Building a Spanish–Greek Parallel Speech Dataset

Adriana Rodríguez Flórez
ÚFAL, Charles University in Prague
EM LCT
adrirflorez@gmail.com

Abstract

Speech translation research has predominantly focused on high-resource, English-centric pipelines, resulting in the underrepresentation of many language pairs. In particular, Greek lacks annotated speech corpora and reliable ASR tools, posing challenges for building and evaluating direct speech-to-speech (S2S) translation systems. This paper presents a pipeline for the construction of a small, time-aligned Spanish–Greek parallel speech corpus using parliamentary recordings from the VoxPopuli dataset. Audio is transcribed using Whisper, a timestamp-aware ASR system provided by OpenAI. Then, the resulting transcripts are aligned at the sentence level using HunAlign. Despite significant development and processing challenges — particularly for processing of Greek utterances — the results demonstrate the feasibility of constructing bilingual speech corpora for low-resource language pairs. This work serves as a proof of concept for this, emphasizing the need for more inclusive resources in multilingual speech translation.

1 Introduction

Building on the momentum of large-scale multilingual resources, recent S2S systems have achieved remarkable progress — but mostly for higher-resource languages, particularly those that align well with English-centric pipelines. As a result, language pairs such as Spanish-Greek remain underrepresented, especially in the domain of spoken language resources.

Although Spanish benefits from extensive support in automatic speech recognition (ASR) and speech corpora, Greek remains a low- to mid-resource language with few robust ASR models or transcribed audio datasets. This imbalance not only limits the development of direct S2S systems but also forces researchers to rely on pivot-based translation pipelines (e.g., Greek → English → Spanish), which can introduce errors in meaning, prosody, and pragmatic nuance.

To address this gap, a small Spanish-Greek parallel speech corpus was extended by using the VoxPopuli data set (Wang et al., 2021), which contains recordings of the European Parliament in 23 languages, accounting for thousands of hours of speech data. While Spanish includes transcribed segments that are matched to others in 15 other languages, Greek is not one of these – and it remains only available as raw, unannotated interpretation audio. Using enhanced versions of OpenAI's Whisper ASR system (Bain et al., 2023), with word-level timestamps, both languages' audio data is transcribed and then aligned with HunAlign (Halácsy et al., 2008).

This project explores the viability of building S2S resources for underrepresented languages in a low-resource setting. It highlights both practical obstacles, such as noisy ASR output, error multiplication in speech and text processing, alignment of interpretive translation structures, and the broader need for inclusive speech technologies.

2 Motivation

Heavy reliance on English as an interlingua in human and machine translation tasks is a practice that introduces subtle but critical translation inaccuracies. This is especially problematic for language pairs like Spanish and Greek, where grammatical structure, formality, and prosody do not align naturally with English.

Pivoting Issues. Many machine translation systems still follow a translation path such as Greek → English → Spanish, and vice versa.

This introduces a number of issues:

- Loss of formal registers and politeness markers
- Over-literal translation of grammatical tense and aspect
- Inaccurate handling of idioms or culturally loaded phrases

These issues can be identified in the following examples:

> **EL:** Κύριε Πρόεδρε, θα ήθελα να επισημάνω την εξαιρετική συμβολή της κυρίας Καρντάση.
> *(Polite and formal: "Mr. President, I would like to point out the excellent contribution of Ms. Kardasi.")*
>
> **EN:** *Mr. President, I want to mention the great work of Ms. Kardasi.*
> *(More direct and casual; omits formal politeness and vocative register)*
>
> **ES:** *Señor Presidente, quiero mencionar el gran trabajo de la señora Kardasi.*
> → *Better: Señor Presidente, me gustaría mencionar la extraordinaria contribución de la señora Kardasi.*

> **EL:** Αυτό θα μπορούσε να δημιουργήσει προβλήματα.
> *(Even if using a future simple tense, this represents a 'cautious' future and the idea of possibility: "This could create problems.")*
> **EN:** *This will create problems.*
> **ES:** *Esto creará problemas.*
> → *Better: Esto podría crear/provocar problemas.*

In these examples, Greek speech is often flattened into more casual or ambiguous English expressions, which are then translated again into Spanish without recovering the original nuance. This process accumulates and multiplies errors, distorting the original intended meaning.

Such distortions illustrate the need for direct translation paths and parallel corpora, without relying on English as an intermediary.

Personal Motivation. I am a native speaker of Spanish and an advanced speaker of Greek, and I am particularly interested in low-resource language processing. Greek, while widely spoken in specific contexts, is under-represented in multilingual datasets, with very few resources supporting direct speech translation tasks.

This personal linguistic background and the academic gap in direct ES–EL speech corpora motivated this project. The goal is not only to generate usable parallel data for these languages, but to encourage future research that prioritizes direct translation paths over English-mediated pipelines, particularly in the speech domain.

3 Related Work

Multilingual Speech Datasets. Several large-scale multilingual corpora have become foundational in S2S and ASR research. VoxPopuli (Wang et al., 2021), the primary dataset used in this study, includes over 4000 hours of European Parliament speeches in 23 languages, with automatic transcriptions and aligned translations for only a subset of them. Spanish is well supported with aligned and transcribed content, whereas Greek is included only as raw unlabeled audio. Other datasets like FLEURS (Conneau et al., 2022) (used for initial testing in early stages of development) offer multilingual benchmarks with limited coverage for Greek, but it lacks the domain consistency and general parallelism that we can obtain from political, parliamentary recordings.

Automatic Speech Recognition. WhisperX (Bain et al., 2023), an extension of OpenAI's Whisper (Radford et al., 2023), is designed for accurate, timestamped transcription across multiple languages. It improves upon the original Whisper by incorporating voice activity detection (VAD) and phoneme-level alignment, making it suitable for constructing parallel corpora from raw audio. In our setup, WhisperX models were run on a CTranslate2 format suitable for FasterWhisper pipelines, for speed and memory optimization, due to a lack of computational resources. While Spanish transcription is generally reliable, Greek remains a challenge due to limited training data and custom models, often resulting in hallucinated or incomplete outputs.

Alignment Techniques. To preliminarily align transcribed segments at the sentence level, HunAlign (Halácsy et al., 2008), a lightweight statistical aligner that combines sentence-length heuristics and bilingual dictionary matching. Though efficient and lightweight, it struggles with interpretive translations (and their potential free word order and structural variations), which are common in spoken language settings. Neural alignment methods such as SeamlessAlign (Communication, 2023) and SONAR-based embeddings (Duquenne et al., 2023) offer improved semantic alignment but require large datasets and computational resources. The choice for HunAlign represents an initial, resource-conscious approach suitable for early-stage corpus construction and bootstrapping.

Low-Resource Speech Technologies. Despite advances in multilingual NLP, many languages —Greek among them— remain marginalized in speech technology pipelines. Lack of ASR models, absence of aligned speech-text pairs, and minimal S2S translation data hinder inclusive system development. This work contributes a small, manually validated parallel speech corpus for Spanish–Greek and demonstrates a proof of concept for a minimal but reproducible pipeline for similar low-resource settings.

4 Data and Processing

This project uses audio data from the 2009 subset of the VoxPopuli corpus (Wang et al., 2021), which includes multilingual European Parliament plenary session recordings. The processing involved collecting audio files for both languages based on parallel sessions, converting them to a compatible format, and aligning them based on session metadata.

4.1 Raw Audio Data.

- Total raw audio files: 1565
- Spanish raw audio files: 965
- Greek raw audio files: 600

4.2 Segmented Audio Data.

- Total audio segments extracted with VAD: 4940
- Spanish audio segments after processing: 2604
- Greek audio segments after processing: 2470
- Matched session pairs in parallel-el-es.tsv: 2470
- Average file duration: ca. 20s

4.3 Data Processing

Spanish and Greek Data. All raw unlabeled Greek and Spanish audio available for the year 2009 was downloaded, being made up of 1565 full-length plenary session audios for both languages. Unlike Spanish, Greek files were provided only as raw .ogg audio without any transcription, segmentation, or metadata matching them to specific Spanish/any other language utterances. Spanish metadata was used to gather representative identified portions of plenary sessions. After preprocessing these files into audio formats acceptable by the transcription pipeline (.wav), the Spanish parallel portion contained 2604 files and the Greek 2470 respectively, ranging from short utterances lasting 10s to full-length session recordings of a minimum of half an hour.

Session Matching. To enable parallel processing, we automatically matched Greek and Spanish audio files by their date-based plenary session identifiers embedded in the filenames. All audio was converted from .ogg to .wav to ensure compatibility with WhisperX and to minimize processing errors such as sample-rate mismatches or decoding failures.

Subsequently, a parallel mapping file was designed, where each line listed a matched pair of Greek and Spanish audio segments from the same session, e.g.:

data/parallel/el/0090401-0900-PLENARY-11-el_20090401-15-32-14_5.wav →

data/parallel/es/20090401-0900-PLENARY-11-es_20090401-15-32-14_55.wav

Manual inspection of a subset of these pairs—based on native Spanish fluency and advanced Greek level—confirmed that the content was thematically and contextually parallel, despite structural and interpretive differences inherent in live translation.

Segment Preparation. These matched audio files formed the input for the transcription and alignment pipeline. Before transcription, audio files were pre-segmented using voice activity detection (VAD) embedded in OpenAI's Whisper. This ensured that each segment contained a clean speech utterance without long silences or noise artifacts, improving downstream alignment quality.

These preprocessed and matched files were then fed into the transcription pipeline described in the next section.

Transcription Validation. Timestamped transcriptions were exported in Audacity-compatible formats (eg. .srt) and manually inspected using Audacity. Hallucinations, boundary truncation, and ASR noise were manually flagged and analyzed, and subsequently removed. This removal was either done by small detail changes in the audio segmentation for transcription, or by changing the model used for the task itself.

The latter decision stemmed from the fact that transcription quality varied significantly between languages: Spanish had high ASR accuracy, while Greek performance suffered from limited training data and vocabulary mismatches, leading to many errors in transcription and careful choices with regards to the transcription models to use.

External Datasets. For initial pipeline testing and validation, we also used the test sets of FLEURS (Conneau et al., 2022), a multilingual benchmark from Google, and two audiobook recordings (e.g., LibriVox). However, due to alignment inconsistencies and stylistic mismatches, these were not included in the final corpus, and were merely used to validate core logic modules.

5 Methodology

The methodology consists of three main stages: (1) segmentation of Greek audio guided by Spanish VAD, (2) transcription of both languages using WhisperX and Faster-Whisper, and (3) sentence alignment using HunAlign. The entire pipeline was implemented in Python, with modular components for preprocessing, transcription, cleaning, and alignment.

- **Audio:** VAD-guided from Spanish timestamps
- **Transcription:** WhisperX + Faster-Whisper, and subsequent cleaning based on regexp filtering and hallucination removal with sensible model choices
- **Alignment:** HunAlign with manual inspection

5.1 Audio Processing

Since Greek recordings are only provided as raw .ogg files without segmentation, we used timestamp metadata from the Spanish ASR manifest to segment the Greek audio. Each Spanish segment's VAD timestamps were applied to the corresponding Greek file using waveform slicing via torchaudio. Segments shorter than one second or containing silence or invalid samples (e.g., NaNs or all zeros) were skipped.

All Greek and Spanish files were converted to .wav using ffmpeg before segmentation. Finally, a parallel manifest parallel-el-es.tsv was generated, matching each parallel plenary session audio for this language pair.

5.1.1 Audio Chunking

Audio duration was also computed to dynamically determine whether to chunk a longer file or transcribe the full file in one go. Long audio files were automatically split into overlapping chunks (e.g., 10 seconds with 1-second overlap, in order to avoid potentially cutting boundary words out of transcription) and transcribed in parallel using a thread pool. Transcription outputs included timestamps for each one of these manually or automatically-generated chunks, and were saved as Audacity-compatible .srt files, for their subsequent validation.

5.2 Transcription Pipeline

Each language was transcribed independently using models optimized for its linguistic features. The transcription system supports both HuggingFace custom models (for Greek) and Whisper-native models (for Spanish) using FasterWhisper.

This transcription pipeline was in fact the first implemented core element of the project, as the viability of different audio durations, their respective language, as well as the models used were evaluated with regards to future design and implementation choices.The functional tests for this core element were done in two publicly-available completely parallel audios: two audio book recordings of the first chapter of the first part of Miguel de Cervantes' Don Quixote.

The choices for the following models stem from the need for balancing out the resource hunger from the transcription pipeline, the lack of strong-enough computational resources, and the obvious need for a language-specific transcription model that minimized ASR errors in this first stage of the pipeline.

- **Spanish (es):** Whisper large-v3 model (in C2Translate format from HuggingFace's Systran/faster-whisper-large-v3 [1])
- **Greek (el):** Custom Whisper large-v2 model from HuggingFace (Sandiago21/whisper-large-v2-greek [2])

A crucial error that appeared in the early stage of development, and highlighted the necessity for such a custom Greek transcription model, was the following:

> 130.031 139.99 Υπότιτλοι authorwave

which was a default, erroneous transcription result for entire segments not recognized during VAD, that appeared multiple times during transcriptions of longer Greek audio files. When this issue was fixed and a working pipeline was achieved, it was tested on two long completely parallel speech files, achieving optimistic results:

> 0.031 9.99 Μέρος πρώτων, κεφάλων πρώτων Της ο χαρακτήρ και τίνες εασχολείέ του εξακουστού δον Κιχώτου του Μαγκησίου
>
> 0.031 9.99 Capítulo I. Que trata de la condición y ejercicio del famoso Hidalgo don Quijote de la Mancha.

5.2.1 Timestamp Alignment

Alignment was performed with WhisperX using preloaded language-specific aligners. The transcription segments were aligned to the audio waveform at the word level. This produced time-aligned segments for both languages, which were further cleaned and saved in JSON and plain text formats.

5.2.2 Cleaning and Filtering

Errors multiply and accumulated throughout the ASR and timestamp alignment steps, into the final results:

- Faulty ASR output (e.g., hallucinated names, wrong punctuation) impacting alignment. For example:

  > η ισότητας για ο λαός αναπτή, η ισότητας, (...), η δω, (...), η δ -0.3

- Misaligned timestamps causing inaccurate VAD-based segmentation. It must also be taken into account that segment duration might not be 1:1 between languages in terms of what each segment says.

To mitigate these issues, segments were manually reviewed and filtered; and cleaning pipelines (e.g., hallucination filters) were applied. Efforts were put into removing lines that included consecutive repeated words and phrases, based on regular expressions.

The filtered outputs were stored under a dedicated directory structure for both Greek and Spanish clean transcriptions. However, more and stronger efforts remain to be dedicated to this part, as even despite filtering of the results, there still remain results with intense hallucinations and other crucial errors originating by the ASR transcription alignment steps.

5.3 Sentence Alignment

Final textual sentence-level alignment was performed using HunAlign[3], a statistical aligner that uses sentence-length heuristics and optional bilingual dictionaries. In this case, the

[1] https://huggingface.co/Systran/faster-whisper-large-v3

[2] https://huggingface.co/Sandiago21/whisper-large-v2-greek

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

-text mode was used without dictionaries due to the lack of such a quickly-available resource for this language pair.

Cleaned transcription files for each language were paired based on their shared segment identifier, and alignment was carried out for each matched pair. The final alignments for each one of the plenary sessions consist on textual files with each line containing the two languages texts, side by side, and their respective alignment score.

6 Limitations

Building this Spanish–Greek speech corpus required significant computational resources and time. Despite using optimized tools such as WhisperX and FasterWhisper, multiple challenges emerged during implementation:

- Processing power — transcribing and aligning large datasets is resource-intensive.
- ASR performance — especially for Greek, which lacks robust, widely-available models.
- Segment alignment — interpretive translations are harder to align than literal ones.

Resource Constraints. The primary bottleneck in development and testing was compute availability.Whisper-based models are computationally expensive, especially for long-form audio and under-resourced languages like Greek.

During the initial development of the transcription pipeline, once a working version was achieved, transcribing a single 10-minute file from the Don Quixote audiobook took several hours on standard hardware. As a result, all model inference had to be offloaded to cloud infrastructure.

Google Colab was used as the primary computing environment, providing limited but sufficient access to higher-efficiency and memory RAM, and different GPU and CPU. However, despite using a premium-tier Colab plan, frequent memory limits and worst of all, session expiration, significantly slowed overall performance and efficiency. Moreover, even then, the pipeline's code had to be carefully adjusted to correctly allocate and utilize resources at each step in its execution.

Processing Time. The combined pipeline of transcription + alignment was time-intensive, which clashed with the inherent lack of real time resources for the devleopment of the project. For each language:

- Transcribing a single segment often took between 5 to 15 minutes, depending on length and model size.
- Alignment using WhisperX added additional overhead, especially for Greek due to lower model accuracy and longer audio segments.
- Overall, processing a small sample of segments (transcription + alignment) took approximately 2.5 hours per language.

Scaling Limitations. Due to the above constraints, we were only able to fully process and align 33 Spanish–Greek segment pairs. A full run over the remaining 2440 pairs (only for the identified parallel segments in the 2009 timeline) was deemed infeasible within the available timeframe and infrastructure. Expected runtime for each language was estimated to be up to 200 human hours.

7 Results

While all previous steps in development certainly helped improve final quality of transcription and alignment of these parallel speeches' transcriptions, the fundamental constraint remained: without substantial GPU resources and time, large-scale transcription and alignment of raw speech remains a bottleneck in low-resource pipeline construction.

7.1 Processed Data

From a total of 1565 downloaded plenary session audio files across Spanish and Greek from VoxPopuli's 2009 records, the system processed 965 Spanish and 600 Greek audio files.

Using a voice activity detection (VAD)-based segmentation and session-matching strategy, this yielded:

- 4940 total usable audio segments

- 2470 matched Spanish–Greek segment pairs (approximately 50% of total segments)

These matched segment pairs represent a significant retention rate of usable parallel data, considering the lack of direct alignment and metadata in the available VoxPopuli data for Greek.

7.2 Transcription Efficiency

Due to hardware limitations, transcription was a major time bottleneck. Of all the identified parallel segment pairs, only a sample of 33 of them was successfully transcribed and aligned, in order to achieve an example of the pipeline's reach. This corresponds with a ∼1.3% final yield of parallel segments processed.

Spanish transcription was comparatively faster when using WhisperX's default large-v3 model, but this had to be compromised due to the extensive resources it required. This made processing each language equivalent in terms of the time each required.

- Average transcription time per segment: 5–8 minutes
- Total transcription time for all Spanish and Greek segments: ca. ∼200 hours per language
- Segment pairs transcription time: ca. 2.5 hours per language

On top of this, the need for post-processing and cleaning of Greek transcriptions — aimed at removing hallucinated or malformed segments — further slowed down the tasks.

7.3 Alignment Results

HunAlign's numerical scores are deemed unreliable, considering the acknowledgement of the structural difference of these pairs, but with some knowledge of the two languages and manual evaluation of the obtained sample, it is visible that the transcribed sentences are in fact parallel.

Although syntactically divergent, aligned segment pairs show thematic overlap between source and target, which validates their inclusion in a manually verified parallel dataset. In general, errors and negative results in textual sentence-level alignment come as a result of errors propagated all the way from the ASR task at the start of the pipeline.

For example, consider the following examples:

> **ES:** No solo vivimos una crisis económica y financiera, vivimos una crisis de valores.
> *Translation (EN/ES): Not only are we living through an economic and financial crisis, we are living a crisis of values.*
>
> **EL:** Σωστή διάγνωση της κατάστασης είναι δεν είναι μόνο οικονομική και χρηματοπιστωτική κρίση, είναι και κρίση αξιών.
> *Translation (EN/EL): A correct diagnosis of the situation is that it is not only an economic and financial crisis, it is also a crisis of values.*
>
> **ES:** El primer elemento común es que hoy hay una gran convergencia en la opinión de que hay que reducir las emisiones y que hay que desarrollar energías limpias.
> *Translation (EN/ES): The first common element is that today there is a great convergence in the opinion that we must reduce emissions and develop clean energies.*
>
> **EL:** Το πρώτο κοινοστοιχείο είναι πως σήμερα υπάρχει σύγκληση στο ότι θα πρέπει να περιοριστούν.
> *Translation (EN/EL): The first common element is that today there is a consensus that they [emissions] should be limited.*

Despite the small number of successfully aligned pairs, this sample demonstrates feasibility and provides a base for expanding the dataset, provided that the necessary computational resources are available.

8 Conclusions

This project presents a practical and resource-aware attempt to build a Spanish–Greek parallel speech corpus, constructed from European

Parliament interpretation audio via the VoxPopuli dataset. Despite significant obstacles —such as limited Greek ASR quality, interpretive rather than literal translations, and computational resource constraints— we were able to successfully align 33 high-quality segment pairs out of an initially processed pool of 2470.

While small in size, this aligned subset demonstrates the feasibility of direct speech-to-speech dataset construction for underrepresented language pairs. Our pipeline combines WhisperX-based timestamped transcription, custom-trained Greek model usage, and HunAlign-based alignment, forming a baseline architecture for similar future initiatives.

This work also sheds light on broader issues in multilingual translation research, particularly the bias toward English-pivoting pipelines and the exclusion of less-represented languages from direct training data. Through this corpus and its methodology, I hope to contribute not only data but also attention toward the challenges and needs of inclusive speech technology development.

9 Future Work

To scale this project beyond its current, identified limitations, several directions are proposed:

- Scale the dataset: Transcribe and align the full 2009–2020 VoxPopuli sessions and extend to additional Spanish–Greek interpretation data.

- Upgrade alignment: Use neural alignment systems like SeamlessAlign and cross-lingual embeddings from SONAR to improve sentence matching accuracy.

- Interpretation paths metadata: Automatically detect interpretation direction, original speech language, and speaker roles from session data.

- Release and collaboration: Publish the cleaned and aligned corpus via HuggingFace or LINDAT, and invite community contributions to extend it – and potentially apply it to other low-resource language pairs.

These efforts would greatly expand the reach and usability of the dataset, and the extension of the pipeline to a wider range of languages, contributing to more equitable multilingual speech translation systems.

Acknowledgments

I would like to thank Ondřej Bojar, my instructor for the NPFL087 Statistical Machine Translation course at ÚFAL in Charles University, as well as Dominik Macháček, for their support, guidance and motivation throughout the development (and obstacles) of this project.

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Seamless Communication. 2023. Seamlessm4t: Massively multilingual multimodal machine translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Péter Halácsy, András Kornai, Péter Németh, and Dániel Varga. 2008. Parallel creation of Gigaword corpora for medium density languages - an interim report.

Alec Radford, Jong Wook Jeong, and et al. 2023. Robust speech recognition via large-scale weak supervision. https://github.com/openai/whisper. OpenAI Whisper.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.