Twitter-MPhon: Studying morphophonological variation with Twitter data

Michael Dow, François Lareau, Patrick Drouin · Université de Montréal

Introduction Variation has come to take an important place in linguistics, and with the rising ubiquity of quantitative methods, data-driven observations are becoming increasingly central to this endeavour. For instance, while non-normative use of *a* + vowel-initial words has long been noted (Wright, 1905), rates vary drastically according to sociolinguistic factors such as dialect (Orton and Halliday, 1963), race (Labov, 1972) and class (Lass, 2002), as well as stress pattern (Raymond et al., 2002). Such variation is not an end to itself; its study directly informs formalism and theory, such as the use of French *h-aspiré* variation in Stochastic OT (Gabriel and Meisenburg, 2009), listener-oriented OT (Boersma, 2007) and learning algorithms with gradient phonological representations (Tessier and Jesney, 2021). In this project, we explore the use of Twitter as another potential source of data in the study of morphophonological variation. In particular, we look at rates of variation in two phenomena, namely *h-aspiré* in French and indefinite allomorphy in English, in comparison with pre-existing experimental or corpus studies.

Methodology Wordlists for each phenomenon were constructed semi-automatically according to a combination of phonetic and orthographic factors and lexicographic information from Wiktionary (Ylonen, 2022) and frequency from the most recent French and English Web corpora in Sketch Engine (Kilgarriff et al., 2014). Criteria for **h-aspiré** nouns included type of initial phoneme and documented blocking of external sandhi. Expressions with these words were then created with their corresponding full definite article and the elided variant (e.g., **le hôtel, l'hôtel*; *le hibou, *l'hibou*). English nouns and adjectives were chosen according to initial phoneme, presence of initial stress, and vowel height, where relevant. These words were then combined with each of the two variants of the indefinite article. Via accounts with Academic Research-level access, we used the Python package `tweepy` to gather tweet counts for each literal expression, controlling for language, from the full Twitter archive dating from March 2006. Exhaustive counts of each expression were then gathered, and for each word, a ratio of normative-to-total forms was calculated for the whole corpus with respect to the above factors. Collection of the full tweets and their metadata remains ongoing.

Results Use of normative forms is high for both phenomena, with some notable exceptions. In French, of the prescriptively aspirated words, only *hélas* 'alas', *harcèlement* 'harassment' and *handicap* showed rates of the full article beneath 95% (59%, 76% and 87%, respectively), and vowel-initial months of the year suggested patterning with aspirated words. Of the vowel-initial English words, we observed the lowest rates of *an* before stressed high vowels and mid vowels, though still fairly high (92% and 93%, respectively). Specific /h/- and /j/-initial words showed lower rates of *a* (e.g., *historic* (82%) and *euphoria* (76%)).

Discussion Results so far suggest that, despite boasting a diverse set of users and being an often informal (albeit written) medium, Twitter data show relatively little variation for these phenomena. In more traditional, oral corpora, English allomorphy boasts ranges of *a* + vowel usage from 5% (Fox, 2015) and 15% (Gabrielatos et al., 2010) up to 90% (Ash and Myhill, 1986), depending on previously mentioned factors. Meanwhile, Moisset (1996) finds application of external sandhi to 13% (92/686) of *h-aspiré* tokens, and higher rates are found by Gabriel and Meisenburg (2009) and Tessier and Jesney (2021), though on much fewer tokens. Further research is needed to clarify our data and these comparisons. First, metadata should allow us to infer and explore crucial sociolinguistic variables, as well as to use device information (mobile vs. web) to address the potential confound of autocorrect. Finally, we note a disparity in size between our corpus and others; further discussion on its potential effect on percentage-based variation may be worthwhile. In conclusion, Twitter data allow us to control for linguistic variables and gather massive corpora at a fraction of the time and cost of traditional corpora, though their viability in studying morphophonological variation and their extension to formal, theoretical accounts is still uncertain.

References

Ash, Sharon and John Myhill. 1986. Linguistic correlates of inter-ethnic contact. *Diversity and diachrony* 53: 33–44.

Boersma, Paul. 2007. Some listener-oriented accounts of h-aspiré in French. *Lingua* 117(12): 1989–2054.

Fox, Susan. 2015. *The New Cockney: New ethnicities and adolescent speech in the traditional East End of London*. Palgrave Macmillan.

Gabriel, Christoph and Trudel Meisenburg. 2009. Silent onsets? an optimality-theoretic approach to french h aspiré words. *Variation and gradience in phonetics and phonology* : 163–184.

Gabrielatos, Costas, Eivind Nessa Torgersen, Sebastian Hoffmann, and Susan Fox. 2010. A corpus-based sociolinguistic study of indefinite article forms in London English. *Journal of English Linguistics* 38(4): 297–334.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1: 7–36.

Labov, William. 1972. *Language in the inner city: Studies in the Black English vernacular*. 3. University of Pennsylvania Press.

Lass, Roger. 2002. South African English. *Language in South Africa* 104126: 104–126.

Moisset, Christine. 1996. The status of ‘h aspire’ in French today. *University of Pennsylvania Working Papers in Linguistics* 3(1): 17.

Orton, Harold and Wilfrid J. Halliday. 1963. *The survey of English dialects, volume 1: The six northern counties and the Isle of Man*. Leeds, UK.

Raymond, William D, Julia A Fisher, and Alice F Healy. 2002. Linguistic knowledge and language performance in English article variant preference. *Language and Cognitive Processes* 17(6): 613–662.

Tessier, Anne-Michelle and Karen Jesney. 2021. Learning French liaison with gradient symbolic representations: Errors, predictions, consequences. In *Proceedings of the Annual Meetings on Phonology*, vol. 8.

Wright, Joseph. 1905. *The English dialect grammar*. Oxford University Press.

Ylonen, Tatu. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1317–1325.