandro Blasco Sánchez
6	Alexander Edberg Thorén	Alexander Edberg Thorén
7	Álvarez	Alexandrov
8	Ana Martínez	Ana Martínez Vidal
9	Angel Dzhambazki	Angel Dzhambazki
10	Arianna Colonello	Antoaneta Rizova-Kalapish
11	Arias Cañete	Arianna Colonello
12	Auken	Arias Cañete
13	B. K.	Auken
14	B.E.	B. E.
15	Catherine the Great	B. K.
16	Christel Schlebusch	Catalina la Grande
17	Cristina Maestre Martín De Almagro	Christel Schlebusch
18	Domènec Ruiz Devesa	Clara Aguilera
19	Dzhambazki	Cristina Maestre Martín De Almagro
20	Fernando López Miras	Domènec Ruiz Devesa
21	Fernando Novella Asensio	Eduardo Salazar Ortuño
22	Fragkos	Fernando López Miras
23	Giovanni Cortese	Fernando Novella Asensio
24	Isabel Rubio Perez	Giovanni Cortese
25	Ismael Antonio López Pérez	Gloria
26	J.B.	H. E.
27	Joaquín Pérez Gómez	Isabel Rubio Pérez
28	Jordi Cañas	Ismael Antonio López Pérez
29	José Luis Álvarez	J. B.
30	José Luis Álvarez Rubio	Joaquín Pérez Gómez

Figure 4: Screenshot of the Excel file with extracted PER NEs (English/Spanish).



Nothing found

Please, try to change search criteria.

PETIMOD_ORIG_ES

Simple Urtasun 3 (224.63 per million)

Details	Left context	KWIC	Right context
doc#15	uchísimas gracias // me ha pedido la palabra el señor	Urtasun	// que aún no ha intervenido // si le parece // luego tan
doc#15	o dar la palabra... ah no has intervenido? vale // pues	Urtasun	la señora Sira y luego... perfecto </s><s> No // muy b
doc#18	// creo... coincido con lo que ha dicho mi compañero	Urtasun	que es lamentable convertir este espacio en un espa

PETIMOD_MEDIATED_EN

Left context	KWIC	Right cont
hand in hand with us // to try to resolve this extremely serious situation // </s><s> thank you very much // now Mr.	Urtasun	has asked
who's already taken the floor // but i would like to give the floor... ah no? you haven't taken the floor already? okay	Urtasun	then mada
uable area ecologically speaking // and some of the things I've heard are regrettable and I can only echo what Mr.	Urtasun	has said /

Figure 5: Four-window panel in Sketch Engine to track shifts in the corpus.

5 Shift analysis and results

The term "shifts" commonly refers to "changes which occur or may occur in the process of translating [and interpreting]" (Bakker et al. 2009: 269). Shifts of translation (and interpreting) can be distinguished from the systemic differences which exist between source and target languages and cultures. Systemic differences, which pertain to the level of competence, are part of the opening conditions for translation (and interpreting). Shifts, on the other hand, result from attempts to deal with systemic differences (ibid). In this study, only NEs that experienced shifts during translation/interpreting were analysed, whereas translations/interpretations where no shifts in NEs occurred were ignored. As stated in the previous section, the bottom-up transfer operations typology from Bernardini (2016: 140), used to categorise shifts in the intermodal corpus EPTIC, was chosen for this analysis. It includes register shifts (either upwards or downwards), quantitative meaning shifts (contraction, expansion, clarification, broadening), and transformational meaning shifts (partial and total), as well as cases akin to normalisation.[22] In the next paragraphs, each of these categories will be

[22] For the sake of clarity, the original name of this category was rephrased for this chapter (from "more collocational" to "normalisation").

described and illustrated with examples from our corpus.[23] However, as Bernardini (ibid.) puts it:

> As often happens with language in use, some instances were impossible to assign indisputably to one category only. In these cases a decision was made based on a close reading of the co-text and, inevitably, intuition as to the main reason for making a certain choice. (Bernardini 2016: 140)

The first type of shifts, categorised under "register" (up and down), could indeed be sometimes confused with contraction and expansion changes. Illustrating them with the use of acronyms helps establish a clear-cut separation between register shifts (formal) and meaning shifts. In example (1), the acronym is avoided in the EN-ES direction, which increases the level of formality. It is shifted to the modifier *de la Unión*, which in the Spanish eurolect can be considered even more formal than the alternative *de la Unión Europea* because of its specificity. Exactly the same change can be further found in the same sentence (from *EU Member States* to *Estados Miembros de la Unión*).

(1) Register up shift (EN-ES translation)
  a. The EU Delegation in Japan and the authorities of EU Member States [PETIMOD_ORIG_EN]
  b. La <Delegación de la Unión> en Japón y las autoridades de los Estados miembros de la Unión [PETIMOD_MEDIATED_ES]

On the contrary, in the ES-EN interpretation in example (2), the acronym *ENVI* is preferred instead of the denomination *Comisión ENVI* (already shortened in the original). As the Spanish ISG always recommend the use of the word *Comisión* when referring to these bodies (cf. European Union 2021: 172), this can be considered a shift which downgrades register. In fact, some shifts of the same nature can be observed in the surrounding verbs (*pedimos* ⟶ *pass*, *realice* ⟶ *carry out*).

(2) Register down shift (ES-EN interpretation)
  a. le pedimos a la Comisión ENVI que realice una visita [PETIMOD_ORIG_ES]
  b. we should pass it on to <ENVI> and ask them to carry out a study… visit [PETIMOD_MEDIATED_EN]

[23] We followed the same conventions of Bernardini (2016: 140). The underlined NE in the source roughly corresponds to the NE or segment in the target (in angle brackets).

Moving to quantitative meaning shifts, contraction implies changing from an informative detailed NE or NE sequence to a shorter and more under-defined equivalent (Bernardini 2016: 141). Although the author does not put it explicitly, it can be deduced from the given examples that contraction and expansion are related, as the reduction (or addition) of meaning also conveys a reduction or addition in the number of words (ibid.). In example (3), the English word referring to the region (*Galicia*) is omitted in the Spanish translation, as it is (supposedly) not necessary for a standard Spanish reader.

(3) Contraction shift (EN-ES translation)

a. in an existing business park, on a green field plot, in Curtis-Teixeiro, La Coruña , Galicia, Spain. [PETIMOD_ORIG_EN]

b. en un parque de actividades económicas ubicado en un terreno no urbanizado de Curtis-Teixeiro <(La Coruña, España)> [PETIMOD_MEDIATED_ES]

A similar example, but this time of expansion, could be extracted from the ES-EN direction. Here we also have a LOC NE referred to a quite specific Spanish area (*Campo de Cartagena*), but the interpreter's decision is the opposite one: adding the modifier *region* to specify the nature of the named entity, thus increasing the number of words.

(4) Expansion shift (ES-EN interpretation)

a. él estaba contentísimo con el modelo agrícola del Campo de Cartagena [PETIMOD_ORIG_ES]

b. they were very happy with the agricultural model in the <Campo de Cartagena region> [PETIMOD_MEDIATED_EN]

Like expansion shifts, clarifications are instances of addition, in which meanings that are implicit in the sources are made explicit in the targets. As a rule of thumb, Bernardini (2016: 140) states that "in the case of clarification words used are more explicit, whereas in the case of expansion there is also an increase in the number of words (though admittedly the difference is not always clear-cut)." For improved distinction, it could be added that clarification seemingly implies adding *less* words that any expansion would. Again, the LOC label provides a suitable example in the EN-ES translation. In example (5) the unit *municipality of Real*, which initially refers to a geopolitical entity and could imply demanding information from any office contained in these borders, is shifted to a more explicit reference (*Ayuntamiento de Real* or town hall). Interestingly, by performing this operation, the nature of the NE is also shifted (from LOC to ORG).

(5) Clarification shift (EN-ES translation)

a. the Environmental Inspection Service requested the municipality of Real to inform [PETIMOD_ORIG_EN]

b. En 2012, el Servicio de Inspección Medioambiental pidió al <Ayuntamiento de Real> información [PETIMOD_MEDIATED_ES]

The third possible case of quantitative meaning shift is broadening, or generalisation through vaguer or emptier terms. In example (6), two PER NEs are generalised through the common, more neutral noun *petitioners*. This is a quite prototypical example, as additionally the first PER (*Eduardo Salazar Ortuño*) is not one of the petitioners, but a lawyer who is present on behalf of them (this is contextual information which can be found in the corpus some interventions before). Other aspects worth mentioning are the double nature of the shift and the extended broadening phenomena in the two MISC NEs *dos minutos*, which are suppressed in favour of the more general idea conveyed by *conclude*.

(6) Broadening shift (ES-EN interpretation)

a. para concluir esta petición le daríamos la palabra por dos minutos al señor Eduardo Salazar Ortuño // y luego le daríamos dos minutos más al señor José Luis Álvarez-Castellanos Rubio [PETIMOD_ORIG_ES]

b. let's close the debate on that and we will conclude this point by giving the floor back to <our two petitioners> [PETIMOD_MEDIATED_EN]

Transformational shifts include two different grades (partial and total). Partial transformation involves a reformulation with approximately the same co-textual meaning, but using an unrelated expression with a different out-of-context meaning (Bernardini 2016: 142). Again, the ES-EN interpretation provides a prototypical example of partial transformation. The collocation *flourishing ecosystem* in example (7) does not convey the same specialised meaning as *Zona de Especial Conservación*, but serves as equivalent in the context of the inversion operated in the target sentence. As already observed in example (6), the shift affects more than one particular NE and can be analysed even at the sentence level.

(7) Partial transformation shift (ES-EN interpretation)

a. en la cuenca del Mar Menor la Red Natura 2000 es una etiqueta formal que no responde a una gestión eficiente de lo que sería una Zona de Especial Conservación [PETIMOD_ORIG_ES]

b. Natura 2000 is an official label that should lead to efficient management of what should be a <flourishing ecosystem> [PETIMOD_MEDIATED_EN]

Total transformation, on the other hand, may sometimes override the limits of equivalence and fall closer to the notion of translation error (see for example Hurtado 2017). In example (8), the translator seems to have looked for the real (and very different) equivalent of the generic NE underlined in (8), but has made a mistake in the process (*Consejería de Turismo y Cultura* instead of *Consejería de Turismo, Cultura y Medio Ambiente*).[24] This is a very similar case to the one illustrated by Bernardini (2016: 142), in which an error is produced in the search of a salient collocation (here, a NE) in the target language.

(8) Total transformation shift (EN-ES translation)

a. the creation of a specific Directorate-General for the Mar Menor, within the regional Department for the Environment [PETIMOD_ORIG_EN]

b. la creación de una Dirección General del Mar Menor, dentro de la <Consejería de Turismo y Cultura> [PETIMOD_MEDIATED_ES]

The last shift category presented in this typology is normalisation. In the words of Bernardini (Bernardini 2016: 142), here "the difference from source to target seems to be one of *collocationality*: i.e., the inherent motivation for using a certain turn of phrase seems to be its salience as a phrase, or status as a collocation, in the target language." In this study, however, the analysed normalisation shifts are not performed on collocations, but on multi-word terms or NEs, such as the ones in example (9). In this translation, a subtle shift in a preposition (*National Assembly in France* → *Asamblea Nacional de Francia*) reveals a more frequent[25] multi-word term in the target language than *Asamblea Nacional en Francia.*

[24] See https://www.borm.es/services/anuncio/ano/2017/numero/3482/pdf?id=757271. In fact, the name of the supervising office is now "Consejería de Agua, Agricultura, Ganadería, Pesca y Medio Ambiente" (https://administracion.gob.es/pagFront/espanaAdmon/directorioOrganigramas/fichaUnidadOrganica.htm?idUnidOrganica=123379&origenUO=comunidadesAutonomas&comunidadesAutonomas=true&volver=comunidadesAutonomas&idCCAA=14#.X-DOye1Ce00).

[25] For example, a search in the Spanish reference corpus CORPES (https://webfrl.rae.es/CORPES/view/inicioExterno.view) yields five results against zero.

(9) Normalisation shift (EN-ES translation)

a. on 16 February 2019, the National Assembly in France has adopted the law of programming 2019–2022 and the justice reform [PETIMOD_ORIG_EN]

b. la <Asamblea Nacional de Francia> adoptó el 16 de febrero de 2019 la ley de programación 2019–2022 y la reforma judicial [PETIMOD_MEDIATED_ES]

6 Discussion

In this section, the overall results of our analysis are discussed, focusing on three different quantifications for both translation and interpreting: 1) distribution of the type of shifts retrieved; 2) distribution of the labels of the shifted entities; and 3) the detailed shift entity relationship with all the subcategories of shifts as described above. Then, we will relate our findings to results reported in related literature on intermodal corpora.

Figure 6 quantifies certain tendencies within English-Spanish translations and Spanish-English interpretations in the Committee on Petitions. The most prominent shifts are quantitative shifts (75 instances in both language pairs) and register shifts (33), followed by transformational shifts (18) and normalisations (16). There is a predominance of register shifts in EN-ES translations (20 against 13) and a fairly more balanced number in the case of quantitative meaning shifts (36 against 39). Transformational meaning shifts are more numerous on ES-EN interpretations (2 against 16); inversely, normalisations are more present in the translations into Spanish (11 against 5).

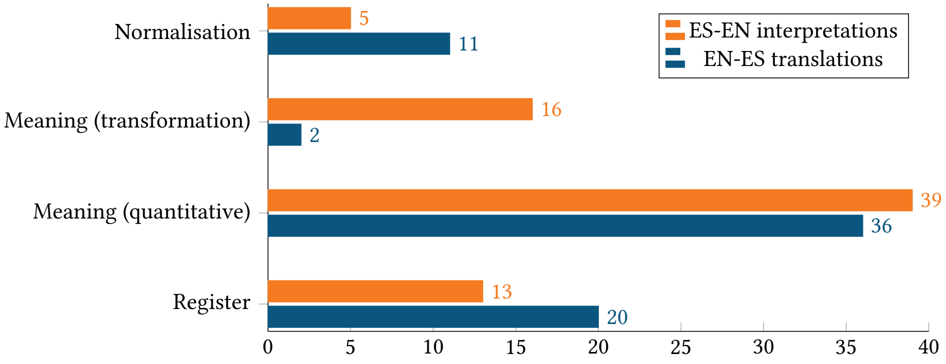


Figure 6: Type of shifts distribution.

Figure 7 shows the distribution of shifted NEs per label, as illustrated in §3 of this chapter. MISC entities are the most frequent (58), closely followed by ORG (51); LOC (22) and PER (11) are considerably less represented in the shifts. The miscellaneous entities are more subject to shifts in the interpretations into English (26 against 32); conversely, organisational entities are prone to shifts in the translations into Spanish (30 against 21). The number of locations remains fairly equal in both directions (12 against 10). Finally, shifts in named persons are almost inexistent in EN-ES translations (only 1 result) in comparison to ES-EN interpretations (10).

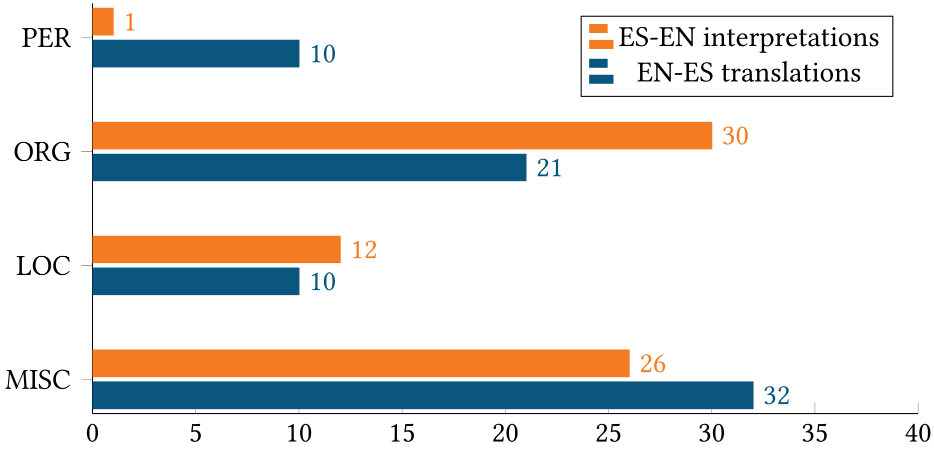


Figure 7: Shifted entities distribution.

In Table 8, 9 and 10, the two types of data commented above (type of shifts and type of entities) are cross-related and broken down into the nine shift subcategories used for this study.

Table 10 contrasts the subcategories of shifts encountered in both directions (EN-ES translations vs. ES-EN interpretations).

In this comparison, major differences can be found which help characterising the shifting profile of each type of transfer separately. It appears that, when operating with named entities:

- English-Spanish translations tend to upgrade register (19), change meaning by contracting (15) and expanding (14), and to normalise multi-word terms (11).

- Spanish-English interpretations, contrarily, tend to downgrade register (12), change meaning by contracting (10) and broadening (24), and to present more transformations, be they partial (8) or total (8).

Table 8: Detailed shift-entity relationship (EN-ES translations)

Type of shift	PER	ORG	LOC	MISC	Total
Register up	0	6	3	10	19
Register down	0	0	1	0	1
Contraction	0	4	2	9	15
Clarification	1	0	2	1	4
Expansion	0	9	1	4	14
Broadening	0	1	1	1	3
Partial transformation	0	0	0	1	1
Total transformation	0	1	0	0	1
Normalisation	0	9	2	0	11

Table 9: Detailed shift-entity relationship (ES-EN interpretations)

Type of shift	PER	ORG	LOC	MISC	Total
Register up	0	0	0	1	1
Register down	0	5	2	5	12
Contraction	4	2	1	3	10
Clarification	0	1	0	2	3
Expansion	1	0	1	0	2
Broadening	3	6	6	9	24
Partial transformation	1	1	0	6	8
Total transformation	1	1	0	6	8
Normalisation	0	5	0	0	5

In general, the results show clear differences in the nature of shifts between EN-ES translations and ES-EN interpretations in the Petitions Committee. Translations from English into Spanish present more frequently register (e.g. *RAMSAR* ⟶ Convención de Ramsar)[26] and normalisation shifts (e.g. *Government of Valencia's Ministry of Agriculture* ⟶ *Consejería de Agricultura de la Generalitat Valenciana*). In the case of register, practically all changes are upwards (*Bulgarian Ministry of Environment and Water* ⟶ *Ministerio Búlgaro de Medio Ambiente y*

[26] These examples were extracted from the most common NE categories in each case according to the correlation Table 8 and Table 9.

Table 10: Comparison of shift subcategories in both directions

Type of shift	EN-ES translation	ES-EN interpretation
Register up	19	1
Register down	1	12
Contraction	15	10
Clarification	4	3
Expansion	14	2
Broadening	3	24
Partial transformation	1	8
Total transformation	1	8
Normalisation	11	5

Recursos Hídricos), as opposed to the downward tendency of the shifts in the interpretations into English (*Comisión de Medio Ambiente del Parlamento Europeo ⟶ ENVI Committee*). The fact that Spanish translators tend to be more formal than English interpreters was a previous intuition confirmed by the data, similarly to the results obtained by Bernardini (2016: 143–144) in her comparative analysis of Italian-English translations. Moreover, results in normalisation bring a new perspective to previous studies, as this is a newly introduced shift category which focuses on changes in specialised multi-word terms instead of general-language collocations. In the case of Bernardini, results showed an increased tendency by Italian-English translators to insert general language collocations. Our data show that normalisation of specialised phraseology is preferred when translating into the Romance language (Spanish) instead.

Moving on to quantitative meaning shifts, the interpretations present a slightly higher amount of them, although it must be specified that they are not of the same type in both directions and modes. While contraction and clarification are more or less equal, expansion prevails overwhelmingly in the EN-ES translations, as in the example: *Association for the Renaissance of Craiova (ARC) ⟶ «Association for the Renaissance of Craiova » (ARC) (Asociación para el Renacimiento de Craiova).* Inversely, broadening is much more numerous in the ES-EN interpretations (*nueve_mil_seiscientas hectáreas ilegales ⟶ considerable illegal construction*). Considering that broadening shifts could be regarded as a simplification feature, our results for the English-Spanish/Spanish-English pair are in line with the bidirectional English<>Italian study of Bernardini et al. (2016), in which interpreters were found to simplify the input more than translators.

Finally, transformations are substantially more present in the ES-EN interpretations, where they are equally distributed among partial (*Ley de Protección Integral del Mar Menor law for iterative protection of the Mar Menor*) and total (*Planes de Ordenación de los Recursos Naturales → natural protection ehm plans*). This is an interesting finding because it presents both similarities and divergences with previous intermodal studies. In Bernardini (2016), for example, transformations were also absent from English-Italian translations, but far more present in the other subcorpora, and the "partial" category outnumbered the "total" one. Although this could be the result of different conceptualisations by the researchers on what "transformation" means, it can also be argued that dissimilarities in transformational behaviour are connected to Ferraresi & Miličević's (2017: 1) "cognitive and task-related constraints" characterising the translation and interpreting processes. In other words, the number and nature of the transformations operated by the translator and/or interpreter could be strongly dependent on factors beyond language direction or mode, such as the communicative situation in which they are working (e.g., whether the context is a plenary session of the Parliament or a Committee meeting) or even the topic of the source text.[27]

Precisely with the goal of shedding some light on the connections between topic (or specialisation field, etc.) and the shifts involved in translation and interpreting, discussion should also centre on the shifted NEs label distribution shown in Figure 7. A clear majority of miscellaneous and organisational entities over locations and proper nouns of persons can be observed both in EN-ES translation and ES-EN interpreting. These results picture a cognitive domain of a rather political nature, in which parties, public platforms and similar organisations discussing policies and agreements are more important than the places where the problem occurred or the persons who complained in the first place. This perspective suits the function of the Committee on Petitions and points indeed towards a supranational way of making politics which permeates through the shifts encountered in translation and interpreting. What is more, a closer examination of Table 8 and 9 reveals that there is a high degree of relationship between the frequencies of shifts and entities in both analysed directions and modes. For example, the frequent upward register shifts in EN-ES translation often occur in MISC NEs (*EU law → Derecho de la Unión*), and the numerous broadening shifts in the ES-EN interpretations are usually operated on ORG NEs (*departamentos de Ecología de la Universidad de Murcia → the University of Murcia and its researchers*). Introducing this new parameter in the analysis of shifts could add a

[27] These factors could also affect the degree of relation between total transformation shifts and translation/interpreting errors suggested in §3.

new variant to the conclusions of Ferraresi et al. (2018) and lead us to hypothesise that simplification is a contingent feature which depends not only on the mediation mode and the source languages involved, but also on the topic of the source text. This is in line with calls for multifactorial research designs in empirical translation/interpreting studies (Corpas Pastor 2008, De Sutter & Lefer 2020), since studies that take into account only one or two explanatory factors fall short of explaining the complexity of real-world translation/interpreting phenomena. Under this view, the analysed ES-EN interpretations of the Committee on Petitions would be more simplified than the EN-ES translations not just because they are an oral mediation performed into English, but also because the interpreters (consciously or not) would apply certain strategies aimed at approaching the content of their message to a broader audience than translation. This would imply neutralising or simplifying institutional-specific MISC and ORG NEs (EU legislation, international agreements, public bodies, etc.), paradoxically the most unfamiliar in the ears of the European citizens who could also exercise their right of petition.

7 Conclusion

The study presented in this chapter can be regarded as innovative for various reasons. To the best of our knowledge, it is one of the first corpus-based studies which relies on translated and interpreted documents from the European Parliament Committee on Petitions. Secondly, it does not only build and employ a type of resource which is still in its infancy (intermodal corpora), but also introduces a new methodological layer through manual and state-of-the-art automatic named-entity recognition (the latter performed by spaCy). This approach added new aspects to the analysis of translation and interpreting shifts (a new shift category called "normalisation" and the possibility of correlating shifts to the semantic labels of the NEs involved), which in turn helped establish interesting findings in relation with previous studies (normalisation as a language-dependent feature of translation, transformation and simplification as contextual, topic-dependent features of interpreting).

Our study presents some limitations, though. Concerning methodology, the suitability of the selected transcription conventions must be revised. Even though we introduced certain modifications to the system, some of the proposed features seem more adequate for multimodal corpora and are counterproductive when recognising NEs (consider for example the hesitation particle *ehm* in *the Socialist for ehm Party ehm from the ehm Murcia region*). NE recognition and corpus-based shift analysis could also be extremely facilitated with the addition of an intermediate alignment phase to cope with terminology variation. In fact, monolingual terminological variation within NEs (e.g., *The Court, Court of Justice of the European Union*, *CJEU*, etc.) turned manual pairing into an exhausting job. As to NE labelling, a more fine-grained taxonomy is also needed for both languages, especially in the MISC category, where additional subtypes not available in the VIP scheme could be traced during the analysis (e.g. agreements like *Ramsar* or quasi-legal documents such as *Estrategia de Gestión Integrada*, among others). Undoubtedly, a tailor-made labelling system like this would considerably increase the quality of the correlating shift-entity results. In addition, the spaCy script integrated in the VIP NER module has been trained on two different language models, which could also account for the differences in precision and recall (685,000 word vectors in English as opposed to 500,000 word vectors in Spanish).

Finally, the NLP-enhanced orientation to the analysis of intermodal corpora presented in this chapter helped envisage a new line of research which does not hold translation and interpreting universals as an unconditional reality, but as a theoretical basis which is given in different degrees in the texts, depending on variants such as the languages and directions involved, the mode of mediation, and even the semantic content of the named entities conveyed. Therefore, multifactorial research designs are needed to capture the multitude of factors that have an influence on the observed phenomena. Although more studies are needed to determine the exact relevance of these semantic categories in translation and interpreting shifts, it would seem that the final goal is finding a transversal set of norms which could break the theoretical differences between translation and interpreting, focussing the discussion on the coordinates or *function* of the mediation instead of the mediating mode itself.

Acknowledgements

The research reported in this article has been funded by the Spanish Ministry of Education and Professional Training (ref. FPU18/05803). The article has also been carried out in the framework of the projects VIP (FFI2016-75831-P), TRIAGE (UMA18-FEDERJA-067) and MI4ALL (CEI-RIS3).

References

Baisa, Vít, Jan Michelfeit, Marek Medved & Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 2799–2803.

Bakker, Matthijs, Cees Koster & Kitty Van-Leuven Zwart. 2009. Shifts. In *Routledge encyclopedia of translation studies*, 269–274. London: Routledge.

Bendazzoli, Claudio. 2010. The European Parliament as a source of material for research into simultaneous interpreting: Advantages and limitations. In Lew N. Zybatow (ed.), *Translationswissenschaft: Stand und Perspektiven*, 51–68. Frankfurt am Main: Peter Lang.

Bernardini, Silvia. 2016. Intermodal corpora: A novel resource for descriptive and applied translation studies. In Gloria Corpas & Miriam Seghiri (eds.), *Corpus-based approaches to translation and interpreting: From theory to applications*, 129–148. Frankfurt: Peter Lang. DOI: 10.3726/b10354.

Bernardini, Silvia, Adriano Ferraresi & Maja Miličević. 2016. From EPIC to EPTIC: Exploring simplification in interpreting and translation from an intermodal perspective. *Target* 28(1). 61–86. DOI: 10.1075/target.28.1.03ber.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard & Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In Mariachiara Russo, Claudio Bendazzoli & Bart Defrancq (eds.), *Making way in corpus-based interpreting studies*, 21–42. Singapore: Springer. DOI: 10.1007/978-981-10-6199-8_2.

Butterfield, Jeremy (ed.). 2015. *Fowler's Dictionary of Modern English Usage.* Oxford University Press. DOI: 10.1093/acref/9780199661350.001.0001.

Calzada-Pérez, María. 2017. Researching the European Parliament with Corpus-Assisted Discourse Studies: From the micro- and macro-levels of text to the macro-context. en. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 30(2). 465–490. DOI: 10.1075/resla.00003.cal. http://www.jbe-platform.com/content/journals/10.1075/resla.00003.cal (17 May, 2022).

Corpas Pastor, Gloria. 2001. Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS. Revista de Traductología* 5. 155–184. DOI: 10.24310/trans.2001.v0i5.2916.

Corpas Pastor, Gloria. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma.* Frankfurt: Peter Lang.

De Sutter, Gert & Marie-Aude Lefer. 2020. On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives* 28(1). 1–23. DOI: 10.1080/0907676X.2019.1611891.

European Parliament. 2020a. *European Parliament Committees: About PETI.* https://www.europarl.europa.eu/committees/en/peti/about.

European Parliament. 2020b. *Petitions FAQ.* https://petiport.secure.europarl.europa.eu/petitions/en/faq/det?questionor=1&sectionor=1.

European Union. 2012. *Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union: charter of fundamental rights of the European Union.* Publications Office. DOI: 10.2860/58644.

European Union. 2021. *Interinstitutional style guide.* Luxembourg: Publications Office of the European Union. DOI: 10.2830/400520.

Ferraresi, Adriano, Silvia Bernardini, Maja Milicevic Petrovic & Marie-Aude Lefer. 2018. Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta : Journal des traducteurs / Translators' journal* 63. 717–738. DOI: 10.7202/1060170ar.

Ferraresi, Adriano & Maja Miličević. 2017. Phraseological patterns in interpreting and translation: Similar or different? In Gert De Sutter, Marie-Aude Lefer & Isabelle Delaere (eds.), *Empirical translation studies: New methodological and theoretical traditions*, 157–182. Berlin: Mouton De Gruyter. DOI: 10.1515/9783110459586-006.

Gaber, Mahmoud, Gloria Corpas Pastor & Ahmed Omer. 2020. Speech-to-Text technology as a documentation tool for interpreters: A new approach to compiling an ad hoc corpus and extracting terminology from video-recorded speeches. *TRANS. Revista de Traductología* 5. 263–281. DOI: 10.24310/TRANS.2020.v0i24.7876.

Goffin, Roger. 1994. L'eurolecte: Oui, jargon communautaire: Non. *Meta: Journal des traducteurs* 39(4). 636–642. DOI: 10.7202/002930ar.

Hurtado, Amparo. 2017. *Traducción y traductología: Introducción a la traductología.* 9th. Madrid: Cátedra.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, vol. 5, 79–86. Phuket: AAMT.

Martin, Anne. 1997. *Tratamiento de las referencias de carácter institucional del mundo de habla inglesa en la prensa española.* Universidad de Granada. (Doctoral dissertation).

Martínez, José & Iris Serrat. 2012. ECPC: El discurso parlamentario europeo desde la perspectiva de los estudios traductológicos de corpus. *Linguamática* 4(2). 65–73.

Mayoral, Roberto. 1999. La traducción de referencias culturales. *Sendebar: Revista de la Facultad de Traducción e Interpretación* 10-11. 67–88.

Mori, Laura (ed.). 2018. *Observing Eurolects: Corpus analysis of linguistic variation in EU law*. Amsterdam: John Benjamins. DOI: 10.1075/scl.86.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy & James R. Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence* 194. 151–175. DOI: 10.1016/j.artint.2012.03.006.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London; New York: Routledge.

Ortega, Juan Miguel. 2002. La traducción de referencias culturales de carácter institucional y político a través de un caso práctico. *Puentes* 1. 21–32.

Prieto Ramos, Fernando (ed.). 2019. *Institutional translation for international governance: Enhancing quality in multilingual legal translation* (Bloomsbury Advances in Translation Studies). New York: Bloomsbury Publishing.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli & Nicoletta Spinolo. 2012. The European Parliament Interpreting Corpus (EPIC): implementation and developments. In Francisco Straniero Sergio & Caterina Falbo (eds.), *Breaking ground in corpus-based interpreting studies*, 53–90. Bern: Peter Lang. DOI: 10.3726/978-3-0351-0377-9.

Saldanha, Gabriela. 2009. Linguistic Approaches. In Mona Baker & Gabriela Saldanha (eds.), *The Routledge encyclopedia of translation studies*, 2nd edn., 148–152. London: Routledge.

Shlesinger, Miriam. 2008. Towards a definition of interpretese: An intermodal, corpus-based study. In Gyde Hansen, Andrew Chesterman & Heidrun Gerzymisch-Arbogast (eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile*, 237–253. Amsterdam: John Benjamins. DOI: 10.1075/btl.80.18shl.

Ustaszewski, Michael. 2019. Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives: Studies in Translation Theory and Practice* 27(1). 107–123. DOI: 10.1080/0907676X.2018.1485716.

Veroz González, María Azahara. 2014a. El Registro Público de Documentos del PE como recurso documental en la traducción especializada: Elaboración de bases de datos terminológicas con corpus en Multiterm. *Hikma* 13. 125. DOI: 10.21071/hikma.v13i.5229.

Veroz González, María Azahara. 2017. Translation in the European Parliament: The study of the ideational function in technical texts (EN/FR/ES). *Meta* 62(1). 19–44. DOI: 10.7202/1040465ar.

Veroz González, María Azahara. 2014b. *La traducción en el Parlamento Europeo. Estudio de los textos técnicos y de comunicación administrativa*. Universidad de Córdoba. (Doctoral dissertation).