

Prediction of liaison in French by measures of Information Theory

Thomas Culetto
University of Oxford
Centre for Linguistics and Philology, Walton Street, Oxford OX1 2HG, UK
thomas.culetto@ling-phil.ox.ac.uk

Although the general phonological context in which liaison occurs is well known, the interaction of other factors, mainly syntactic and stylistic, makes this phenomenon hard to predict with sufficient accuracy. Different domains have been proposed: syntactic (Selkirk 1974), prosodic (Selkirk 1986, Scarborough and Jun 2003) and lexical (Bybee 2001). Experimental results suggest that liaison is linked to syntactic and idiomatic levels (Boula de Mareüil 2003, Fougeron 2001b), but they do not identify a single factor. A probabilistic model combining different sources of uncertainty could account for such a variable phenomenon, especially the fuzzy area of optional liaison.

Using measures of Information Theory, I generalise Zelig Harris' approach of morphological segmentation (1955). My hypothesis is the following: in a text tagged using Parts of Speech, we consider a token; if it proves easy to predict the next token, entropy is low and liaison is more likely to occur. For instance, with a model of 14 syntactic categories the determiner is nearly always followed by a noun (80%), its entropy is low (0.9 bits) and liaison is compulsory: un enfant /*õnãfã*/. Conversely when the uncertainty (and entropy) is high, liaison is less likely to occur. For instance, conjunctions of coordination have high entropy (3.1 bits) and do not trigger compulsory liaison. Using a tagged corpus, I reached the conclusion that the categories with conditional entropy under a certain threshold were generally followed by an obligatory liaison. These categories include the determiner, the pronoun, the preposition and the conjunction of subordination. One of the advantages of such a model is that it could better account for optional liaison and sociolinguistic variation by raising or lowering the threshold where liaison occurs. However as this model suffers from shortcomings, further improvements might include calculating entropy of syntactic categories after lexical items.

References

- Adda-Decker, M., Boula de Mareüil, P. and Lamel, L. 1999. Pronunciation variants in French: schwa & liaison. *ICPhS*, 2239-2242, San Francisco.
- Boula de Mareüil, Ph., Adda-Decker M. and Gendner, V. 2003. Liaisons in French: a corpus-based study using morpho-syntactic information. *15th ICPhS*, 1329-1332, Barcelona.
- Bybee J., 2001. Frequency effects on French liaison. In J. Bybee and P. Hopper (eds). *Frequency and the emergence of linguistic structure*, 337-359. John Benjamins, Amsterdam.
- Fougeron C, Goldman JP and Frauenfelder U. 2001a. Liaison and schwa deletion in French: an effect of lexical frequency and competition. *Eurospeech conference*. Aalborg, Denmark.
- Fougeron C, Goldman JP, Dart A, Guelat L and Jeager C 2001b. Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. *8ème TALN*, Tours.
- Harris, Z., 1955. From phoneme to morpheme. *Language* 31, 190-222.
- Scarborough, R. & Jun S-A, 2003. Accentual Phrase and the domain of liaison in French. Poster presented at 15th ICPhS, Barcelona.
- Selkirk E., 1974. French liaison and the X-bar convention. *Linguistic Inquiry* 5, 573-590.
- Selkirk, E. 1986. On derived domains in sentence phonology. *Phonology Yearbook* 3, 371-405.