

Anti-markedness patterns in French deletion and epenthesis: An information-theoretic account

It is widely observed that the quality of the segment that deletes or is epenthesized is unmarked (e.g. Rice 1999). French vowel deletion and epenthesis challenge this view, however, since the vowel in question is front rounded, widely considered to be universally marked (Chomsky & Halle 1968). In particular, a mid front rounded vowel (tense or lax, depending on factors such as speaker and region) commonly deletes (e.g. *tu devenais* 'you became' [tyd.vœ .nɛ]), and is used to break-up ill-formed clusters (e.g. *un contact penible* [œkʔtaktœpenibl]) (e.g., Noske 1993). From a markedness perspective, one would predict deletion or epenthesis to involve one of the front unrounded vowels such as [e] or [ɛ], which also occur in the French inventory, rather than their rounded counterpart.

In this talk, we suggest that the French pattern receives a straightforward account when viewed through the lens of *information theory* (Shannon 1948). In particular, the concepts of *information value* (or surprisal) and *entropy* allow for a deeper understanding of why the deleted and epenthesized sounds need not be universally unmarked. The calculation of a segment's contribution to the entropy of a system shows that the quality of the preferred epenthetic or deleted segment corresponds to the segment with the lowest information content. It thus contributes little to the entropy of the system.

But why should a low contribution to entropy matter for epenthesis and deletion? If we view language as a system shaped by linguistic evolution to meet the competing demands of robustness and efficiency in communication, it becomes clear that the least informative vowel can be reduced in the interest of efficiency, or inserted in the interests of robustness (e.g. to help disambiguate otherwise perceptually-masked cues) while minimally affecting the information content of the message. Moreover, reducing or epenthesizing a low information vowel can help to smooth the entropy rate (aka information density (Levy & Jaeger 2007) or redundancy profile (Aylett & Turk 2004)), further maximizing communicative effectiveness.

Using a corpus of formant data of French vowel productions (Gendrot & Adda-Decker 2005), we constructed an exemplar model based on Nosofsky (1986). The model provides information values for both acoustic similarity, frequency, and a combined value. Normalizing the data to account *only* for acoustic similarity yields lower information values for the mid front unrounded vowels [e] and [ɛ]. Note that [ɛ] is the least marked vowel of this set, hence predicted by traditional markedness-based accounts. However, combining the frequency and acoustic data consistently points to [œ] as the least informative vowel, with [ø] (the other commonly-epenthesized vowel in French) close behind. A model using token frequency alone predicts [a] as the epenthesized or deleted vowel.

To investigate the average contribution of each vowel to the entropy of a corpus of French (New et al., 2001), we calculated the average systemic change in entropy associated with merging a given vowel phoneme with each other vowel phoneme in turn. Whether measured at the lexical, triphone or segment levels, average entropy change was much lower for [ø, œ] than for [ɛ]. These results suggest that the unexpected preference in French for epenthesizing and deleting [ø, œ] over the less universally marked vowel [ɛ] can be explained by reference to information. All three vowels are predicted to be highly confusable with other vowels in the French vowel system, but our results suggest that epenthesis or deletion of [ø, œ] has a lower impact on the information density of an utterance.

These results support an approach in which markedness can be evaluated at the level of the individual language. Further, it allows us to predict the quality of the epenthetic/deleted vowel using a quantitative information-theoretic approach taking into account language specific frequency and the phonetic properties of the segments in question.

References

Aylett, M. & A. Turk (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence, and Duration in Spontaneous Speech. *Language and Speech*, 47(1). pp. 31-56.

Chomsky & Halle 1968. *The Sound Pattern of English*. New York: Harper and Row.

Gendrot, C. & M. Adda-Decker (2005). Impact of duration on F1/F2 formant values of oral vowels: an automatic analysis of large broadcast news corpora in French and German. In *Proceedings of Eurospeech*. Lisbon (Portugal), September 2005. pp. 2453-2456.

Levy, R. & T.F. Jaeger (2007). Speakers optimize information density through syntactic reduction. *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

New B., Pallier C., Ferrand L. & R. Matos (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101. pp. 447-462.
<http://www.lexique.org>

Nosofsky, R. (1986). Attention, Similarity and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, 115(1). pp. 39-57.

Noske, R. (1993). 1993. *A Theory of Syllabification and Segmental Alternation. With studies on the phonology of French, German, Tonkawa and Yawelmani*. Tübingen : Niemeyer (*Linguistische Arbeiten* Series nr. 296).

Rice, Keren. (1999). Featural markedness in phonology: variation. *Glott International* 4:7. 3-6, 4:8. 3-7.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. Vol. 27. pp. 379-423, 623-656.