On the Advantages and Challenges of Working with Large Corpora of Naturalistic Speech

Johanna Cronenberg[1] and Ioana Chitoran[2]

[1]Université Paris Cité, [2]Université Paris Diderot

In 1972, William Labov stated that “[t]he aim of linguistic research in the community must be to find out how people talk when they are not being systematically observed; yet we can only obtain this data by systematic observation” [1, p. 209]. More than fifty years later, we are finally in the fortunate position of no longer having to resort to conducting secret recordings in New York City department stores. We are now equipped with both the computational capabilities and vast collections of naturalistic data to elude the Observer’s Paradox – yet corpus studies still seem to be relatively rare in phonetics. In this talk we plan to illustrate some of the advantages and challenges associated with large corpora of naturalistic speech, hoping to spark a discussion that will be of interest to both corpus phoneticians and those who want to become one.

Our recent work uses recordings of radio and TV shows of five Romance languages (981 hours total; for more information on the corpora see [2]) which were not collected for the purpose of linguistic analysis. The aim of our ongoing acoustic analyses is to provide a comprehensive picture of the distinction between diphthongs (e.g. /ja, jo/) and hiatuses (e.g. /ia, io/) in European French, Italian, Spanish, Romanian, and Portuguese. Large-scale studies and crosslinguistic comparisons [3, 4, 5], but also longitudinal investigations [6, 7, 8] are much more feasible and ecological when using existing corpora instead of collecting new data. In addition, naturalistic speech in combination with larger sample sizes make it “possible to test whether effects that arise in experimental or intuition-based studies are widespread and meaningful” [9, p. 8]. Our analyses, for instance, include 104,667 occurrences of /ia/ and 135,182 occurrences of /io/ across the five languages which will allow us to test whether there are robust differences between the languages with regard to the diphthong-hiatus distinction, as has been claimed from historical and phonological perspectives [10].

Since it is too time-consuming to listen to a large speech corpus in its entirety, getting to know the data needs to take on a different form. This was challenging because the corpora we are using came without metadata. However, information about the speech style (broadcast news vs. interviews or debates), the speaker’s gender and region of origin, as well as the recording date were encoded in the file names for parts of the data and could be extracted using regular expressions and data manipulation techniques. While the data had already been parsed through an automatic speech recognition system [11, 12], the resulting phonemic segmentation and alignment cannot be checked manually. A perhaps underestimated tool to find possible alignment errors is summary statistics: for example, unreasonably long segments or those that have been assigned the default duration of 30ms might have been misaligned and are worth checking by hand. We have also encountered data quality issues, such as background noise and music, which require more complex solutions such as training a classifier to recognise poor audio quality.

Despite the challenges associated with corpus work, we ultimately encourage speech scientists to use and re-use available corpora as well as to mine data from public sources such as radio archives and make them available to other researchers. Corpus studies complement laboratory studies because they can provide insights into spoken language under naturalistic circumstances and they facilitate the usage of larger, multilingual, and/or longitudinal samples.

References

[1] Labov, W. (1972): *Sociolinguistic Patterns.* Blackwell.

[2] Vasilescu, I., Wu, Y., Jatteau, A., Adda-Decker, M., and Lamel, L. (2020): Alternances de voisement et processus de lénition et de fortition: Une étude automatisée de grands corpus en cinq langues romanes. *Traitement Automatique des Langues* 61(1), 11-36.

[3] Ahn, E., and Chodroff, E. (2022): VoxCommunis: A Corpus for Cross-Linguistic Phonetic Analysis. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 5286-5294.

[4] Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. (2021): VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 993-1003.

[5] the Authors: To be added after review.

[6] Puggaard-Rode, R., Horslund, C. S., and Jørgensen, H. (2022): The rarity of intervocalic voicing of stops in Danish spontaneous speech. *Laboratory Phonology* 13(1), 1-47.

[7] Harrington, J., Palethorpe, S., and Watson, C. I. (2000): Does the Queen speak the Queen's English? *Nature* 408(6815), 927-928.

[8] Quené, H. (2013): Longitudinal trends in speech tempo: The case of Queen Beatrix. *The Journal of the Acoustical Society of America* 133(6), EL452-EL457.

[9] Cohn, A. C., and Renwick, M. E. L. (2019): Doing phonology in the age of big data. *Cornell Working Papers in Phonetics and Phonology 2019*, 1-36.

[10] the Authors: To be added after review.

[11] Lamel, L., and Gauvain, J.-L. (1992): Continuous Speech Recognition at LIMSI. *Proceedings of the Final Review of the DARPA ANNT Speech Program*, 1-7.

[12] Adda-Decker, M., and Lamel, L. (1999): Pronunciation variants across system configuration, language and speaking style. *Speech Communication* 29, 83-98.