**Quantifying near-homophony induced by French liaison**

Victor Antoine (LSCP, ENS, PSL University, EHESS, CNRS, France)
Rory Turnbull (School of English Literature, Language and Linguistics, Newcastle U., UK)
Sharon Peperkamp (LSCP, ENS, PSL University, EHESS, CNRS, France)
chez.vantoine@gmail.com

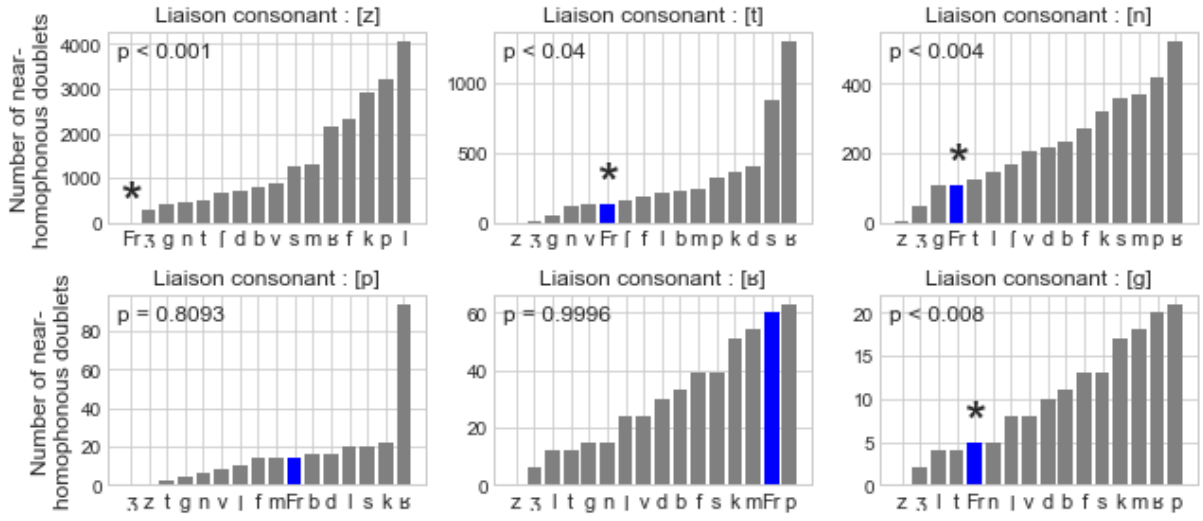
French has a number of words ending in an underlying so-called liaison consonant (one of /z,n,t,ʁ,p,g/), which is pronounced only before vowel-initial words (cf. *dernier chat* [dɜʁ.nje.ʃa] 'last cat' - *dernier achat* [dɜʁ.nje.ʁa.ʃa] 'last purchase'). Liaison can create near-homophony; for instance *dernier achat* [dɜʁ.nje.ʁa.ʃa] is pronounced like *dernier rachat* 'last repurchase'. Listeners are sensitive to the subtle acoustic cues distinguishing such pairs, yet they activate the unintended as well as the intended word (Spinelli et al., 2002). Hence, for cases like *dernier achat* liaison increases the difficulty of word segmentation. We examine whether French is structured in such a way as to minimize liaison-induced near-homophony.

We used the electronic dictionary Lexique (New et al., 2001) to (i) extract all words with an underlying liaison consonant e.g. *dernier* and *trop*, and (ii) all pairs of words differing only by the presence vs. absence of an initial consonant, e.g. *achat/rachat, uni/puni* 'united/punished'. We created doublets by combining liaison words with the relevant pairs, e.g {*dernier achat / dernier rachat*} and {*trop uni* / *trop puni*}, and filtered out all ungrammatical ones, such as {**dernier acheter* 'last to purchase' / **dernier racheter* 'last to repurchase'}. To investigate whether liaison consonants are the consonants that provide the lowest number of doublets, we substituted each liaison consonant with each of the other French consonants and applied the same procedure to obtain lists of near-homophonous doublets for the relevant liaison words. To illustrate, by replacing [ʁ] with, say, [f], we obtained new doublets for *dernier* (and the other words ending in liaison-[ʁ]), for instance {*dernier acteur* 'last actor' / *dernier facteur* 'last postman'}. For each true liaison consonant, we then compared the number of near-homophonous doublets to the number of such doublets for all the substitutes, using one-tailed one-sample Wilcoxon signed rank tests. We found that for four of the six liaison consonants, the number of near-homophonous doublets in real French is lower than the median of the number of such doublets obtained by the substitutions (Figure 1A). These four liaison consonants include /z/, and /t/, which are inflectional morphemes (plural of nouns and adjectives, and third person singular and plural for verbs, respectively) and which therefore are the ones that occur by far in the largest number of words. Thus, these results suggest that French is indeed structured such as to minimize liaison-induced near-homophony.

Previous work has shown that across languages, homophonous word pairs tend to involve words from different syntactic or semantic categories (Dautriche et al., 2018), suggesting that homophones are hardly detrimental for everyday speech comprehension. In a similar vein, we additionally examine the extent to which near-homophonous liaison doublets induce real-life processing difficulty. Using frequency data from Google Ngram Viewer (Michel et al., 2011), we computed two parameters, (i) the sum and (ii) the ratio, for each doublet (Formulas 1A and 1B), assuming that the real-life processing difficulty increases (i) the more the frequency of each of the members of the doublet increases and (ii) the closer the two members are in terms of frequency. We then extracted so-called troublesome doublets, i.e. those with a ratio and a sum among the 50% highest ones, and compared the number of troublesome doublets in French to the number of such doublets obtained with substitutes. We found that for three of the six liaison consonants, again including /z/ and /t/, the number of near-homophonous troublesome doublets in real French is significantly lower than the median of the number of such doublets obtained by the substitutions (Figure 1B).

We tentatively conclude that French is structured such as to avoid adding to the difficulty of word segmentation by near-homophonic doublets induced by liaison. In further research, we plan to use Latent Semantic Analysis to examine processing difficulty using a second real-life difficulty score, based on the semantic similarity of word pairs involved in doublets, such as *achat*/*rachat* 'purchase/repurchase' (high semantic similarity) and *uni*/*puni* 'united/punished' (low semantic similarity).

**A** - Number of near-homophonous doublets




**B** - Number of troublesome near-homophonous doublets

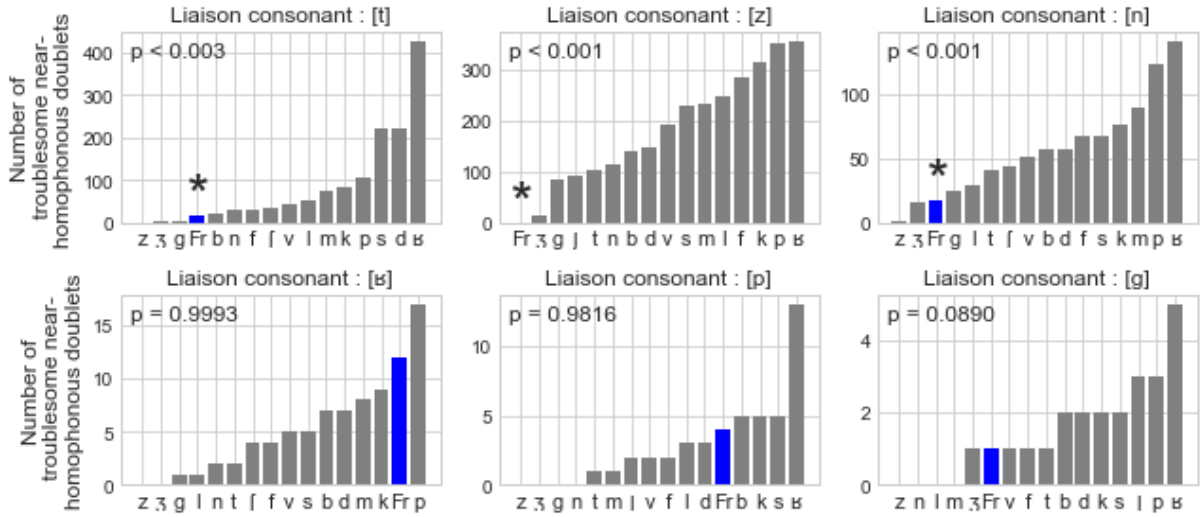



**Figure 1**: Number of near-homophonous doublets (A) and number of troublesome near-homophonous doublets (B) for each liaison consonant (subplots) and each substitution (individual bars). The scales of the y-axis are sorted from the largest ([z] and [t]) to the smallest ([g]). Blue bars represent data for real French – labeled 'Fr' – and grey bars represent data for the various substitutions. There is no blue bar for the subplots for [z] as there are no near-homophonous doublets for this liaison consonant in real French.

**References**


Dautriche, I., Fibla, L., Fievet, A. C., & Christophe, A. (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive psychology*, *104*, 83-105.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176-182.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516-524.

Spinelli, E., Cutle, A., & McQueen, J. M. (2002). Resolution of liaison for lexical access in French. *Revue française de linguistique appliquée*, *7*(1), 83-96.

**Formula 1**: Sum and Ratio

Let *a* and *b* be the frequency values of members of a doublet, then the doublet's sum and ratio are respectively defined as:

**A** – Sum

$$\text{Sum} = a + 1 + b + 1$$

**B** – Ratio

$$\text{Ratio} = \frac{min(a+1, b+1)}{max(a+1, b+1)}$$

The doublet's sum increases as *a* and *b* increase, and the doublet's ratio approaches 1 the closer *a* and *b* are i.e. the smaller the difference between their frequencies. By hypothesis, these parameters reflect the real-life processing difficulty experienced by listeners (i) for whom a frequent doublet implies a frequent segmentation difficulty and (ii) for whom selecting the intended word segmentation is most difficult for doublets whose members have a similar high frequency, such as {*est en* 'is in' / *est tant* 'is so'}, and least difficult for doublets in which one member has a very low and the other a very high frequency, such as {*un nerf* 'a nerve' / *un air* 'a tune'}.