EPTIC — Past, Present, Future

2023-24 Project Report

Alice Fedotova

January 23, 2025

Introduction

This document details my activities as a Research Fellow at the University of Bologna from October 1, 2023, to October 1, 2024. It includes a description of the workflow developed during the project, instructions for future EPTIC updates, suggestions for future work, and additional resources that could be useful for further refinements. Additionally, it presents the research output derived from the work conducted during my fellowship. This research position was supported by the NextGeneration EU programme, ALMArie CURIE 2021 - Linea SUpER, Ref. CUPJ45F21001470005.

Project Objectives

Considering these challenges, the following activities are necessary to improve and expand the EPTIC corpus. Advances in Artificial Intelligence, such as Speech-to-Text systems and automatic alignment tools, provide an opportunity to increase both the size and language coverage of the corpus.

1. Curation of the existing CoLiTec corpora in their NoSketch Engine versions (e.g., compilation of missing metadata, checks on NoSketch Engine links, etc.).
2. Annotation and indexing of EPTIC data currently stored on the MySQL-based tool SkEPTIC.
3. Testing of neural Text-to-Speech systems to transcribe European Parliament spoken data, with a view of enlarging existing EPTIC subcorpora and extending its language coverage.
4. Production of guidelines:
   4.a. For the enlargement of EPTIC.
   4.b. For extending EPTIC-specific methods to other multimodal corpus designs.
5. Research activities resulting in the submission of at least one journal article or book chapter and one proposal for a conference presentation.

Notes on Terminology

NoSke, NoSketchEngine: free version of the corpus management tool Sketch Engine `https://nlp.fi.muni.cz/trac/noske`.

SkEPTIC: EPTIC's data annotation platform, i.e. `https://skeptic.dipintra.it/users/login`.

The database, DB, SkEPTIC database: backend of the SkEPTIC platform, i.e. where some of the latest annotated texts were stored at the beginning of the project: `https://skeptic.dipintra.it/users/login`. It consists in a MariaDB database hosted on the DIT's servers.

EPTIC's NoSketch Engine: web-hosted, NoSketch Engine-based public version of the EPTIC corpora, i.e. `https://bellatrix.sslmit.unibo.it/noske/eptic/`. The format to upload the corpora on Sketch Engine's requires: the vertical files, the alignments to be indexed, and the registry files.

Raw data: small portion of unprocessed, unorganized data, provided at the beginning of the project.

NoSketch Engine's data: the EPTIC data already published on NoSketch Engine at the beginning of the project. Was missing some raw data and some recent texts from the database behind SkEPTIC.

SkEPTIC data: the latest annotated texts that were present on SkEPTIC's database at the beginning of the project. They had to be merged with the raw data and EPTIC's NoSketch Engine data.

Data sources: the three abovementioned data sources. They were combined, deduplicated, cleaned, and uploaded as a renewed suite of corpora on EPTIC's NoSketch Engine. Next, the combined data was also uploaded as an unified and updated MariaDB database, replacing the one behind SkEPTIC.

Text-to-video alignment: timestamps indicating the start and the end of each spoken sentence.

Text-to-text alignment: multilingual alignments between a text in a given language and all available texts pertaining to the same event in the other languages.

Contents

1 Extract, Transform and Load

Extract, transform, load (ETL) is a three-phase process where data is extracted from an input source, transformed (including cleaning), and loaded into an output data container [2]. In the context of the project, the data had to be extracted from three sources: 1. the SkEPTIC database, 2. the NoSketch Engine version of EPTIC, and 3. the raw data provided at the beginning by the project collaborators. Subsequently, it had to be combined, deduplicated, cleaned, and prepared for upload into: 1. a new NoSketch Engine set of corpora and 2. a new MariaDB database to replace the one behind SkEPTIC.

1.1 Data Extraction

The database containing the data to be extracted and merged with the raw data and NoSketch Engine's data was the SkEPTIC database, i.e. the MariaDB database behind the SkEPTIC platform, where the latest annotated texts provided by the EPTIC collaborators were stored. These annotations were created after the release of the EPTIC corpus on NoSketch Engine, and therefore had to be extracted in order to be merged with the other two abovementioned data sources. In order to obtain the data from SkEPTIC's database, I used a dump [3] of the original SkEPTIC's database provided by an EPTIC collaborator. I then installed a local instance of MariaDB, and used `pipelines/extract_tables_db.py` to connect to it and extract the tables from the database using SQLAlchemy[1]. A choice had to be made between CSV and Excel. I chose to use Excel because it allowed us to easily analyze and inspect the data throughout the project. Its features for extracting descriptive statistics and performing manual transformations made it a more practical and versatile choice, despite not being the most lightweight option. Furthermore, even after merging the data sources (see Section 1.2), the dataset remained relatively small, allowing for quick processing regardless of the chosen format [1].

1.2 Data Transformation I

During this part of the project, I aggregated all the provided data into a single SOT (Source of Truth)[2], represented as an Excel file for each table expected by SkEPTIC's database. These files, stored in `eptic.v3/database_tables`, contain data from three sources: 1. the SkEPTIC database, 2. the NoSketch Engine version of EPTIC, and 3. the raw data. To achieve this, I first standardized the attributes from the tags in the NoSke files to align with the database table columns. For instance, the NoSketch Engine version of the corpus used a different convention for uniquely identifying the texts (e.g., `0001en_sp_st`), so I assigned new integer-only, SkEPTIC-compliant unique IDs to these entries to align with the database's conventions. Next, I proceeded to merge the datasets. Duplicate entries, present across the merged dataset, were identified and removed using the Levenshtein distance [6] algorithm, which calculates the minimum number of single-character edits required to match strings. Texts with a Levenshtein distance lower than 10 were considered for further inspection, which was

[1] `https://www.sqlalchemy.org/`

[2] `https://en.wikipedia.org/wiki/Single_source_of_truth`

conducted manually. If a text was present both in the NoSkeptic and SkEPTIC data, priority was given to the version added on SkEPTIC. Besides deduplication, this method was also used to identify which older texts from the NoSketch Engine version of the corpus should have been assigned to the "events" added in SkEPTIC, an additional ID introduced at a later stage of EPTIC's development [4].

1.3 Data Transformation II

Once all data sources were merged, some metadata, texts, videos, text-to-video alignments, and text-to-text alignments were identified as missing. This part of the project was therefore dedicated to restoring as many missing components as possible before uploading the dataset. To expedite the task, existing tools were adapted to automate the retrieval of most missing elements, while leaving tasks requiring human intervention to future work. The focus was on achieving a practical level of completeness within a reasonable timeframe rather than striving for absolute perfection; the following paragraphs will discuss the approach taken to address each of these missing components.

The missing **text-to-video alignments** were performed using aeneas[3], a Python/C library designed to automatically synchronize audio and text. A part of the results was manually checked and found to be satisfactory, though some manual adjustments were applied in cases where the speeches began later in the videos, a scenario that aeneas is unable to handle effectively. No automatic evaluation was conducted, as this step will not be necessary in future EPTIC updates. Whisper and Whisper-derived models, which will be employed to streamline the corpus construction process in future updates, natively support timestamping, making this aspect of the process inherently automated.

As for the missing **text-to-text alignments**, these were performed using Bertalign[4], the state-of-the-art model for multilingual sentence alignment at the time of the project [5]. Though an automatic evaluation would have been warranted, the results were deemed considerably reliable after human inspection. Due to inconsistencies in the existing data, we ultimately decided to re-align all texts from scratch. To adapt Bertalign for this project, I configured `pipelines/align_texts_bertalign.py` to output alignments in Intertext format, adjusted Bertalign's parameters to preserve the existing sentence splitting, and implemented custom logic to correctly handle the required language combinations.

As **metadata related to the interpreters** were mostly missing, diarization was added with `pyannote-audio`[5] and voice gender recognition was performed with `wav2vec2-large-xlsr-53-gender-recognition-librispeech`[6]. The scripts I employed are located at `pipelines/pyannote.py` and `pipelines/assign_genders.py` respectively. It should be underscored that this addition was done automatically and with limited audio samples, therefore human verification is warranted for researchers intending to conduct further studies involving these specific variables. Regarding the nativeness of the interpreters, no model capable of determining this feature was found. Therefore, apart from the

[3] `https://github.com/readbeyond/aeneas`

[4] `https://github.com/bfsujason/bertalign`

[5] `https://github.com/pyannote/pyannote-audio`

[6] `https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech`

annotations provided by the EPTIC collaborators, this metadata remained almost entirely incomplete.

Another area where completeness could not be achieved relates to the **videos of speeches** delivered at the European Parliament and their interpretations. An experimental script was used to retrieve some of the missing videos; however, due to restrictions on the website, the script did not work consistently, making it more practical to leave this task to human intervention. At the time of publishing, 73 texts remain without videos, either because the videos on the website were corrupted or could not be retrieved. A full list of these videos can be found in Section 5.9 of the Appendix.

Lastly, three cases of **missing texts** were not addressed in this version of the corpus. The first case involves missing texts pertaining to already added events, such as instances where either the spoken or written version of a language is missing for a given event. The second case concerns events that only contain source texts, where the absent target texts could be considered missing parts. Lastly, the third case comprises subcorpora where the number of texts was considered too scarce, such as Hungarian as a target. A full list of these events can be found in Section 5.9 of the Appendix.

1.4 Uploading the Data on NoSketch Engine

The files required for uploading the latest version of EPTIC are located at `eptic-pipelines/eptic.v3`. More specifically, `eptic.v3/bertalign_alignments` contains the alignments in XML format as required by the script `noske_scripts/intertext2noske.py`, which converts InterText alignment files into alignment files[7] usable by NoSketchEngine. Vertical, POS-tagged files[8] are stored in `eptic.v3/pos_tagged_files`, and lastly, registry files[9] are located at `noske_files/registry`. The procedure adopted to prepare the vertical files for upload is largely similar to the one described in `docs/pipeline_corpus_tagging_indexing.txt`, with the exception that the pre-processing steps previously adopted before POS-tagging were handled by `pipelines/database_to_pretgd.py`, a script designed to prepare the subcorpora for tagging by putting together the tabular data contained in `eptic.v3/database_tables`. These intermediate files are illustrated in `eptic.v3/pre_pos_files` as an example. POS-tagging was then performed on the basis of these files on Sketch Engine's official website, and the individual vertical files were saved for later upload on DIT's instance of NoSketch Engine. The only downside to this approach was the inconsistency in the available taggers on NoSketch Engine: for instance, not all languages could be tagged using the TreeTagger, and some were tagged using RFTagger or FreeLing.

For the most part, the procedure to upload the updated subcorpora on NoSketch Engine did not undergo major modifications. Once all required files had been uploaded in the respective folders, I:

1. Renamed registry files to, e.g. `eptic3_en_sp_st`, to avoid conflicts with the previous version of che corpus, and adapted the paths accordingly.
2. Prepared the script for the NoSke alignment files `noske_scripts/intertext2noske_cambiato.py`,

[7] `https://www.sketchengine.eu/guide/setting-up-parallel-corpora/#tab-id-4`
[8] `https://www.sketchengine.eu/glossary/vertical-file/`
[9] `https://www.sketchengine.eu/documentation/corpus-configuration-file-all-features/`

which takes as arguments the registry files of the corpora being aligned and an alignment file:

```
python intertext2noske_cambiato.py \
    '/var/lib/manatee/registry/eptic3_sl_sp_tt' \
    '/var/lib/manatee/registry/eptic3_fr_sp_tt' \
    'eptic_fr_sp_tt.eptic_sl_sp_tt.xml'
```

3. Applied `noske_scripts/intertext2noske_cambiato.py` to all possible alignment combinations required by the XML files using `noske_scripts/process_alignments_eptic3.sh`.
4. Applied `noske_scripts/fixgaps_and_rename.py` to all output NoSke alignment files, a script which applies the NoSke helper script fixgaps.py[10] to all NoSke alignment files and renames them to the common format referenced in the registry files, e.g. `alignment.de_sp_tt.en_sp_st.txt`.
5. Moved the postprocessed alignment files to the `/manatee/alignedf_files/EPTIC.V3` folder.
6. Moved the videos to the `/video` folder on `amelia.sslmit.unibo.it`.

2 Automatic Speech Recognition Experiments

See Section 5.7 for recent updates about work on verbatim transcription in the field of ASR, which makes this part of the project somewhat obsolete at the time of writing. Nevertheless, the following sections describe the experiments conducted during the project attempting to obtain a model capable of providing a verbatim transcription of the speeches and the interpretations.

2.1 Data Preparation

I accidentally deleted the data I used for fine-tuning, but it can be easily reobtained using the same code that I used.

2.2 Whisper Fine-Tuning

Only managed to fine-tune Whisper-small. The results were interesting and were reported in the paper presented at CLiC-it 2024 (see Section ...).

3 Pipeline and Production of Guidelines

3.1 Publishing the Codebase on GitHub

The codebase is publicly available at `https://github.com/ffedox/eptic-pipelines`. It contains five main folders: "docs", containing the technical documentation for tagging, indexing and uploading EPTIC; "eptic.v3", the new, current version of the corpus; "eptic.v4", demonstrating the workflow designed for future updates; "pipelines", containing the code developed during the project;

[10] `https://www.sketchengine.eu/documentation/mn-mapping-helper-scripts/`

"tests", containing scripts to test for possible bugs that might arise during the corpus construction process. The code can also be found in the directory `/home/afedotova/EPTIC25` on DIT's server at `john.sslmit.unibo.it`.

3.2 Guidelines for the Enlargement of EPTIC

A problem that was identified early on during the project concerned the availability of the videos. As of 2023/2024, older videos (of debates around 2011) could not be obtained from `https://www.europarl.europa.eu/plenary/en/debates-video.html` and had to be downloaded from `https://multimedia.europarl.europa.eu/en/webstreaming` instead. This meant that no automated approach could not be implemented for the retrieval of the videos, as the new website is configured in such a way that the videos can only be obtained via email. Additionally, it is not possible to download all available tracks at once anymore. SkEPTIC has to be updated to allow multiple videos to be uploaded at once, keeping track of whether the video is a speech or an interpretation, and the language spoken in the video.

4a. The suggested workflow for future additions to EPTIC would be (4a, 4b):

TBA

4 Papers, Conferences, Other

4.1 Data Analysis and ICL Abstract

This part of the project was dedicated to examining the data and submitting an abstract to the 21st International Congress of Linguists (ICL)[11]. Following Bernardini et al., 2016, I proposed an abstract analyzing ...

Though it was not presented, the abstract was accepted to be presented at the conference.

4.2 JTDH Extended Abstract (May 2024)

In May 2024, I submitted an extended abstract for the 14th Conference on Language Technologies and Digital Humanities (JT-DH-2024). The abstract was focused on ...

The extended abstract was accepted and presented at the 14th Conference on Language Technologies and Digital Humanities (JT-DH-2024).

4.3 CLiC-it 2024 Paper

Paper for CLiC-it 2024 on EPTIC construction and ASR experiments accepted and presented. The paper discussed ...

[11] `https://icl2024poznan.pl/`

4.4 Other

EPTIC v2-Expanding an intermodal, multidirectional parallel corpus of European Parliament debates. CoLiTec Seminar 01/03/24.

Linear and logistic regression in LDA and NLP. (? Should I mention this here?)

5 Future Work

(For now just a list of things I should remember to mention)

5.1 SkEPTIC Database Update

SkEPTIC's database should be updated.

5.2 Consistent tagging

Possibility of using the Treetagger instead of uploading the files on Sketch Engine for tagging.

5.3 CrisperWhisper Streamlit GUI

There's a GUI already available for transcribing audios with CrisperWhisper. Possibility of hosting it like we did with the other Whisper models?

5.4 Update Python script to not align all texts again next time

Will do as part of the things to finish for the project. In short: adapting the code for future updates (the workflow will be slightly different)

5.5 Adding texts to incomplete events

Mentioned at the end of Section 1.3.

5.6 SkEPTIC has to be updated to allow multiple videos

Maybe. Reasons explained in Section 3.2

5.7 Research into CrisperWhisper

Though it is likely that CrisperWhisper exhibits better performance than the models developed during the project, due to the availability of more data and resources, it is still unclear how it performs compared to a human annotator. More research could be conducted to compare the performance of CrisperWhisper against the gold transcriptions performed by EPTIC's annotators.

5.8 EPTIC vs ParlaMint

What is the ParlaMint format and how does it compare to EPTIC? More research into this

5.9 The pipeline as learning content

Just an idea we had

References

[1] Stack Overflow Contributor. *What are the disadvantages of using an Excel worksheet instead of a CSV with pandas?* Accessed: 2023-01-11. 2025. URL: https://stackoverflow.com/questions/61692226/what-are-the-disadvantages-of-using-an-excel-worksheet-instead-of-a-csv-with-pan.

[2] Wikipedia contributors. *Extract, transform, load.* Accessed: 2024-12-09. 2024. URL: https://en.wikipedia.org/wiki/Extract,_transform,_load.

[3] MariaDB documentation. *mariadb-dump.* Accessed: 2024-12-09. 2024. URL: https://en.wikipedia.org/wiki/Extract,_transform,_load.

[4] Marta Kajzer-Wietrzny and Adriano Ferraresi. *SKEPTIC Guidelines for EPTIC Collaborators.* Available upon request from the authors. EPTIC Project. 2020.

[5] Lei Liu and Min Zhu. "Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts". In: *Digital Scholarship in the Humanities* 38.2 (June 2023). Published online: 29 December 2022, pp. 621–634. DOI: 10.1093/llc/fqac089. URL: https://doi.org/10.1093/llc/fqac089.

[6] Wikipedia contributors. *Levenshtein distance — Wikipedia, The Free Encyclopedia.* [Online; accessed 17-Dec-2024]. 2024. URL: https://en.wikipedia.org/wiki/Levenshtein_distance.

Appendix

Lists of missing texts or videos to handle in the following updates.

Missing texts

Events: 1 (pl_tt_wr missing), 48 (fr_tt_sp missing), 140 (sl_tt_wr missing), 141 (sl_tt_wr missing), 142 (sl_tt_wr missing), 143 (sl_tt_wr missing), 150 (sl_tt_wr missing), 151 (sl_tt_wr missing), 154 (sl_tt_wr missing), 177 (sl_tt_sp missing), 178 (sl_tt_sp missing), 179 (en_st_sp missing), 200 (sl_tt_sp missing), 209 (sl_tt_sp missing), 222 (it_st_sp and en_tt_sp missing).

Only sources

Events: 158, 159, 160, 161, 164, 165, 180, 182, 183, 186, 188, 189, 190, 191, 193, 194, 196, 197, 198, 199, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255.

Missing videos

Texts: 512, 514, 515, 1030, 1019, 776, 265, 267, 269, 782, 271, 784, 273, 275, 277, 279, 535, 281, 536, 283, 538, 1317, 935, 296, 936, 298, 937, 938, 943, 1330, 959, 961, 1985, 837, 2374, 582, 2376, 584, 586, 588, 972, 590, 1998, 592, 594, 602, 604, 477, 606, 479, 485, 486, 488, 489, 491, 619, 493, 621, 2287, 623, 2289, 1017, 1011, 366, 374, 631, 1016, 505, 1018, 635, 1020, 509, 1022.