Creating a corpus of web-data with Pyrlato. A demonstration.

Giuseppe Magistro and Claudia Crocco

Ghent University

The use of corpora in acoustic analyses has become a standard practice in phonetic phonological research, offering high ecological validity (see e.g. Beckman, 1997; Warner, 2012; Tucker & Mukai, 2023 for a discussion on validity). However, compiling corpora and looking for specific phenomena can be time and resource-consuming. In response to this challenge, we developed a program named *Pyrlato* (author&author), which we aim to demonstrate in the workshop. Pyrlato is a novel tool designed for creating corpora of real-world spoken data from the web. The tool extracts audio files from YouTube, cutting and extracting desired segments such as specific phonemes, syllables, or words found in YouTube videos. This enables the creation of corpora with tens of thousands of tokens within a few computational hours. Pyrlato works across Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Turkish, Ukrainian, and Vietnamese, i.e. those languages for which YouTube provides automatic subtitles. The software searches for the desired string in the subtitles and, upon finding the match, extracts the relevant audio extract containing the string in .mp3 format (other formats are also possible). Pyrlato relieson theexternallibrariesMoviepyand PytubetoconvertYouTube videos into Python objects containing a method to extract subtitles with their respective timestamps.

The demonstration will showcase Pyrlato's online version (link: https://colab.research.google.com/drive/1zv67DpKehySkJLd0Vk7hpG7__Je4fbi4?usp=sharing), requiring no installation to ensure a smooth and reproducible participation by the workshop attendees. We will initially showcase its basic functionalities: we will build a corpus of specific words in the purpose of phonetic and phonological studies (e.g., the German modal particle "doch," see Egg and Zimmermann, 2012; Repp & Seeliger, 2023) and decide the size of the cut interval which contains the desired word. Subsequently, the tutorial will delve into more complex queries, showing that Pyrlato can search phonetic combinations in specific morphological contexts, such as geminate consonants in Italian or the realization of the suffix "-ig" in German, hence limiting the search to positional constraints. The session will then guide participants in more intricate searches, e.g. on how to express alternative strings and instructing the program for more fine-grained output in diverse contexts.

We will also address some other complications that are useful to study sociolinguistic variation, e.g. we will illustrate techniques to limit the output to specific YouTube channels (e.g., BBC channel), particular varieties (e.g., British English), or defined situational contexts (e.g., audiobooks or interviews), or the exclusion of unwanted speakers or topics. Advanced features, including video downloads for multimodal research, will also be shown. At the end of the session, we will show to quickly convert the files in another format, rename them and download the corpus in batch. We would like to integrate the demonstration with a Q&A, session encouraging participants to explore Pyrlato's application in their research domains and to provide feedback on the tool. We believe Pyrlato holds substantial potential, and while the project is in its infancy, widespread dissemination, feedback, and collaborative input are crucial for its improvement to build a tool which can cater to the needs of linguists in corpus creation, labeling, deployment and maintenance.

References

• Author&Author. Information removed for ensuring the anonymity.

• Beckman, M.E. (1997).A typology of spontaneous speech. In Y. Sagisaka, N. Campbell, & N. Higuchi (Eds.), Computing Prosody: Computational Models for Processing Spontaneous Speech (pp. 7–26). Springer. http://dx.doi.org/10.1007/978-1-4612-2258-3_2.

• Egg, M.,& Zimmermann, M. (2012). Stressed out! Accented discourse particles: The case of DOCH. In A. Aguilar Guevara, A. Chernilovskaya, & R. Nouwen (Eds.),Proceedings of Sinn und Bedeutung 16(pp. 225–238). MIT Press.

• Repp, S.,& Seeliger, H. (2023) Reject?! On the prosody of non acceptance. In R. Skarnitzl, & J. Volín (Eds.), Proceedings of the 20thInterna-tional Congress of Phonetic Sciences (pp. 1355–1359). Guarant International.

• Tucker, B.V., & Mukai, Y. (2023). Spontaneous speech. Cambridge University Press. http://doi.org/10.1017/9781108943024.

• Warner, N. (2012). Methods for studying spontaneous speech. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), The Oxford Handbook of Laboratory Phonology (pp. 621–633). Oxford University Press.