ES COCO: Removing Barriers to Language Research With an Open-Access Database for Spanish-English Speech Data

Natasha Vernooij[a], Natalie Robbins[bcd], Logan Walls[a], Eric Martell[a]
Departments of: [a]Psychology; [b]Cognitive Science; [c]Linguistics; [d]Romance Languages and Literature
Contact: vernooij@umich.edu
Presented at the 2022 U-M data science and AI Summit

LSA PSYCHOLOGY
UNIVERSITY OF MICHIGAN

Introduction and Proposal

We will create an open-access data repository of Spanish-English bilingual speech called **ES COCO** (**E**nglish-**S**panish **CO**de-switching **CO**rpus).
Some Spanish-English bilingual speech data sources already exist, such as BilingBank (MacWhinney, 2019). However, these data sources are not easily amenable to analysis because researchers must aggregate data across many files and sites.

ES COCO will contain:

- **annotated speech** from podcasts and already-existing corpora
- **metadata** such as speaker and demographic information

This will be the largest self-contained Spanish-English corpus, enabling researchers to analyze the data without manually aggregating across many disparate sources.

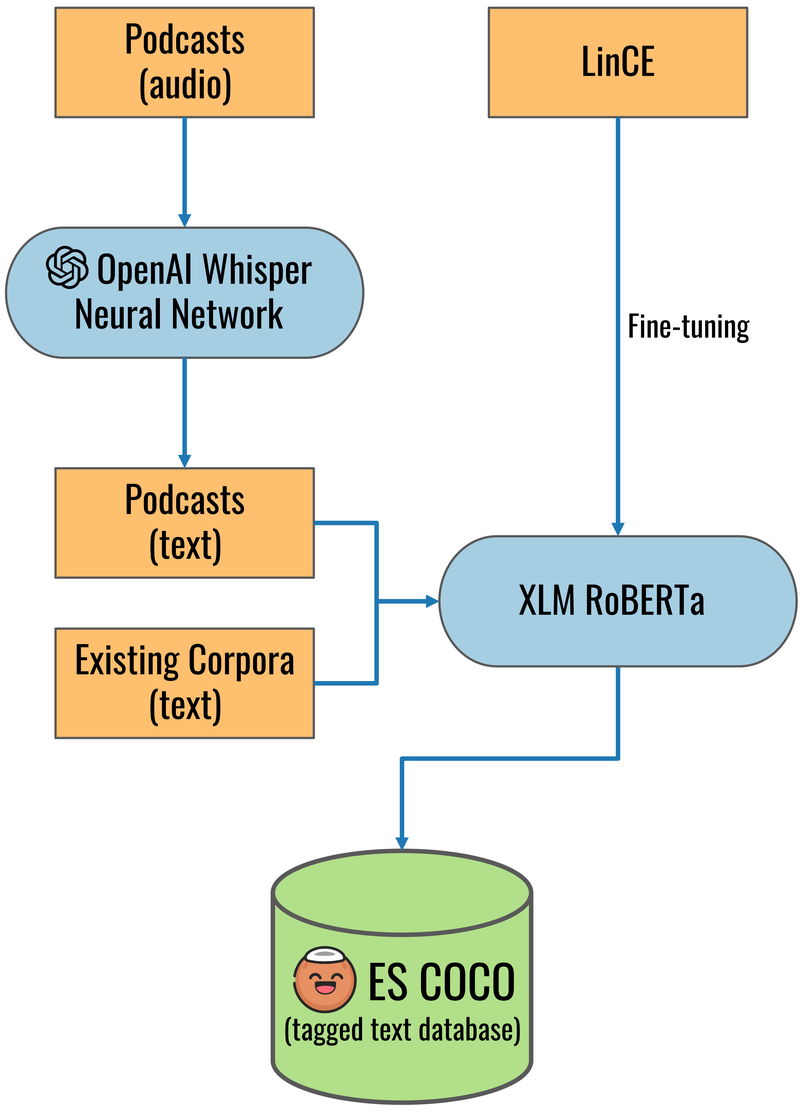
Method

Speech-to-text conversion

ES COCO will draw much of its data from recorded audio conversations between bilingual speakers. We will convert this data to text so that it can be annotated using natural language processing techniques. To accomplish this, we will use **OpenAI's Whisper** (Radford et al. 2022), a neural network model for automatic speech recognition.

Text annotation

For both transcribed audio and existing textual corpora, we will annotate each word with linguistic features such as which language it comes from, what part-of-speech it is, and its sentiment in context. We will achieve this using **XLM RoBERTa,** a transformer language model (Conneau et al., 2019), that we will fine-tune using bilingual-focused datasets from **LinCE** (Aguilar et al., 2020).



References

Aguilar, G., Kar, S., & Solorio, T. (2020). Lince: A centralized benchmark for linguistic code-switching evaluation. arXiv preprint arXiv:2005.04322. https://doi.org/10.48550/arXiv.2005.04322
Conneau, A., et al. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.https://doi.org/10.48550/arXiv.1911.02116
MacWhinney, B. (2019). TalkBank and SLA. In N. Tracy-Ventura & M. Paquot (Eds.), *The Handbook of SLA and Corpora*. Routledge.
Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. https://openai.com/blog/whisper/

Product

ES COCO will have a user-friendly interface that can run locally on a user's machine or be accessed via web browser.

Users can:

- **search** and **filter** the corpus by linguistic features (e.g., part of speech, language, code switch)
- view the **linguistic context** (i.e., the sentence)
- access speaker information (e.g., demographics, proficiency)
- access file information (i.e., source)
- download data for their own analyses

Conclusion

ES COCO removes the largest barriers in language research: the time and financial cost of collecting, transcribing, and tagging data. This corpus is particularly beneficial for researchers who are not at R1 institutions and who have limited access to funding, personnel, time, and the language communities required for language research.
See our project and progress at the QR code below.