EXPANDING THE EUROPEAN PARLIAMENT TRANSLATION AND INTERPRETING CORPUS: A MODULAR PIPELINE FOR THE CONSTRUCTION OF COMPLEX CORPORA

Alice FEDOTOVA,[1] Adriano FERRARESI,[1] Maja MILIČEVIĆ PETROVIĆ,[1] Alberto BARRÓN-CEDEÑO[1]

[1]University of Bologna

Keywords: translation and interpreting corpora, corpus construction, parliamentary corpora, natural language processing

1 INTRODUCTION

The present paper introduces an expanded version of the European Parliament Translation and Interpreting Corpus (EPTIC), a multimodal parallel corpus comprising speeches delivered at the European Parliament along with their official interpretations and translations (see Bernardini et al., 2016; Bernardini et al., 2018). Constructing multimodal and parallel corpora for translation and interpreting studies (TIS) has been acknowledged as a "formidable task" (Bernardini et al., 2018), which – if automated, as we propose – involves a number of subtasks such as automatic speech recognition (ASR), multilingual sentence alignment, and forced alignment, each of which poses its own challenges. Yet tackling these subtasks also offers a unique way to evaluate state-of-the-art natural language processing (NLP) tools against a unique, multilingual benchmark. In this paper we discuss the development of a modular pipeline adaptable for each of these subtasks and address the broader implications of this work for the field of corpus construction.

While multilingual sentence alignment is particularly relevant for translation and interpreting corpora, transcriptions of spontaneous or planned speech aligned with recordings are also essential for linguistic research more broadly, in particular for spoken corpora used in phonology, conversational analysis, dialectology and so forth (Lemmenmeier-Batinić, 2023). However, the adoption of NLP tools for corpus construction has often lagged behind due to

technological hesitancy and tool requirements, i.e. lack of re-use of tools developed for specific corpus construction needs, which impeded interoperability.

To address these challenges, this paper aims to raise awareness of the software available for the purposes of corpus construction, create reusable resources, and facilitate the adoption of NLP tools for researchers interested in speech transcription, sentence alignment, and more generally multimodal corpora. We highlight the potential benefits of automatic alignment and transcription for different types of corpora, elaborating on the increasing interest in tools such as OpenAI's Whisper (Radford, 2022) and their suitability for linguistic research. We find that satisfactory results can be achieved with ASR, although challenges remain, especially with regards to the verbatimness of the transcription, where by verbatimness we mean the level of detail where all words are transcribed, along with disfluencies and some extra-linguistic information (Wollin-Giering, 2023). Sentence alignment can be facilitated through state-of-the-art embedding-based tools, whereas forced alignment can be considered a largely solved problem. This makes the construction of EPTIC more streamlined and requiring less human intervention, with wider implications for multilingual corpus construction in the field of TIS and beyond.

2 EPTIC 1.0 AND TOOLS FOR ITS EXPANSION

Within EPTIC, the corpus construction process revolves around individual speech events, where edited "verbatim" reports published by the European Parliament and transcriptions of the speeches are accompanied by transcriptions of interpretations and official translations. Figure 1 uses parallel boxes to represent, both vertically and horizontally, different facets of the same events (source or target, written or spoken). Empty space in the English subcorpus represents the potential for more languages to be added. Corpora containing translations in both directions (e.g., from English to French and from French to English) are referred to as bidirectional, while those with translations in only one direction are referred to as unidirectional.

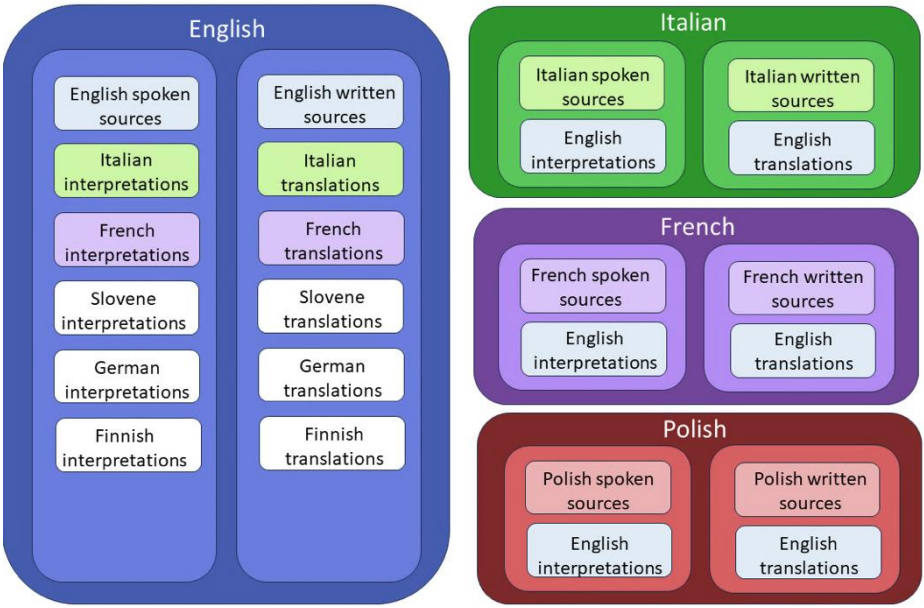

Figure 1: Structure of EPTIC.

The languages included and the size of the previously published EPTIC corpus, before the expansion discussed in the present work, are shown in Table 1.

Our approach to corpus expansion began with a review of previous guidelines for developing the EPTIC corpus (Bernardini et al., 2018; Kajzer-Wietrzny and Ferraresi, unpublished). The first step in the construction of the previous version involved obtaining transcripts, verbatim reports, translations, and interpretations of European Parliament speeches from the official website.[1] Transcripts of the original speeches and interpretations were manually adapted following editing conventions to annotate features of orality such as disfluencies and timestamped using Aegisub. Then, the texts were automatically segmented into sentences and aligned across languages using the Intertext Editor alignment tool.

The creation of the new workflow started with the previous procedure as a basis. It was first subdivided into separate tasks, the main ones being

[1] https://www.europarl.europa.eu/plenary/en/debates-video.html

Table 1: Token counts, by language, of the previously published version of EPTIC.

	Sources		**Targets**	
	Spoken	*Written*	*Interpretations*	*Translations*
English	24,136	22,782	53,615	58,561
French	27,713	26,674	23,185	25,855
Italian	20,016	19,591	20,352	23,234
Polish	11,011	10,616	*TBA*	*TBA*

automatic speech recognition, multilingual sentence alignment, and forced alignment. Software selection was based on criteria such as ease of use and installation, compatibility with the Python programming language, linguistic coverage, and compatibility with Sketch Engine, an established corpus query tool for teaching and research (Rychlý, 2007; Kilgarriff, 2014). Python version 3.11.5 was used along with the Poetry[2] package manager for portability. The resulting workflow is planned for release by the end of 2024 as an open-source GitHub repository and a Python package installable via pip.[3] The following paragraphs discuss the tasks and the considerations made in order to design a new pipeline for EPTIC.

Automatic Speech Recognition has seen recent advancements, with the introduction of Whisper (Radford, 2022) and Wav2Vec 2.0 (Baevski, 2020). However, achieving a reasonable level of transcription quality is complex and context-dependent, as it can be interpreted and evaluated differently depending on the specific domain, task, and application (Kuhn et al., 2024). For EPTIC, we require an ASR system to produce a verbatim transcription where all words are transcribed, along with disfluencies and some extra-linguistic information. “Verbatimness” is, however, also a broad concept (Wollin-Giering, 2023), given the variety of transcription conventions existing in the field of linguistics, and Whisper has been observed to produce transcripts “often almost comparable to the final read through of a manual (verbatim to gisted) transcript” (Wollin-Giering, 2023). We further explore this

[2] https://python-poetry.org/
[3] https://pypi.org/project/pip/

claim by testing a variant of Whisper, WhisperX,[4] on our data. Given its exceptional performance in long-form transcription (Bain et al., 2023), we hypothesize that WhisperX could be especially beneficial when dealing with parliamentary speeches. In this regard, EPTIC serves as a challenging and unique benchmark due to its multilingual nature, inclusion of non-native speech, and interpreted text, which has been found to<z be particularly difficult to transcribe (Wang and Wang, 2024).

Sentence Alignment involves identifying and aligning parallel sentences, both monolingually and multilingually. For this task, we used Bertalign (Liu & Zhu, 2022), a tool for aligning parallel corpora based on sentence embeddings. Unlike predecessors like Hunalign[5] that rely on lexical translation probabilities, Bertalign employs sentence embeddings to identify parallel sentences, providing a more robust approach for handling semantic similarities across languages.[6] Some changes were necessary as the default settings were not appropriate out-of-the-box. For instance, we were required to change the value of the variable is_split to False, as the corpus was already sentence-split in a previous step. The tool produces alignments in the format of a list of tuples, and it has been extended with an additional Python script to convert its output into the Sketch Engine alignment format based on corpus-internal indexing.

3 EPTIC 2.0: PIPELINE AND CORPUS PROPERTIES

The Python pipeline, aimed at facilitating the expansion of the EPTIC corpus, has been structured in a modular fashion. This process begins with the extraction of text and video data, either manually or through the use of ad-hoc scripts, depending on the amount of data that the researcher intends to add.[7]

[4] https://github.com/m-bain/whisperX

[5] https://github.com/danielvarga/hunalign

[6] The settings for Bertalign are not documented, but information about the code is available as part of comments in its aligner.py script, which is provided in a GitHub repository.

[7] Download of the video through the European Parliament's interface can be hindered by the lengthy process, requiring a personal e-mail address which is then used to obtain a URL for the purpose.

Table 2: Token counts, by language, of the expanded version of EPTIC.

	Sources		**Targets**	
	Spoken	*Written*	*Interpretations*	*Translations*
English	43,138	41,047	55,109	58,651
French	35,648	34,063	31,935	35,566
Italian	21,208	20,646	27,329	31,816
Polish	9,458	9,193	*TBA*	*TBA*
Slovene	*TBA*	*TBA*	19,717	22,476
German	*TBA*	*TBA*	18,258	19,822
Finnish	*TBA*	*TBA*	11,624	12,045

Transcription is then performed using WhisperX, which concurrently provides timestamp information. To remove mistranscriptions and to ensure adherence to the transcription guidelines, the transcripts undergo manual review to incorporate disfluencies and rectify potential mistranscriptions. Ongoing research is being conducted to determine whether official transcripts can be leveraged to enhance the quality of the transcriptions in terms of word recognition, and whether it is possible to include disfluency markers such as hesitations, false starts, and repetitions.

Once all required texts have been transcribed, they undergo sentence splitting and sentence alignment using Bertalign. Subsequently, relevant metadata, encompassing session topics, are automatically retrieved from the European Parliament website. The only metadata item requiring manual input is the speech type, which can be defined as impromptu, read out, or mixed if both delivery types are present. After exporting the alignments in the Intertext format[8] and performing Part-of-Speech tagging with Sketch Engine, the texts and metadata are converted to the .vert format, rendering them ready for indexing in Sketch Engine (Kilgarriff, 2014; Rychlý, 2007). The following part-of-speech taggers were used to annotate the texts within Sketch Engine: MULTEXT-East Slovenian (version 4)[9] for Slovene, Polish NKJP[10] for Polish,

[8] https://wanthalf.saga.cz/intertext

[9] http://nl.ijs.si/ME/V6/msd/html/msd-sl.html

[10] https://nkjp.pl/poliqarp/help/ense2.html

Table 3: Performance of WhisperX by language, expressed in WER.

	English WER	Italian WER	French WER	Slovenian WER
Whisper small	0.212	0.219	0.162	0.463
Whisper medium	*0.196*	*0.173*	0.213	0.327
Whisper large-v2	*0.194*	*0.152*	0.118	0.262

FreeLing[11] for Italian and French, TreeTagger[12] for Finnish and English, and German RFTagger[13] for German.

We now highlight the substantial expansion of the EPTIC corpus across different languages and all subcorpus types, as compared to the original corpus presented in Table 1. Table 2 shows the size, in tokens, of the updated bidirectional English, French, Italian subcorpora, with the English subcorpus exhibiting the largest increases. The Polish unidirectional subcorpus has not yet received an expansion, though text-to-video alignments were added as an additional feature. Furthermore, new additions to EPTIC include the unidirectional Slovenian, German, and Finnish subcorpora, with the Slovenian interpretations (19,717 words) and translations (22,476 words) representing the largest target subcorpus to be added as part of this update.

WhisperX has demonstrated robust performance across diverse speech types, languages, and accents. Consistent with prior research findings (Wollin-Giering, 2023), it does not provide verbatim transcriptions, though the word error rate (WER) metric alone may be insufficient to fully ascertain the extent of this limitation.[14] We are currently investigating additional evaluation metrics and conducting qualitative analyses to comprehensively assess WhisperX's transcription capabilities. The results in Table 3 indicate that good performance can be achieved for certain languages, such as Italian, which exhibits a low WER of 0.118, while performance degrades for Slovenian.

[11] https://freeling-user-manual.readthedocs.io/en/latest/tagsets/

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[13] https://www.cis.uni-muenchen.de/~schmid/tools/RFTagger/

[14] WER evaluates the accuracy of transcribed text compared to the ground truth, with lower WER indicating better performance.

Table 4: Disfluency types and transcription errors made by Whisper large-v2.

	Human transcription	Whisper large-v2
Contraction	**I'm** encouraged that the interim leadership has promised substantial reforms because embarking on such a path will greatly strengthen Tunisia's relationship with the European Union.	**I am** encouraged that the interim leadership promised substantial reforms because embarking on such a path will greatly strengthen Tunisia's relationship with the European Union.
Truncation	E.g., polygamy was banned, veils were banned, foreign direct **in-** ehm **investment** was encouraged, tourism was encouraged.	E.g. polygamy was banned, veils were banned, foreign direct **investment** was encouraged, tourism was encouraged.
Discourse marker	Presidente, la sommossa in Tunisia è senz'altro un riflesso della frustrazione della gente di fronte alla situazione politica, poi c'è anche la corruzione la la con- e la conduzione della famiglia regnante **diciamo**.	La sommossa in Tunisia è senz'altro un riflesso della frustrazione della gente di fronte alla situazione politica, poi c'è anche la corruzione e la conduzione della famiglia regnante.
Filled pause	**Ehm** importanti sono le riforme… solo questo potrà rinforzare le relazione con la Tunisia.	Importanti sono le riforme solo questo potrà rinforzare le relazioni con la Tunisia.
Empty pause	Il Parlamento europeo deve condannare queste azioni che **…** rivelano il volto opprimente e aggressivo della Turchia a Cipro.	Il Parlamento europeo deve condannare queste azioni che rivelano il volto frimente e aggressivo della Turchia a Cipro.

Furthermore, we evaluated WhisperX on a subset of English speech data to examine whether factors such as speaker nativeness or interpreted speech influence WER. Our findings indicate a WER of 0.104 for native English speakers, 0.110 for non-native speakers, and a notably higher WER of 0.222 for interpreted speech. This is consistent with Wang and Wang (2024), where the higher WER in interpreted speech is attributed to the increased presence

of disfluencies such as filled and unfilled pauses and the challenges posed by mispronunciations, which are more prevalent in interpreted speech, leading to greater difficulties in accurate ASR transcription.

The task of aligning transcriptions with their corresponding audio files across multiple languages and subcorpora within EPTIC has formerly been carried out manually, a laborious and error-prone process, especially at a large scale. Leveraging Bertalign appears to be effective in tackling this complex challenge. This allowed for the introduction of automatic alignment across all subcorpora within EPTIC - a significant improvement over the previous version of the corpus, where text-to-text alignments were often absent due to the manual efforts required.

4 CONCLUSIONS AND FUTURE WORK

In conclusion, the development of a state-of-the-art, NLP-based modular pipeline has resulted in a significant expansion of the European Parliament Translation and Interpreting Corpus. The resulting corpus features a significant increase in size and language coverage compared to the previous version of EPTIC (Bernardini et al., 2016; Bernardini et al., 2018), making it a more comprehensive and valuable resource for research in TIS, as well as related fields such as linguistics and NLP. The development of this pipeline has demonstrated the potential for automating various aspects of corpus construction by including ASR, multilingual sentence alignment, and forced alignment. By integrating tools like WhisperX, Bertalign, and aeneas, the process of transcribing, aligning, and timestamping the audio-video data has been streamlined, reducing the time and effort required for manual intervention.

Limitations remain, such as the relatively small overall corpus size and the challenges in evaluating punctuation prediction by WhisperX, i.e. how closely it aligns with the sentence boundaries that a human would conceive of as natural and appropriate for a given spoken content. Additionally, ongoing work includes conducting a more thorough evaluation of the tools' performance using metrics including the F1 score in the case of sentence alignment, as well

as experimenting with methods aimed at improving ASR performance, for instance by fine-tuning Whisper on our transcriptions. The inclusion of automatic alignment between all subcorpora in EPTIC, facilitated by Bertalign's robust performance, represents a significant advancement, enabling analyses and comparisons that were previously impractical or impossible due to the substantial manual effort required.

Looking ahead, future work could involve adding numerical features to represent prosodic aspects of speech, which could enable more research avenues leveraging the spoken EPTIC data. These numerical features, such as mean pitch, pitch range, intensity contours, duration of speech segments, voice quality measures like jitter and shimmer, and spectral characteristics, could be obtained using acoustic analysis tools like Praat (Boersma & Weenink, 2024). Additionally, the release of the Python-based software will provide a valuable resource for other researchers and corpus builders working with the Sketch Engine platform or similar tools. By lowering the barrier to entry and increasing the efficiency of multimodal corpus development, this work paves the way for more comprehensive and representative language resources, ultimately driving progress in various domains of linguistic and computational research.

5 ACKNOWLEDGMENTS

The work of Alice Fedotova is supported by the NextGeneration EU programme, ALMArie CURIE 2021 - Linea SUpER, Ref. CUPJ45F21001470005.

6 REFERENCES

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460. Retrieved May 19, 2024, from https://10.48550/arXiv.2006.11477

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). Whisperx: Time-accurate speech

transcription of long-form audio. *arXiv preprint.* Retrieved May 20, 2024, from https://arxiv.org/pdf/2303.00747

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building Interpreting and Intermodal Corpora: A how-to for a Formidable Task. *Making Way in Corpus-Based Interpreting Studies*, 21-42. https://doi.org/10.1007/978-981-10-6199-8_2

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC — Exploring Simplification in Interpreting and Translation from an Intermodal Perspective. *Target*, *28*(1), 61-86. https://10.1075/target.28.1.03ber

Boersma, P. & Weenink, D. (2024). Praat: Doing Phonetics by Computer [Computer program]. Version 6.4.12. Retrieved May 14, 2024, from http://www.praat.org/

Della Corte, G. (2020). Text and Speech Alignment Methods for Speech Translation Corpora Creation: Augmenting English LibriVox Recordings with Italian Textual Translations [Master's thesis]. Retrieved May 14, 2024, from http://www.diva-portal.org/smash/get/diva2:1440026/FULLTEXT01.pdf

Jones, C., Li, W., Almeida, A., & German, A. (2019). Evaluating Cross-Linguistic Forced Alignment of Conversational Data in North Australian Kriol, an Under-Resourced Language. *Language Documentation and Conservation*, *13*, 281-299. Retrieved May 15, 2024, from http://hdl.handle.net/10125/24869

Kajzer-Wietrzny, M. & Ferraresi, A. (2020). *Guidelines for EPTIC collaborators.* Unpublished manuscript.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten Years on. *Lexicography, 1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the Accuracy of Automatic Speech Recognition Solutions. *ACM Transactions on*

Accessible Computing, 16(4), 1-23. https://doi.org/10.1145/3636513

Lei, L. & Zhu M. (2022). Bertalign: Improved Word Embedding-Based Sentence Alignment for Chinese–English Parallel Corpora of Literary Texts. *Digital Scholarship in the Humanities, 38*(2), 621-634. https://doi.org/10.1093/llc/fqac089

Lemmenmeier, D. (2023). Spoken Language Corpora: Approaches for Facilitating Linguistic Research [Doctoral dissertation]. Retrieved May 13, 2024, from https://www.zora.uzh.ch/id/eprint/235310/1/Lemmenmeier_Dolores_Dissertation.pdf

Pettarin, A. (2018). Forced-alignment-tools [Computer program]. Version 1.0.9. Retrieved May 16, 2024, from https://github.com/pettarin/forced-alignment-tools

Radford, A., Kim, J.W., Xu, T., Brockman, G., Mcleavey, C. & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518). Proceedings of Machine Learning Research. Retrieved May 13, 2024, from https://proceedings.mlr.press/v202/radford23a.html

Rychlý, P. (2007). Manatee/Bonito-A Modular Corpus Manager. *RASLAN*, 65-70. Retrieved May 14, 2024, from https://www.sketchengine.eu/wp-content/uploads/Manatee-Bonito_2007.pdf

Wang, X., & Wang, B. (2024). Exploring Automatic Methods for the Construction of Multimodal Interpreting Corpora. How to Transcribe Linguistic Information and Identify Paralinguistic Properties? *Across Languages and Cultures.* Retrieved 14 May, 2024, from https://eprints.whiterose.ac.uk/212127/3/2024%20Across%20-%20Accepted%20version_Binhua%20Wang.pdf

Wollin-Giering, S., Hoffmann, M., Höfting, J., & Ventzke, C. (2023). Automatic Transcription of Qualitative Interviews. *Sociology of Science Discussion Papers*. https://10.13140/RG.2.2.14480.38404

Wu, H., Yun, J., Li, X., Huang, H., & Liu, C. (2023). Using a Forced Aligner for Prosody Research. Humanities and Social Sciences Communications, 10(1), 1-13. https://doi.org/10.1057/s41599-023-01931-4

RAZŠIRITEV KORPUSA PREVODOV IN TOLMAČENJ EVROPSKEGA PARLAMENTA (*EUROPEAN PARLIAMENT TRANSLATION AND INTERPRETING CORPUS*): MODULARNI PROCES ZA GRADNJO KOMPLEKSNIH KORPUSOV

Članek opisuje razširjeno različico večmodalnega vzporednega korpusa EPTIC, ki vsebuje govore iz Evropskega parlamenta, njihova tolmačenja in prevode. Predstavlja modularen proces za avtomatsko prepoznavanje govora, večjezično poravnavo povedi in časovno žigosanje z uporabo orodij kot so WhisperX, Bertalign in aeneas. Čeprav ostajajo nekateri izzivi, kot so majhna velikost korpusa za manj bogata jezikovna okolja, rezultati kažejo, da je mogoče z naprednimi orodji za obdelavo naravnega jezika doseči zadovoljive rezultate pri gradnji korpusa. Razširjena različica EPTIC zajema večji obseg podatkov in jezikov, kar jo naredi bolj celovit vir za raziskave. Prihodnje delo vključuje ovrednotenje uporabljenih orodij in razvoj metod za samodejno izboljšanje jezikovnih prepisov.

Keywords: korpusi prevodov in tolmačenj, gradnja korpusov, parlamentarni korpusi, obdelava naravnega jezika