Analyzing linguistic variation in a Romanian speech corpus through ASR errors

Variable phonological rules have been the object of extensive attention in Romance languages such as French and Spanish (Bürki et al., 2011; Meunier & Espesser 2011; Candea et al., 2013; Torreira et al., 2010; Torreira & Ernestus 2011, 2012). In contrast, such phenomena have received almost no attention in Romanian. In this paper we focus on two morpho-phonological phenomena in spoken Romanian, whose source of variation has not been empirically studied.

The first one concerns the variable deletion of the final lateral of the masculine definite article /-ul/: omul 'the man' is realized as [omul] or [omu]. Impressionistically, [l] deletion is very frequent in a casual, conversational style of speech. Romanian [l] is light in all contexts, so the variation cannot be attributed to misperception of a vocalized [l]. It is more likely that the production of [l] in the definite article ranges from complete absence to a very strong, consonantal [l]. We are studying the variable production of [l] between these two extremes and the frequency of its deletion in a large spoken corpus. The second phenomenon we study is final palatalization in plural nouns and adjectives (lupi 'wolves' realized as [lup^{j}] or [lup$^{(j)}$]), and second person singular verb forms (e.g. sapi 'you dig' realized as [sap^{j}] or [sap$^{(j)}$]. In this case we hypothesize that the palatalization is regularly and systematically produced, but it is only variably perceived, possibly as a function of the consonantal context (based on Spinu et al., 2012). The existence of this type of variability in spontaneous speech has occasionally been mentioned, but there has been no empirical work on its incidence in natural discourse or on the factors that may favor or disfavor its occurrence.

Phonetic variation of this kind cannot be elicited in laboratory settings. We therefore decided to study the extent to which it can be linked to errors produced by an automatic speech recognition (ASR) system, in an approach proposed by Adda-Decker & Lamel (1999). An automatic speech recognition (ASR) system is currently being developed for spoken Romanian (Vasilescu et al., 2014). We use the ASR system as a tool to detect instances of inter- and intra-speaker variation. In an ASR system, the training process of the acoustic model generates segmentations into words and on a subword level, into phone segments. The phonemic labeling resulting from the segmentation process depends on the acoustic model configuration, and on the pronunciation variants included in the dictionary. Such annotated corpora have been shown to be valuable resources for phonetic or more generally linguistic studies (Adda-Decker & Lamel, 1999; Vasilescu et al., 2012).

From a linguistic point of view, the errors allow us to identify the contexts in which they occur, and the acoustic manifestation of the variants. Ultimately, this essential first step will allow us to assess whether the observed variation is strictly contextually predictable, or tends to generalize, potentially leading to sound change (Ohala 1989; 1996).

The data consists of 3.5 hours of recorded broadcast news and debates, for which manual reference transcriptions are provided. These data were processed by the ASR system. The predicted variation was observed as ASR transcription errors, and the hypothesized cause was missing variants in the pronunciation lexicon. To test this hypothesis we authorized pronunciation variants according to linguistic hypotheses for each case. The results were (manually) verified by the authors: selected variants were analyzed in terms of: (i) the acoustic realization of the expected phenomenon and (ii) the system's performance as an indicator of the effective realization of the hypothesized variation. On the entire corpus the ASR performance resulted in 17% WER (word error rate) overall, which is relatively low.

For the definite article the authorized transcription variants are –ul and –u. The results for /-ul/ show that the system was able to robustly detect the final [l] whenever it was realized with a strong release burst, and rarely detected it when it was realized with approximant-like formant structure, as well as when it had been deleted according to our visual spectrographic analysis. That is, the ASR system marked as deleted both instances where there was no evidence of [l] in the acoustic signal and also when it was unreleased. This failure of the ASR system to detect unreleased final [l] may shed light on the origin of the deletion phenomenon (by listener misinterpretation of non-salient articulations, cf. Ohala 1996). Specifically, it allows us to test hypotheses on the phonetic environment of the [l] deletion.

We hypothesized that the definite article would be more likely absent and consequently not transcribed (i.e., no error) in the following contexts:

- before a C-initial word compared to a V-initial word. – True (67% vs. 13%, respectively)

- in spontaneous and casual connected speech (debates) more than in formal, prepared connected speech (broadcast news) – True (36.9% for debates vs. 20% for the entire corpus)

We hypothesized that a pronounced [l] is more likely to be missed (transcription error) before a C-initial word than a V-initial word – True (15.9% vs. 4.1%, respectively)

The acoustic analysis confirms two different realizations of [l], when it is pronounced: it is systematically detected when realized with a strong release burst, and rarely detected when realized with approximant-like formant structure. The low detection of [l] may be interpreted as a weakening of its morphological marking of definiteness. We propose that the marking of definiteness may be transferred to the preceding [u], which is historically a desinence vowel (Chitoran, 2001).

In the case of word-final palatalization, three transcription variants were authorized: C, Ci, Cj (the latter modeled by Russian Cj). The results so far indicate that the Romanian ASR system provided the following alignments: C (undetected palatalization) 45.5%; Ci (detected palatalization with Romanian model) 32.3%; Russian Cj (detected palatalization with Russian model) 20.2%. These initial results confirm that word-final palatalization has low perceptual salience, consistent with the results of human perception experiments by Spinu et al. (2012) with stimuli based on speech recorded in the laboratory.

We conclude that the study of phonetic variation starting from ASR system errors is an efficient first step in analyzing spoken corpora. It tests the robustness of the hypothesized variability, it can identify the contexts in which it occurs, and the main acoustic realizations. Based on this initial information, the data can further be used for more fine-grained acoustic analysis and perception testing.

References

Adda-Decker, M. & Lamel, L. 1999. Pronunciation variants across system configuration, language and speaking style. *Speech Communication* 29(2-4):83-98.

Bürki, A., C. Fougeron, C. Gendrot & U.H. Frauenfelder (2011). "Phonetic reduction versus phonological deletion of French schwa: Some methodological issues". *Journal of Phonetics* 39: 279-288

Candea, M., M. Adda-Decker & L. Lamel. 2013. Recent evolution of non-standard consonantal variants in French broadcast news. *Proceedings of Interspeech 2013*, Lyon, France, August 25-29, 2013, p. 412-416

Chitoran I. 2001. *The phonology of Romanian*. Mouton de Gruyter.

Meunier, C. & R. Espesser. 2011. Vowel reduction in conversational speech in French: The role of lexical factors. *Journal of Phonetics* 39:3, 271-278.

Ohala, J.J. 1989. Sound change is drawn from a pool of synchronic variation. In: Breivik, LE, Jahr, EH (eds.) *Language change: Contributions to the study of its cause*s. Mouton de Gruyter. 173-198.

Ohala, J.J. 1996. The connection between sound change and connected speech processes. *Arbeitsberichte (AIPUK 31)* Universität Kiel. 201-206.

Spinu, L., Vogel, I., Bunnell, H.T. 2012. Palatalization in Romanian – Acoustic properties and perception. *Journal of Phonetics*, (40)1: 54-66

Torreira, F. & M. Ernestus. 2011. Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology* 2: 331-353.

Torreira, F., & Ernestus, M. 2012. Weakening of intervocalic /s/ in the Nijmegen Corpus of Casual Spanish. *Phonetica*, 69. 124-148

Torreira, F., M. Adda-Decker, M. Ernestus. 2010. The Nijmegen Corpus of Casual French. *Speech Communication* 52. 201-212

Vasilescu, I., Adda-Decker, M., Lamel, L. 2012. Cross-lingual studies of ASR errors: paradigms for perceptual evaluations. *Proceedings of LREC*, Turkey 2012.

Vasilescu, I., B. Vieru, L. Lamel. 2014. Exploring pronunciation variants for Romanian speech-to-text transcription. Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced languages (SLTU'14) St. Petersburg, Russia, May 2014