Synchronic variation and sound change in Romance languages: a corpus-based study of lenition phenomena in Romanian and Spanish

Ioana Vasilescu, Lori Lamel
LIMSI CNRS, Université Paris-Saclay

Many diachronic phonological changes mirror synchronic phonetic variations, and as J. Ohala proposed, some historical processes can be “duplicated” in laboratory conditions through acoustic and perceptual experiments [1]. We compare two lenition phenomena in Romanian and Spanish through corpus-based analyses, both of which are linked to phonological changes from Latin to modern Romance languages.

In Romanian, we discuss the deletion of the masculine definite article *-l* (pom 'apple' – pomu**l**/pomu 'the apple') whose function is taken over by the desinence *-u-*, in connected speech, and we compare *-l* deletion in all attested contexts (word final +/-definite article, e.g. calu***l*** vs *ca**l*** '(the) horse', word initial, e.g. ***l**ac* 'lake', and internal, e.g. *cale* 'way'). Derived from the Latin personal pronoun *illu(m)*, the definite article is attested since Middle Ages, but an unsettled orthography with and without *-l* persisted until the 20th century when the orthographic norm stabilized and imposed a written *-l*. For Spanish, we address lenition of intervocalic voiced stops /bdg/ in varieties from Spain and Latin America (Caribbean) (e.g. *vi**d**a* 'life' pronounced *via*). The sound change is no longer active, but Spanish still exhibits this phenomenon as synchronic variation.

Our approach is in line with the methodological trend observed in phonetic studies over the last two decades which consists of the increasing use of speech data collections gathered not only in controlled but also in natural settings. In this research framework, corpora have evolved from small, task dependent, to large-scale ones containing dozens to thousands of hours, collected for various purposes, not necessarily linguistic. The large corpora are typically heterogeneous in terms of recording conditions and speaker profiles [2], and provide ideal conditions to study the patterns of phonetic variation, many of which linked to reduction phenomena [3]. Such patterns are strictly speaking reduction, that is deletion of segments, but can also cover a wide variety of coarticulatory events, some of which refer to language specific phonological rules. Large corpora can also be beyond the human capacity of processing for the purpose of linguistic studies and require help from automatic processing. This assistance is provided by tools coming from speech technology, and specifically speech recognition, that through techniques such as forced alignment of the speech signal with the manual transcriptions, are able to convert virtually unlimited quantities of spoken data in material for corpus-based phonetic research.

The corpora used for this study mainly come from research projects dedicated to the development of speech-to-text systems for the two languages of focus [4,5]. For Romanian we are using both semi-prepared and spontaneous speech, consisting of semi-prepared news, televised interviews, dialogues and read speech, totaling 10.5 hours [6]. For Spanish we are using semi-prepared and spontaneous news in Peninsular (13 hours) and Latin American Spanish (Caribbean variety, almost 4 hours). The broadcast data are contrasted with a 5-hour corpus of telephonic conversational speech in Peninsular Spanish [5]. All data are manually transcribed (orthographic level).

On the methodological side, we use speech-to-text systems to estimate the frequency and triggering contexts of the two lenition phenomena. Specifically, we make use of the method which consists of the alignment of the speech signal with both canonical (phonological, here non lenited variants) pronunciations of the words in the dictionary and with their reduced or modified variants (lenited variants, e.g. *pomu* instead of *pomu**l*** for Romanian, and *via* instead of *vi**d**a* for Spanish). This method has been already used in a range of phonetic and laboratory phonology studies, from broad quantification of reduction phenomena as function of the language and speaking style [7,8], to in depth analyses of the relation that can be established between variation and underlying phonological rules (e.g. schwa and liaison in French [9,10,11], dark vs light *-l* occurrence in English [12], and minimal phonological contrast in the Romanian vocalic system [13]).

Using this method, we demonstrate that in Romanian, the highest deletion rates concern the definite article (50% on average) compared to other lexical positions (12%). The former undergoes two constraints: (i) right edge consonantal contexts trigger deletion more often (37%) than vowel contexts (10%) and (ii) the trend is reinforced by the speaking style as spontaneous speech triggers

more deletions (deletion rate ranges from 31% for semi-prepared news to 74% for dialogues). Differences in deletion rates as an effect of the speaking style suggests that *-l* deletion is not a sound change in progress, but that l-realization is a segment insertion triggered by the orthography. Although traces of the definite article *-l* still persist in the nominal inflection (ie. genitive sg. *pomului* 'of the apple', pl. *pomilor* “of the apples”), the deletion rates of word internal *-l-* including such occurrences do not exceed 10%.
In Spanish we measure the extent of the phenomenon as a function of geographical and stylistic repartition. The data show that the distribution of pronunciation variants across Peninsular and Latin American Spanish varieties is consistent with trends depicted by classical linguistic studies. More lenited variants are thus selected by the speech recognition system for Caribbean journalistic speech (40%) compared to Peninsular Spanish (24%). Rates for telephonic conversational (70%) recordings are significantly higher compared to news. In comparison with Romanian, the Spanish data confirms that the synchronic variation follows historical trends which can be highlighted with new methods based on the use of speech-to-text systems as linguistic tools.

[1] Ohala, J. J. 1989. Sound change is drawn from a pool of synchronic variation. In L. E. Breivik & E. H. Jahr (eds.), *Language Change: Contributions to the study of its causes*. [Series: Trends in Linguistics, Studies and Monographs No. 43]. Berlin: Mouton de Gruyter. 173-198.
[2] Ernestus, M. 2011. An introduction to reduced pronunciation variants. *Journal of Phonetics*. 3 (39). 253–260.
[3] Adda-Decker, M., Lamel, L. 2017. Discovering speech reductions across speaking styles and langages. In Cangemi, F., Clayards M., Niebuhr O., Schuppler B., & Zellers M. *Rethinking reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic variation*, De Gruyter Mouton. 101-128.
[4] Vasilescu, I., Vieru, B., Lamel, L. 2014. Exploring pronunciation variants for Romanian speech-to-text transcription. *Proceedings of SLTU-2014 – International Workshop on Spoken Language Technologies for Under-Resourced Languages.*161–168.
[5] Vasilescu, I., Hernandez, N., Vieru, B., Lamel, L. 2018. Exploring Temporal Reduction in Dialectal Spanish : A Large-scale Study of Lenition of Voiced Stops and Coda-s. *Proceedings of Interseech*, Hyderabad, India.
[6] Vasilescu, I., Chitoran, I., Vieru, B., Adda-Decker, M., Candea, M., Lamel, L., Niculescu, O. Accepted. Studying variation in Romanian : deletion of the definite article -l in continuous speech. *Linguistic Vanguard*.
[7] Adda-Decker, M., Snoeren, N. 2011. Quantifying temporal speech reduction in French using forced speech alignment. *Journal of Phonetics*, 3(39). 261–270.
[8] B. Schuppler, M. Ernestus, O. Scharenborg, Boves, L. 2011. Acoustic reduction in conversational Dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics* 3(39). 96-109.
[9] Adda-Decker, M., Boula de Mareüil, P., Lamel, L. 1999. Pronunciation variants in French: schwa and liaison. *Proceedings of ICPhS*, San Francisco. 2239-2242.
[10] Fougeron, C., Gendrot, C., Bürki, A. 2007. On the phonetic identity of French schwa compared to /ø/ and /oe/. *Actes des 5èmes Journées d'Études Linguistiques.* 191-198.
[11] Bürki, A., Ernestus, M., Fougeron, C., Gendrot, C., Frauenfelder, U. 2011. What affects the presence versus absence of schwa and its duration : A corpus analysis of French connected speech. *Journal of the Acoustical Society of America* 130 (6). 3980-3991.
[12] Jiahong, Y., Liberman, M. 2009. Investigating /l/ variation in English through forced alignment. In *Proceedings of Interspeech*, Brighton. 2215-2218.
[13] Renwick, M., Vasilescu, I., Dutrey, C., Lamel, L., Vieru, B. 2016. Marginal contrast among Romanian vowels: Evidence from ASR and functional load. *Proceedings of Interspeech*, San Francisco. 2433–2436.