Data cleaning methods to improve the accuracy of coding in the Projet Phonologie du Français Contemporain *Corpus: A Case Study of Schwa Realization*

Joshua M. Griffiths
The University of Texas at Austin

Studies of French phonology have recently taken a methodological shift to allow for the use of large corpora, particularly since the advent of the *Projet Phonologie du Français Contemporain* (PFC) corpus (Durand, Laks, and Lyche 2002). This corpus, coded specifically for both schwa and liaison, two complex, variable hallmarks of French phonology, has not only been useful in studies of French phonology, but also in studies of phonological variation. While the PFC has been instrumental in shaping how phonology is approached due in large part to its coding system and sheer quantity of available data, the corpus still does have its shortcomings that appear to have gone largely unnoticed in the literature. For instance, when queried for all instances of schwa uttered in the unguided and guided conversations of speakers of continental French, the corpus returned 94,080 instances of schwa; however, once pre-processed for analysis, 31,982 of these data did not contain an underlying or epenthesized schwa. This is problematic considering the sheer amount of research that has relied on this corpus since its initial publication, including many articles, book chapters, doctoral dissertations, and edited volumes (Durand et al. 2009; Gess et al. 2012), many of these works have been quite influential, and without careful attention of the researcher, these results could be misleading or inaccurate. In addition to the inaccuracy of the coding, the data are coded so globally that it is often difficult to make any particular claims pertaining to the phonology of French since the corpus only specifies whether the segment preceding and following schwa is a vowel, a consonant, a reduced consonant cluster, and a strong or weak intonation break.

In this presentation, I highlight the methodological undertaking required to specify which of the data generated by the PFC do actually contain schwa, and how best to annotate them for phonological analysis. The data pre-processing and analysis were done in R (R Core Team 2018). For this particular analysis, I define French schwa as a mid-vowel that can alternate with zero and is represented orthographically as <*e*>. While phonologists have argued extensively over the definition of schwa, the graphemic representation is often a necessary criterion for its definition (Dell 1973). While defining a phonological phenomenon with an orthographic convention may be problematic as it admits non-phonological factors into the grammar, this definition permits a precise identification of schwa in an annotated corpus. Including this criterion into pre-processing eliminated what were erroneously coded as 24,130 instances of deleted schwa. In addition to the removal of non-schwas from the data, the pre-processing script also specified the sound immediately preceding and following schwa in a more granular fashion, providing insight into the natural classes of the neighboring sounds, allowing for a more fine-tuned phonological analysis in any framework.

Along with the methodological considerations discussed, I present the results of a generalized linear mixed-effects regression model to which the data following the new coding specifications were fit (Jaeger 2008, Baayen 2012). Each post-consonantal schwa was subject to analysis. (*n*=54,890). Multiple models were constructed and were compared by their Akaike Information Criterion (AIC). The model which best fit the data at hand (AIC=40,109.63) treated the speaker's region (Northern France v. Southern France), the type of word (lexical v. functional), schwa's distribution in the word, as well as an interaction term of the natural classes of sounds that preceded and followed the schwa as fixed-effects. Speaker, word, and manual coder of the data in question were all treated as random effects. The particular model found that there were in fact multiple natural classes that were predictive of schwa deletion and realization. Concerning the interaction term, the odds ratios of the coefficients were taken to find that schwa deletion was most likely to occur between a voiceless stop and a prosodic break (ie *petite*## /pə.ti.t##/ 'small.FEM.SING') , and schwa realization was most likely to occur between a nasal consonant and voiced fricative (ie *mesure* /mə.zyʁ/ 'measure') . In addition to the natural classes and their interactions, an expected regional effect in which southern French was significantly more predictive of schwa realization than Northern French, as well as an effect of word type in which lexical words were significantly more predictive of schwa realization than functional words, indicating that schwa is often not maintained in monosyllabic clitics, determiners, and the conjunction *que* in addition to other polysyllabic prepositions. Furthermore, there was a significant effect of schwa's distribution in the word. The model found schwa is most likely to be realized in word-internal position.

While the PFC is an invaluable contribution to French and corpus linguistics, this study suggests that the PFC may need some other means for tagging the data, be it some protocol for subjecting the annotations to measures of inter-rater reliability or adapting an automated tagging algorithm for identifying schwas for coding. With over one-third of the data being unusable, the current coding protocol needs to be seriously re-evaluated.

References

Baayen, R. Harald. 2012. Mixed-effects models. *The Oxford Handbook of Laboratory Phonology*, ed. by Abigail C. Cohn, Cécile Fougeron, Marie K. Huffman, and Margaret E. L. Renwick, 668-678. Oxford: Oxford University Press.

Dell, François. 1973. *Les règles et les sons.* Paris: Hermann.

Durand, Jacques; Bernard Laks; and Chantal Lyche. 2002. La phonologie du français contemporain : Usages, variétés, et structure. *Romanistische Korpuslinguistik – Korpora und gesprochene sprache.* ed. by Claus D. Pusch and Wolfgang Raible, 93-106. Tübingen: Gunter Narr Verlag.

Durand, Jacques; Bernard Laks; and Chantal Lyche (eds.) 2009. *Phonologie, variation et accents du français*. Paris: Hermès.

Gess, Randall; Chantal Lyche; and Trudel Meisenburg (eds.) 2012. *Phonological variation in French: Illustrations from three continents.* 11; (Studies in Language Variation.) Amsterdam: John Benjamins Publishing.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59.434-46.

R Core Team. 2018. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Online: https://www.R-project.org/.