Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4044–4053
Marseille, 11–16 May 2020

EMPAC: an English–Spanish Corpus of Institutional Subtitles

Iris Serrat Roozen*, **José Manuel Martínez Martínez**†
*Universidad Católica de Valencia San Vicente Mártir
Valencia, Spain
iris.serrat@ucv.es

†Universität des Saarlandes
Saarbrücken, Germany
j.martinez@mx.uni-saarland.com

Abstract

The EuroparlTV Multimedia Parallel Corpus (EMPAC) is a collection of subtitles in English and Spanish for videos from the European Parliament's Multimedia Centre. The corpus has been compiled with the EMPAC toolkit. The aim of this corpus is to provide a resource to study institutional subtitling on the one hand, and, on the other hand, facilitate the analysis of web accessibility to institutional multimedia content. The corpus covers a time span from 2009 to 2017, it is made up of 4,000 texts amounting to two and half millions of tokens for every language, corresponding to approximately 280 hours of video. This paper provides 1) a review of related corpora; 2) a revision of typical compilation methodologies of subtitle corpora; 3) a detailed account of the corpus compilation methodology followed; 4) a description of the corpus; 5) the key findings are summarised regarding formal aspects of the subtitles conditioning the accessibility to the multimedia content of the EuroparlTV; and 6) some final remarks and ideas for future research.

Keywords: subtitles, corpus, institutional translation, accessibility, English, Spanish

1. Introduction

The Internet has brought about vertiginous changes in the current social structures and has generated profound transformations in almost all social spheres. Nowadays, the Internet is an indispensable tool for working, interacting with public institutions, and having fun, and it is undoubtedly an instrument that can foster a more inclusive, heterogeneous, and fairer form of coexistence.

Advances in technology and the Internet have also generated great changes in the audiovisual industry and in audiovisual translation, so much so that, as (Orrego, 2013) points out, the ever-increasing speed of data transfer and availability of data storage have turned the Internet into the perfect habitat for audiovisual content, to the detriment of purely textual content. Therefore, it seems logical to remember that if we wish for a greater number of users to be able to enjoy videos hosted on the web, there is a need to develop fundamental techniques to make access to published material possible, such as subtitling for deaf people and hard of hearing (SHD), audio description (AD), and conventional subtitling, among other forms of audiovisual translation.

Therefore, we are faced with a scenario in which audiovisual translation (AVT) on the Internet should be central, which, in turn, requires that substantial research is conducted in this area. However, despite the proliferation, in recent years, of works on audiovisual accessibility in other media (Pedersen, 2017; Romero-Fresco, 2019), studying audiovisual content on the Internet is still, at present, practically uncharted territory.

With the aim of promoting studies in this area, we have analysed material hosted on the European Parliament's website, specifically on the EuroparlTV online television channel. This platform, which distributes audiovisual material subtitled in 24 languages, has allowed us, firstly, to study online subtitles generated by translation professionals in the institutional context and, secondly, to analyse compliance with the accessibility requirements for audiovisual content laid out by the Web Content Accessibility Guidelines (WCAG) 2.0—regulations developed by the Web Accessibility Initiative (WAI) of the World Wide Web Consortium (W3C). In order to conduct a study of these characteristics, we have created the EuroparlTV Multimedia Parallel Corpus (EMPAC), which combines the English and Spanish subtitles of the videos broadcast on EuroparlTV between 2009 and 2017.

The parameters that we have decided to study are those that define the subtitling and which will allow us to identify whether the commonly accepted norms for this form of AVT in traditional media (TV, DVD, Cinema, etc.) are also a feature of online subtitling. These parameters are part of the space-time dimension proposed by Díaz Cintas and Remael (2007), on which the pillars of subtitling are based. As Díaz Cintas and Remael (2007, p. 95) note:

> It is very frustrating and disconcerting to see how the subtitle disappears from the screen when we have not yet finished reading it, [...] the typical occasion in which we feel that we have 'read' rather than 'watched' the film.

Thus, in our study, we analyse subtitle reading speed in terms of the following variables: reading speed in characters per second (cps), pause between subtitles, number of lines, and number of characters per line. We go into greater depth by studying segmentation of subtitles in lines (intra-subtitle segmentation) and the segmentation of sentences across subtitles (inter-subtitle segmentation) attending to linguistic criteria, which as various studies point out, is related to the speed at which we read and assimilate the text written in the subtitle.

Our analysis of the segmentation variable is based on a proposal for analysing possible cases of inadequate segmenta-

tion. This is carried out using the CQPweb's graphic interface, firstly, to perform automated searches in EMPAC for possible inappropriate segmentations, which constitutes one of the true innovations of the present work, since, in our opinion, such descriptive and quantitative examinations are entirely missing from the literature, in contrast, to prescriptive recommendations which can be found in different handbooks and guidelines. Secondly, the CQPweb's graphic interface allows us to calculate the number of cases contained within the corpus and to examine these on a year-by-year basis or programm type basis.

The following sections of this paper are organized as follows. The second section offers a brief review of other similar corpora. The third section introduces the design criteria of the corpus. The fourth section explains the methodology to compile the corpus. The fifth section describes the corpus obtained. The sixth section illustrates some empirical findings. The last section concludes the paper offering some final remarks, and directions for future research.

2. Related work

Before creating EMPAC we checked if there were already existing corpora of subtitles fulfilling the following criteria:

1. freely available
2. institutional translation
3. professional translation
4. online content
5. English-Spanish
6. containing enough technical information
7. size

We review first freely available subtitling corpora that ideally should cover institutional online content. We check the contents of institutional translation corpora to see if they contain subtitles. Then we revise subtitling corpora and see if they contain institutional subtitles. Next, we reflect on the formats and technical information provided in the reviewed corpora. Next, we highlight what is missing and justify the need to compile EMPAC and how this corpus fills a gap. Finally, we outline the typical methodologies.

2.1. Corpora

This review focus on corpora that preferably fulfil one or more of the criteria introduced above, specially, to be freely available and to contain an aligned English-Spanish version. The vast majority of research in corpus-based translation studies on subtitling rely upon small *ad hoc* corpora that are not representative enough and that are not freely available due to copyright reasons. These types of corpora are not reviewed in this paper.

To the best of these authors knowledge there is no corpus of subtitles representing institutional translation of online content. The most famous parallel corpora of institutional translation are resources created from textual material produced by international organizations like the European Union (Steinberger et al., 2014) or the United Nations (Ziemski et al., 2016). Translation plays a central role in this institutions and an enormous collections of aligned documents in many different language pairs have been collected for internal use to feed translation memories to train MT systems and, finally, packaged and released for the public. However, we cannot find any collection of subtitles within those corpora where the typical text types are basically legal and administrative documents, press-releases, reports, meeting minutes, official records and other parliamentary documents of different institutions that are publicly available. In some cases, because the purpose of the collections were to create multilingual NLP tools or MT systems the design of the corpora and the information preserved about the texts is not always very helpful to pursue translation research as pointed by Karakanta et al. (2018).

The biggest collection of aligned subtitles freely available is the OpenSubtitles corpus (Tiedemann, 2008; Lison and Tiedemann, 2016; Tiedemann, 2016; Lison et al., 2018). This corpus is made up of subtitles for films and TV series, therefore, it does not cover institutional translation. It is not clear if the subtitles are representative of online content, as it could well be that this subtitles correspond to content distributed in DVD. Moreover, the authorship of the translations is not clear as (Pryzant et al., 2018) explain in a similar experience compiling a parallel corpus of English-Japanese subtitles, the subtitles are often the "official" translation (probably ripped from DVDs or downloaded from public or private streaming platforms). Being that the case, we could assume that the translation is the output of a professional translator; but it could also happen that the subtitles are produced by amateur translators (fansubs) or, what could be worse, by automatic tools, therefore, the quality of the subtitles could be compromised.

Another source of freely available online content subtitled and translated are TED talks. Several initiatives (Cettolo et al., 2012; Zeroual and Lakhouaja, 2018; Di Gangi et al., 2019) have created corpora from this source of data. In any case, the register is not representative of institutional translation. The transcriptions, subtitles and translations are produced by volunteers, not professional translators. We want to mention here the comparative study of a MT system for subtitles trained with a corpus from OpenSubtitles and another trained with TED talks that showed that TED talks tended to yield worse results in the translation of subtitles (Müller and Volk, 2013). This result indicates that there might be differences between subtitles and TED transcripts.

Bywood et al. (2013) describe the compilation of a corpus containing first and foremost professional translations of films and series. The initial corpus was extended with a corpus of texts from the EuroParl. To our knowledge, this corpus is the biggest collection of professional subtitles used to train a MT and to build a queriable parallel corpus. Sadly, it has not been released and it is not available for the public.

2.2. Compilation methods

We have reviewed a series of articles (Xiao and Wang, 2009; Cettolo et al., 2012; Fishel et al., 2012; Bywood et al., 2013; Lison and Tiedemann, 2016; Lison et al., 2018; Zeroual and Lakhouaja, 2018; Pryzant et al., 2018; Di Gangi et al., 2019) to identify common practices regarding the compilation of subtitle corpora. The following steps can be found in most of the proposals:

- acquiring the data (subtitles and metadata)

- model subtitle format (eg. SRT) to some other format (XML, YAML...)
  - parsing the file
  - character encoding conversion
  - text normalization
  - metadata extraction
  - metadata inclusion
- linguistic processing:
  - tokenization
  - sentence splitting
  - PoS tagging
- alignment
  - document: find/match all the versions of the same video
  - sub-document level: caption vs sentence
  - word alignment
- indexation for queries

For the acquisition of the data there are two approaches that have often to be combined: 1) to obtain a dump of the data by the managers of the original resources, and; 2) to scrap the websites or query databases to obtain the required data.
Modelling the subtitles from an input format into the structured target format requires to cope with different input formats and to do some preprocessing of the data. The typical problems that can arise are related to unknown/proprietary formats, character encoding, and strange characters. Some of the metadata of interest have to be extracted/derived from the same subtitles (eg. number of subtitles per text, number of lines per subtitle, number of characters per subtitle...).
As for the output format, most of the corpora are released in XML (Cettolo et al., 2012; Karakanta et al., 2018; Lison and Tiedemann, 2016; Lison et al., 2018; Tiedemann, 2008; Barbaresi, 2014; Zeroual and Lakhouaja, 2018; Steinberger et al., 2014; Ziemski et al., 2016), although YAML has been used in the modelling of recent corpora (Di Gangi et al., 2019; Pryzant et al., 2018).
For the linguistic processing, it is helpful to verify that the data is in the expected language, and then proceed with the tokenization, the lemmatization, the PoS tagging, and the sentence splitting. Being the latter challenging as subtitles follow different conventions regarding punctuation compared to other written formats (which are typically the model for most sentence splitters).
The alignment of this collections of texts is one of the most challenging parts. First, the alignment at document level (to match the different versions of the same video) is difficult as there are sometimes no available metadata. Therefore, an algorithm based on the naming conventions of the documents, on the metadata gathered in the previous steps, or some analysis of the content is needed. The alignment at this level can also imply having several versions in the same language for the same video.
Once a pair of documents have been identified the next question arises: shall the subtitles be aligned at subtitle or at sentence level? Most of the corpora are aligned at sentence level, however, there have been experiments and criticism supporting the alignment at subtitle level.

The main arguments justifying sentence alignment according to (Bywood et al., 2013) are:

> [...] due to varying word order in the source and target languages, word and phrase translations in the target subtitles might actually appear in different subtitles in the source. In addition to this, many annotation tools (such as dependency parsers) expect sentences as input; thus sentence-aligned corpora are more compatible with linguistic annotation. Finally, a sentence-aligned subtitle corpus can be more compatible with other sentence-aligned material, e.g. if used as an out-of-domain corpus.

However, Bywood et al. (2013, pp. 603–604) report better MT translation performance for systems trained with alignments at subtitle level than at sentence level at least for the translations from English into Spanish and vice versa. The poorer performance of MT of subtitles aligned at sentence level might be due to errors in the automatic extraction of sentences from the source subtitles. Most algorithms for sentence boundary detection rely on punctuation and/or syntactic parsing. Sadly, subtitles do not always follow the same punctuation conventions as regular written texts. Moreover, we are not aware of any syntactic parser trained with subtitles or audiovisual texts but more distant registers and genres like news stories.
Regardless of the level of granularity of the sub-document alignment, most algorithms make the most of an additional source of information not available for typical alignment of written text: time. The subtitles include information about the time they appear or disappear on screen. The assumption would be that for the sake of synchrony between what is being said on screen and what is rendered in the subtitles, subtitles will contain similar information for the same time spans. This is more often the case if a master template has been used to keep a common segmentation of subtitles across languages.
Alignments at word level are produced, specially if a MT system has been trained with the data.
Finally, some corpora are prepared and indexed to be queried as a parallel corpus like the OpenSubtitles at OPUS (Tiedemann, 2012) using the Open Corpus Workbench (OCWB) (Evert and Hardie, 2011) as its backend and queries use the CQP query language or the SUMAT corpus (Bywood et al., 2013) using Bilingwis as search interface (Weibel, 2014; Volk et al., 2014).

2.3. Criticism on technical aspects of subtitles

Many of the above mentioned resources have been built to offer to the community parallel corpora as training and evaluation data for multilingual NLP applications, specially machine translation. Both from the machine translation community (Karakanta et al., 2019) and from translation studies there is a gap on corpora including explicit information about the spatial and temporal constraints of subtitles.
The two main features modelling those constraints are the length of the subtitles and the lines, and the reading speed on the one hand; and the line breaks on the other hand.

3. About EuroparlTV

EuroparlTV is the online television channel of the European Parliament, now called the Multimedia Centre of the European Parliament. This channel was born in September 2008 in an attempt to bring closer this institution to the citizens (specially the youngest ones) through a modern and creative medium. The goal was to inform the citizens about the activities and decisions of the European Parliament. From the very beginning, it became unique in its class by offering videos subtitled into more than 20 languages.

The contents offered at EuroparlTV are classified into five main categories: Background, Discovery, History, Interviews and News. The news provide coverage of the business at the European Parliament and also contents on current affairs and hot issues through daily news, interviews, reportages and programmes. Discovery contains a series of videos aimed at teachers and school-age kids in an effort to explain European affairs to youngsters. The background and history channels gather together, as their names indicate, background and historical content on the European Union, in general, and, on the European Parliament, in particular. The interviews section features interviews of Members of the European Parliament and other authorities.

EuroparlTV constitutes a representative sample of the subtitles produced for online content distributed through the Internet by the European Union institutions and agencies. These subtitles are illustrative of the EU's best efforts to accomplish its commitments in regard to accessibility by granting the access and understanding of the videos to the highest number of persons by means of subtitles. An effective tool to overcome linguistic barriers and also to satisfy the needs of citizens with hearing impairments.

4. Corpus Compilation

The corpus was compiled using the EMPAC toolkit[1], a set of utilities devised to compile versions in English and Spanish of the corpus and to align them at document and subtitle level. The toolkit fulfils a three-fold goal: 1) to automate the process as much as possible; 2) to document it; and 3) to enable reproducible research[2]. The EMPAC toolkit is written in Python and released under a MIT License which is very permissive regarding copy, modification and distribution if the terms of the license are observed.

The main steps to compile the corpus are:

1. Download of SRT files
2. Modelling subtitles as XML
3. Annotation of linguistic information:
   (a) tokenisation, PoS tagging, lemmatization
   (b) sentence boundary detection
4. Alignment at document and subtitle level
5. Encoding the corpus for CQPweb

[1] `https://bitbucket.org/empac/toolkit/src/master/`

[2] The crawler at the time of writing this paper does not work anymore because the website of EuroparlTV has changed significantly in 2018.

4.1. Download of SRT Files

The European Parliament's Multimedia Centre provides subtitles for many of its videos in SRT format. These assets can be found by visiting the page presenting the video and looking into the source code for the URL pointing to the SRT file in the desired language. This is a very repetitive and time consuming task prone to errors. Therefore, a simple crawler was written.

The script executes the following actions: 1) builds a query to search in the repository of the Multimedia Centre given several parameters; 2) paginates the results; 3) for each of the results: a) retrieves the URL of the video and gets other metadata; b) downloads the subtitles in SRT format; c) saves the metadata in an Excel spreadsheet.

Several optional parameters can be passed to download only subsets of the materials available at the Multimedia Centre:

- `from` and `to` to filter the results for a given time period;
- `lang` to specify the language of the subtitles;
- `type` to choose the type of programme (loosely equivalent to the notion of register), namely: 1) news, 2) interview, 3) background, 4) discovery, 5) history, and 6) others.
- `category`: classification of the videos by its field or domain, namely: 1) EU affairs, 2) economy, 3) security, 4) society, 5) world, and 6) others.

The metadata fields retrieved at this stage are:
a) the language of the subtitle in two-letter ISO code (eg. `en` for English); b) the title of the video; c) the URL of the audiovisual text; d) the type of programme; e) the category of the audiovisual text; f) the date when the video was (re-)published in the platform; g) a brief description/summary of the contents of the text; h) the duration of the video in hours, minutes and seconds in the format HH:mm:ss; i) the URL of the SRT file; and j) a unique ID based on the name of the SRT file, which must be a valid MySQL handle.

4.2. Modelling Subtitles as XML

The SRT files downloaded during the previous stage are transformed into XML describing explicitly the structure of the document (subtitles and lines). Several textual attributes are calculated at text, subtitle and line level, and the metadata of the text are added. The `SRT` input is a file in plain text format where subtitles are separated by an empty line. The first line corresponds to the position of the subtitle within the document (from 1 to n), the second line contains the start timecode and end timecode of the subtitle, and the next lines (one or two) are the text that is displayed on the screen split into lines. Timecodes provide a time reference to synchronise the display of each subtitle with the video signal indicating the time that the subtitle should appear on the screen, and the time it should disappear. The timecode format used in `SRT` files is indicated in hours, minutes, seconds, and milliseconds in the format `HH:mm:ss,SSS`. Timecodes are crucial also to calculate other time-related metrics like the duration of the subtitle, the reading speed in characters per second, or the pauses between subtitles.

The output is a well formed and valid XML file containing in a structured manner all the information extracted so far. The script taking care of this task executes the following algorithm:

1. loads the metadata file generated during the download stage;
2. loops over all SRT files located in the input folder;
3. checks that the language of the file is right;
4. parses the information contained in the SRT file to create a list of subtitles, and then:
   (a) it creates a root element **<text>**;
   (b) for each subtitle identified it appends to the root a children element **<subtitle>** which, in turn, can contain one or more **<line>** elements which contain the actual text of the subtitles; and, finally
   (c) it adds a number of attributes at text, subtitle and line level describing formal properties of each element.

At text level, it adds the metadata fields collected by the crawler (see the previous Section 4.1.) and it incorporates other information derived from the parsing of the SRT like a) the total number of subtitles (`subtitles`); b) the number of subtitles made of one line (`one_liners`); c) the number of subtitles made of two lines (`two_liners`); d) the number of subtitles made of more than two lines (`n_liners`).
At subtitle level, it provides information like a) the position in the text (`no`); b) the timecode from which the subtitle starts to be displayed (`begin`); c) the timecode from which the subtitle is not displayed anymore (`end`); d) the duration in seconds of the subtitle (`duration`); e) the total number of lines (`n_lines`); f) the total number of characters (`chars`); g) the reading speed in characters per second (`cps`); and h) the time elapsed between the end of the previous subtitle and the beginning of the current one (`pause`).
At line level, it adds the following attributes: a) the line number or its position in the subtitle (`no`); b) the number of characters (`chars`).

4.3. Annotation of Linguistic Information

The annotation of linguistic information consists of two processes: 1) the annotation of morpho-syntactic information (tokenization, lemmatization, and PoS); and, 2) splitting the text into sentences.

4.3.1. Tokenization, Lemmatization and PoS Tagging

The linguistic annotation of the subtitles identifying tokens, lemmas and PoS was carried out using the utilities of the wrapper of the TreeTagger wrapper[3], which in turn uses mytreetaggerwrapper[4], and the TreeTagger (Schmid, 1995). The process consists of four steps:
1) text normalization; 2) linguistic annotation with TreeTagger; 3) postprocessing; and 4) enrichment of the XML elements **<text>**, **<subtitle>**, and **<line>** with token information.
Before annotating any linguistic information with the TreeTagger, the XML files are preprocessed performing character normalization on the XML's text contents to obtain better results with the parser. This version of the XML files is the NORM version of the corpus.
Then, the text is annotated with the TreeTagger using a wrapper that loads the parameter files for English or Spanish just once, and handles the creation of well-formed and valid XML with the text contents in verticalized (VRT) format. The VRT format is the expected input to encode texts as corpus with the Open Corpus Workbench (OCWB)—the corpus index engine and query processor powering CQPweb—and it presents structural information as XML elements with their attributes, and positional attributes corresponding to the tokens, represented as one token per line, and in each line, word form, PoS and lemma delimited by tabulations.
Next, the VRT files are postprocessed fixing some issues introduced during the previous step.
And finally, a `tokens` attribute is added to every **<text>**, **<subtitle>**, and **<line>** element indicating the number of tokens within the scope of the given element.

4.3.2. Sentence Splitting

Sentence boundaries are annotated separately in a different XML file. As the tokenization is the same as in VRT files produced in previous steps, it makes possible to adopt a multi-layer stand-off annotation paradigm which grants well-formed valid XML and eases managing the annotations in a modular fashion.
Sentence splitting is carried out using the `punkt` sentence tokenizer of the NLTK (Bird et al., 2009). The output is an VRT file with the same token stream produced by TreeTagger. The sentences are delimited with the element **<s>** and carry the attributes `no` denoting the position in the text, and `tokens` indicating the length of the sentence in tokens. The output of this step is saved in the SENTS folder of the corpus.

4.4. Alignment at Document and Subtitle Level

Once the texts have been annotated with linguistic information, documents and subtitles are aligned using a Python script. This programme takes all the XML files in one version (English) and all the files from another version (Spanish) and tries to match them at document level. Once the pairs of documents are found, it proceeds with the alignment at subtitle level.
The alignment at document level follows a very simple heuristic. First, the record attribute of the English file is obtained. Then, all documents matching the record attribute in the Spanish target version are found. If more than one file was found, the one with the same number of subtitles is used.
Regarding the alignment of the texts at subtitle level, the number of subtitles should be the same in the both versions of the documents. Moreover, the timecode for the begin and the end of the subtitles tend to be the same for all versions regardless of the language, indicating the usage

[3] `https://github.com/chozelinek/wottw`
[4] `https://github.com/chozelinek/mytreetaggerwrapper`

of a master template (the text is segmented in subtitles for one language, and the same segmentation is used for all the other versions). In general, this means that if the number of subtitles in the two versions to be aligned is the same, and the start and end timecodes of each subtitle are the same in both versions, a one-to-one alignment is assumed. The vast majority of the subtitles follow this alignment pattern. If the number of subtitles is not the same, or some mismatch is found regarding the timecodes, an alignment of the subtitles based on the timecode is used. This alignment method relies in comparing the timecodes of each subtitle of the source version with those of all the subtitles in the target version, if any subtitle in the target version has a start timecode greater or equal than the source start timecode and an end timecode smaller or equal than the source end timecode both subtitles are aligned. The alignments for each segment (the minimal mapping between subtitles of the source text and the target text) are saved in a third file. The format of the alignment file is the one used to import alignments for the OCWB. Each line represents a segment, on the left hand side of the line are recorded the subtitle `id` attribute(s) of the source version, and on the right hand side the attribute(s) of the subtitles of the target version.

4.5. Encoding the Corpus for CQPweb

Finally, a shell script performs the encoding of each version of the corpus (English and Spanish) with the OCWB, it generates the alignment at document and subtitle level, and it imports those alignments for the OCWB indices.
The information encoded are texts' metadata, texts' structure, the linguistic information at token level, and the sentence boundaries. All this information is indexed to enable CQP queries.
The alignments are incorporated into the indexed corpus enabling both the visualization of a subtitle and its aligned version in the other languages, and parallel queries.
The encoded corpus is accessible through an installation of CQPWeb.

4.6. Wrong Subtitle Segmentation Identification

Once the corpus was encoded and accessible through CQPweb, a set of queries were designed to extract and annotate wrong cases of intra- and inter-subtitle segmentation for both English and Spanish. The instances found are organized by grammatical categories. The approach is a rule-based alternative to the Chink-Chunk algorithm used in Karakanta et al. (2019).

5. Corpus Description

The resulting corpus[5] is released in different formats or versions, namely:

a) `SRT`, the original subtitle files as published at the European Parliament's Multimedia Centre in the SRT format, one of the most popular and almost a de facto standard in the web;
b) `XML`, the subtitles modelled in XML containing metadata and formal attributes describing the subtitles and the lines;
c) `NORM`, the XML version for which the text has been normalized to improve the quality of the morpho-syntactic parsing;
d) `VRT`, the texts in the verticalized format expected by the OCWB to encode the corpus—one token per line, and positional attributes (word, PoS, and lemma) separated by tabulations—annotated with TreeTagger; and,
e) `SENTS`, the texts split in sentences with the NLTK in verticalized format.

Moreover, the alignments English-Spanish at subtitle level are also provided in the tabular format that the `cwb-align-import` tool from the OCWB uses to import alignments of structural elements.

5.1. Size of the Corpus

We describe the size of the corpus in this section according to the four main variables: a) language, b) year, c) type, and d) category.

5.1.1. Language

The corpus is made up of two subcorpora: English and Spanish. The English version contains 3,817 texts, 224,402 subtitles, 2,466,812 tokens which is equivalent to 274 hours of video recordings. Its Spanish counterpart shows a similar size containing 3,992 texts, 243,823 subtitles, 2,475,690 tokens and a duration of 283 hours.

Table 1: Summary of EMPAC by language

Language	Texts	Subtitles	Tokens	Duration
English	3,817	224,402	2,466,812	274.03
Spanish	3,922	243,823	2,475,690	283.25

A total of 3778 texts are available and aligned in both languages. There are 38 texts only in English without a translation into Spanish, and 145 texts in Spanish not translated into English.

Table 2: Summary of videos with subtitles aligned.

Category	Videos
Aligned	3,778
Only English	38
Only Spanish	145

5.1.2. Year

A gradual diminution of the volume of texts along the years can be observed for both languages. 2010 is the year featuring the highest volume of texts published (approx. 750), and 2017 is the year featuring the lowest volume of texts (159). During the 2010–2017 period, the volume of videos published has been reduced to a 25%.

5.1.3. Type

The distribution of texts across types is almost identical for both languages. In general, the type featuring more videos

[5] `https://bitbucket.org/empac/corpus/src/master`

Table 3: Summary of EMPAC Spanish

Year	Texts	Subtitles	Tokens	Duration
2009	464	41,533	405,237	50.92
2010	757	50,273	502,100	59.33
2011	566	34,451	348,945	39.99
2012	594	38,390	390,409	43.35
2013	483	35,288	360,968	38.93
2014	347	19,061	200,712	21.87
2015	358	13,918	149,678	16.49
2016	194	6,735	73,374	7.91
2017	159	4,174	44,267	4.46

Table 4: Summary of EMPAC English.

Year	Texts	Subtitles	Tokens	Duration
2009	355	23,115	249,958	40.14
2010	744	48,506	522,283	57.31
2011	563	34,054	372,791	39.90
2012	591	37,362	416,704	43.26
2013	484	34,704	385,908	38.97
2014	357	21,562	244,808	25.12
2015	361	13,939	156,020	16.57
2016	203	7,019	77,521	8.32
2017	159	4,141	40,819	4.45

is by far *news* with a share over the 50% of the corpus. However, if the size of the texts (be it in subtitles or tokens) or their duration is considered, then the type *background* is the one representing a higher proportion of the corpus with a 25% of the total, followed by *news*, while each of the remaining registers covers between 1% to 10%.

Table 5: Summary of EMPAC English by type.

Type	Texts	Subtitles	Tokens	Duration
Background	689	99,938	1,102,105	127.10
News	2,246	74,234	813,665	87.62
Interview	393	30,260	337,252	35.96
Discovery	378	15,605	166,968	17.71
History	111	4,365	46,822	5.64

It is remarkable, that not all types show a similar length. Background videos are the longest ones and tend to be four times longer than the news which is the shorter type of text. The types sorted in descending order by their length in tokens are background, followed by interview, discovery, history and, finally, news.

Table 6: Summary of EMPAC Spanish by type.

Type	Texts	Subtitles	Tokens	Duration
Background	695	109,565	1,098,877	127.72
News	2,320	79,819	828,398	92.03
Interview	413	33,667	339,759	39.82
Discovery	381	16,161	162,019	17.88
History	113	4,611	46,637	5.79

5.1.4. Category

The classification of videos by categories starts in year 2016. All videos published in previous years are labelled as *Other*. If only the videos for the period 2016–2017 are considered, the most frequent category is EU affairs, followed by society, and at a greater distance economy, security and world.

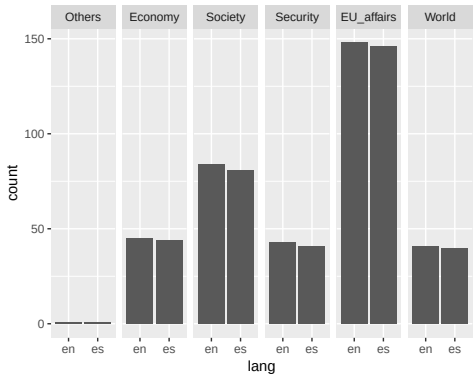

Figure 1: Number of videos by category across languages for 2016 and 2017.

5.2. License

The EMPAC corpus is licensed under an Open Data Commons Attribution License (ODC-BY) v1.0[6]. This license was chosen for two reasons: 1) it applies only to *sui generis* database rights and any copyright in the database structure, while it does not apply to the individual contents of the database; and, 2) it is very permissive, as long as the user observes the terms of the license.

The first reason is of paramount importance, as it allows to apply the license only to the *sui generis* database rights and any copyright in the database structure acknowledging the authorship of the creators, while not applying to the SRT files and their text (the individual contents of the database) which are under copyright as specified in the legal disclaimer of the European Parliament website. By using this license the creators do not attribute to themselves any right on the original subtitles, avoiding any breach of the law, and, thus, reducing all uncertainty for potential users, encouraging maximal reuse and sharing of information.

The ODC-By 1.0 license basically allows the users: a) to share, to copy, distribute and use the database; b) to create, to produce works from the database; and c) to adapt, to modify, transform and build upon the database. As long as the user attributes any public use of the database, or works produced from the database, in the manner specified in the license.

5.2.1. Results

Now, we summarise the key findings of the analysis based on EMPAC corpus which are reported in the PhD dissertation of Serrat Roozen (2019).

[6] https://opendatacommons.org/licenses/by/1.0/

First, we wish to highlight the efforts made by the EU to ensure that, in linguistic terms, the content is accessible to the greatest number of people. However, this study aims to draw attention to the changes that have taken place since 2016. Although subtitles continue to appear in all of the official EU languages, they are no longer necessarily accessible, as the required reading speed is so high that it makes them difficult to read and understand.

In terms of reading speed, the results we obtained show that between 2009 and 2016, the characters per second (cps) rates were distributed similarly in English and Spanish, and we observed a progressive shift towards faster subtitles in both versions. However, in 2016, the Spanish subtitles were slightly faster than the English ones, while the gap grew considerably in 2017. In fact, this same year, 70% of the subtitles in Spanish were displayed at speeds of above 15 cps, with 33% displayed at 21 cps - rates which require reading speeds from the viewer that are almost unattainable. In the case of English, on the other hand, we found that only 59% of the subtitles were displayed at speeds of above 15 cps.

Regarding the number of characters per line (cpl), the trend shows that this variable increased sharply in English in 2017, when it shifted from the 37-cpl standard; however, the rise was even greater in Spanish, so much so that almost 50% of the subtitles surpassed 43 cpl. Hence, we find it interesting that it should be the English version that features the most subtitles in the 42- and 43-cpl group. Moreover, it does not seem far-fetched to hypothesise that this finding is due to an automated transcription process, followed by limited subsequent editing. Naturally, this interpretation of the data requires further confirmation.

As for the number of lines per subtitle, between 2009 and 2016, the trend was practically the same for both English and Spanish, with over 70% of the subtitles consisting of two lines. However, since 2016, the trend has been completely reversed, so that in 2017, up to 80% of the subtitles were of one line.

Lastly, using the CQPweb web interface, we performed an automatic analysis of subtitle segmentation. This constitutes one of the true innovations of this study, since, to our knowledge, it is the first of its kind. In general and in quantitative terms, the results show the low incidence of inadequate segmentation in the EMPAC corpus both in English and Spanish. However, it is necessary to point out that 2017 is the year in which we find the most cases of inadequate segmentation.

Therefore, we can say that the videos hosted on the European Parliament website are not accessible to people with sensory disabilities or to those for whom the language barrier is an impediment to accessing information. The speed at which subtitles are presented makes them largely inaccessible to everyone.

6. Conclusion and Future Research

In this paper, we have presented a new corpus of institutional subtitles in English and Spanish. We have described the methodology to compile the corpus. We have described the corpus, its structure, and its size. And we have provided an overview of the most salient findings revealed by this resource with regard to the accessibility to audiovisual online content produced by the European Parliament through the usage of subtitles. Now, we would like to conclude pointing out two lines of future work.

First, an evaluation on the quality of the automatic linguistic annotation of the corpus was not carried out because it was out of the scope of this research. However, we should assume a sub-optimal quality because the subtitles do not belong to the domain of the texts used to train the English model—the Penn Treebank corpus[7] is made up journalistic texts mainly—and the Spanish model—the Ancora corpus[8] consists of journalistic texts and the CRATER corpus[9] is a collection of telecommunications manuals—neither of them containing subtitles or transcriptions of audiovisual texts or dealing with matters related to the European Parliament. As for the sentence tokenizer (Kiss and Strunk, 2006)—which relies, on the one hand, on an unsupervised model that learns abbreviation words, collocations, and words that start sentences, and, on the other hand, punctuation—we also expect a poor performance as subtitles do not follow exactly the same punctuation conventions as regular written texts. Therefore, an assessment of the quality of the automatic linguistic annotation and, eventually, the training of a language model on a subtitle corpus and its evaluation could be a valuable line for future research.

Second, most subtitle editors already provide feedback on formal features of subtitles when the maximum length of lines in characters, or the maximum reading speed in characters per second is exceeded. However, no feedback on intra-subtitle segmentation attending to linguistic criteria is provided. This functionality could be provided using the insights gained with EMPAC on inappropriate intra- and inter-subtitle segmentation. The corpus or the rules derived from the queries could be used to improve the segmentation of subtitles into lines for automatic captioning or machine translation of subtitles. The implementation of this application would be another line for future research.

We wish to conclude by highlighting that the EMPAC corpus is freely accessible online at `http://hdl.handle.net/21.11119/0000-0006-553B-9`.

Our aim is to contribute to improving accessibility to audiovisual content on different media, and we hope that this corpus can provide the basis for future quantitative and qualitative research.

7. Acknowledgements

We want to thank the CLARIN-D Zentrum of the Universität des Saarlandes for their support archiving the corpus in their repository.

8. Bibliographical References

Barbaresi, A. (2014). Language-classified Open Subtitles (LACLOS): Download , extraction , and quality assessment. Technical report, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 11.

[7] `https://catalog.ldc.upenn.edu/LDC99T42`

[8] `http://clic.ub.edu/corpus/en/ancora`

[9] `https://eprints.lancs.ac.uk/id/eprint/130292`

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Bywood, L., Volk, M., Fishel, M., and Georgakopoulou, P. (2013). Parallel subtitle corpora and their applications in machine translation and translatology. *Perspectives*, 21(4):595–610.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT3: Web inventory of transcribed and translated talks. In Mauro Cettolo, et al., editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation, EAMT 2012*, number May, pages 261–268, Trento, 5.

Di Gangi, M. A., Cattoni, R., Bentivogli, L., Negri, M., and Turchi, M. (2019). MuST-C: a Multilingual Speech Translation Corpus. In Jill Burstein, et al., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pages 2012–2017, Minneapolis, 6. Association for Computational Linguistics.

Díaz Cintas, J. and Remael, A. (2007). *Audiovisual Translation: Subtitling*. Translation Practices Explained. St. Jerome, Manchester.

Evert, S. and Hardie, A. (2011). Twenty-first century Corpus Workbench : Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*, Birmingham, 7. University of Birmingham.

Fishel, M., Georgakopoulou, Y., Penkale, S., Petukhova, V., Rojc, M., Volk, M., and Way, A. (2012). From subtitles to parallel corpora. In Mauro Cettolo, et al., editors, *Proceedings of the 16th Annual Conference of the European Association for Machine Translation, EAMT 2012*, number May, pages 3–6, Trento, 5.

Karakanta, A., Vela, M., and Teich, E. (2018). Preserving and Extending Metadata in Parliamentary Debates. In Darja Fišer, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, 5. European Language Resources Association (ELRA).

Karakanta, A., Negri, M., and Turchi, M. (2019). Are subtitling corpora really subtitle-like? In Raffaella Bernardi, et al., editors, *CEUR Workshop Proceedings*, volume 2481, Bari, 11. CEUR Workshop Proceedings.

Kiss, T. and Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Computational linguistics*, 32(4):485–525.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 923–929, Portorož, 5. European Language Resources Association (ELRA).

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1742–1748, Portorož, 5. European Language Resources Association (ELRA).

Müller, M. and Volk, M. (2013). Statistical Machine Translation of Subtitles: From OpenSubtitles to TED. In Iryna Gurevych, et al., editors, *Language Processing and Knowledge in the Web*, volume 8105 of *Lecture Notes in Computer Science*, pages 132–138. Springer, Berlin/Heidelberg.

Orrego, D. (2013). Avance de la traducción audiovisual: Desde los inicios hasta la era digital. *Mutatis Mutandis*, 6(2):297–320.

Pedersen, J. (2017). The FAR model: assessing quality in interlingual subtitling. *The Journal of Specialised Translation*, (28):210–229.

Pryzant, R., Chung, Y., Jurafsky, D., and Britz, D. (2018). JESC: Japanese-English Subtitle Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1133–1137, Miyazaki, 10. European Language Resources Association (ELRA).

Romero-Fresco, P. (2019). *Accessible Filmmaking: Integrating Translation and Accessibility into the Filmmaking Process*. Routledge, London; New York, 5.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging With an Application To {G}erman. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, Dublin, Ireland.

Serrat Roozen, I. (2019). *Análisis descriptivo de la accesibilidad a los contenidos audiovisuales de webs del Parlamento Europeo*. Ph.D. thesis, Universitat Jaume I, Castelló de la Plana, 7.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., and Gilbro, S. (2014). An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Tiedemann, J. (2008). Synchronizing translated movie subtitles. In Nicoletta Calzolari, et al., editors, *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 1902–1906, Marrakech, 5. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, pages 2214–2218, Istanbul, 5. European Language Resources Association (ELRA).

Tiedemann, J. (2016). Finding alternative translations in a large corpus of movie subtitles. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3518–3522, Portorož, 5. European Language Resources Association (ELRA).

Volk, M., Graën, J., and Callegaro, E. (2014). Innovations in parallel corpus search tools. In Nicoletta Calzolari, et al., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, pages 3172–3178, Reykjavik, 5. European Lan-

guage Resources Association (ELRA).

Weibel, M. (2014). Bilingwis: ein statistikbasiertes Konkordanzsystem für die Systematische Rechtssammlung des Bundes. *LeGes: Gesetzgebung {&} Evaluation*, 25(2):285–291.

Xiao, H. and Wang, X. (2009). Constructing Parallel Corpus from Movie Subtitles. In Wenjie Li et al., editors, *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, pages 329–336, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zeroual, I. and Lakhouaja, A. (2018). MulTed: A multilingual aligned and tagged parallel corpus. *Applied Computing and Informatics*.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In Nicoletta Calzolari, et al., editors, *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3530–3534, Portorož, 5. European Language Resources Association (ELRA).