Corpus analyses of French liaison

Supervised by Sharon Peperkamp & Rory Turnbull

Administrative information

Member of *conseil pédagogique* : Maria Giavazzi
External researcher : Isabelle Dautriche

Introduction

Background and rationale

Lexical segmentation is a complicated process. Unlike written language, oral language is a continuous flow of sounds that individuals must be able to segment in order to extract words from it. One of the great difficulties in this process arises from homophony, i.e. when several words are supported by the same sequence of sounds. French becomes a special case because there exists the phonological phenomenon of *liaison*. Keeping it simple, it consists in the pronunciation of an underlying final consonant (/z/, /n/, /t/, /ʁ/, /p/, /g/) when the next word is vowel-initial - e.g. *petit arbre* 'small tree' pronunced [pə.ti.**t**aʁbʁ] and not [pə.ti.aʁbʁ]. Due to resyllabification - the liaison consonant becoming the onset of the following word - liaison can therefore create cases of homophony. For example *petit ami* 'boyfriend' and *petit tamis* 'little sieve' are both pronunced [pə.ti.ta.mi]. In sum, in addition to the usual instances of homophony within[5] and through[6] words, French provides opportunities for further homophony.

In confusing environments - where several segmentations are possible (e.g. « tulips » can also be « two lips ») - it has been shown (i) that participants also activate the form not intended by the speaker and (ii) that these situations are processed more slowly than control situations (Christophe et al., 2020; Gow Jr & Gordon, 1995; Spinelli et al., 2003; Tabossi et al., 1995). Liaison constitutes therefore an additional difficulty for French speakers by adding new confusing situations.

The lexicons of languages are structured in such a way as to reduce cases of homophony. For instance, Dautriche et al (2018) have shown that homophones tend to be distributed syntactically - i.e. they belong to different syntactic categories - and semantically - i.e. they refer to widely distinct concepts - in the lexicon of several Indo-European languages, including French. As liaison brings an additional difficulty comparable to that of homophony, it is then possible to imagine that French lexicon could be structured in such a way as to reduce this difficulty.

Key research question

Is the French lexicon structured in such a way as to reduce cases of homophony created by liaison: i.e. do liaison consonants provide less homophony than other consonants?

General hypotheses

French provides a set of liaison consonants that helps to minimize the number of confusing sequences (i.e. with other liaison consonants, there would be more homophony).

Methods

Material

We will use the LEXIQUE 3.83 database (New et al., 2004; New et al., 2001) which contains 142,694 French language entries from literature and film subtitles. A cleaning process will be necessary to remove all entries that are not relevant to our study (for instance onomatopoeias).

5 [pɛ̃] can be *pin* ‘pine’ or *pain* ‘bread’.
6 *Le chat laid* ‘the ugly cat’ [ləʃalɛ] can also be *le chalet* ‘the cottage’ [ləʃalɛ].

Measures

We want to extract from LEXIQUE (i) words containing an underlying liaison consonant (e.g. *petit* 'small') and (ii) word pairs that differ by a liaison consonant (e.g. *ami* 'friend' / *tamis* 'sieve'). Sequences will be created (e.g. *petit ami* 'little friend' / *petit tamis* 'little sieve'). We will sort them to determine which ones are ambiguous or not. For example, there are only three words in French that trigger a liaison with [ʁ]: *dernier* 'last', *léger* 'light' and *premier* 'first'. The sequence [*dernier achat* 'last purchase (noun)' / *dernier rachat* 'last repurchase (noun)'] is ambiguous but not [**dernier achètent* 'last purchase (verb)'/ **dernier rachètent* 'last repurchase (verb)'] which is simply impossible in French. We will get a number of ambiguous sequences for our liaison consonant. We will apply the same selection procedure with the other consonants to get a number of ambiguous sequences if the French liaison consonant was changed. For example, noun pairs can be ambiguous after *dernier*, *premier* and *léger*: the sequence [*dernier achat* / *dernier rachat*] is ambiguous in 'real' French, as the sequence [*dernier acteur* 'last actor' / *dernier facteur* 'last postman'] if we assume that the liaison consonant [ʁ] becomes [f]. To summarize :

- Extraction of liaison words (*dernier*, *premier*, etc) and word pairs (*achat*/*rachat*, *acteur*/*facteur*, etc)

For each liaison consonant:

- Setting a filter to select ambiguous sequences (e.g. for [ʁ]: noun pairs but not verb pairs)
- Application of this filter on the sequences to obtain
  - the number of ambiguous sequences in French (*dernier achat*/ *dernier rachat*)
  - but also the number of ambiguous sequences with other consonants (e.g. for [ʁ] becoming [f]: *dernier acteur*/ *dernier facteur*)

Predictions

The prediction is that French liaison consonants provide a significantly lower number of ambiguous sequences than when these consonants are substituted by others.

Analyses

(i) Number of ambiguous sequences with liaison consonants vs. substituted consonants.

One-Sample Wilcoxon Signed Rank Tests will be performed to analyze the data (since count data are unlikely to be normally distributed, non-parametric tests are preferred). For each liaison consonant, we will use the number of ambiguous sequences for that liaison consonant as a theoretical value. The number of ambiguous sequences obtained when this consonant is changed will constitute the dataset to be compared to this theoretical value. We will therefore perform a one-tailed test:

- H_0: number of ambiguous sequences (other consonants) = number of ambiguous sequences of 'real' French (liaison consonant)
- H_1: number of ambiguous sequences (other consonants) > number of ambiguous sequences of 'real' French (liaison consonant)

For example, if we find X ambiguous sequences for the liaison consonant [ʁ] and {Y, Z, W, ...} ambiguous sequences when [ʁ] is changed to {[t], [p], [b] ....} : we will compare the set {Y, Z, W, ...} to our real French value X. If the lexicon is indeed organized to minimize homophony, we expect the median of {Y, Z, W, ...} to be significantly higher than X. Of course, we will perform this analysis for all the liaison consonants in French i.e. [p], [g], [z], [n], [t] and [ʁ].

(ii) Further detailed analysis by adding ambiguity severity factors?
The inclusion of several factors could allow a more precise analysis of the ambiguity of word sequences (precise implementation still under discussion- the data presented for this part in the thesis will thus be exploratory data):

- **Frequency of words:** are *être* and *paître* frequent at all?
- **Frequency of word sequences**: are *être* and *paître* frequently with *trop*?
- **Semantic relatedness**: can the sequence *trop être* 'to be too much' leave the possibility open to interpret it as *trop paître* 'to overgraze' ?

Interpretation

If the number of ambiguous sequences is indeed significantly higher when liaison consonants are substituted, then we will reject our hypothesis H_0. Otherwise, we will not reject our hypothesis H_0.

Expected contributions

(Names are given in alphabetical order and not by amount of contribution).

- Project design : Victor Antoine, Sharon Peperkamp, Rory Turnbull.
- Extraction of required data from LEXIQUE (scripting) : Victor Antoine.
- Statistical analyses : Victor Antoine ; and guided by Sharon Peperkamp and Rory Turnbull.
- Report Writing : Victor Antoine ; and commented by Sharon Peperkamp and Rory Turnbull.

References

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. Journal of Memory and Language, 51(4), 523-547.

Gow Jr, D. W., & Gordon, P. C. (1995). Lexical and prelexical influences on word segmentation: Evidence from priming. Journal of Experimental Psychology: Human perception and performance, 21(2), 344.

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.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary french: LEXIQUE™. L'année psychologique, 101(3), 447-462.

Spinelli, E., McQueen, J. M., & Cutler, A. (2003). Processing resyllabified words in French. Journal of memory and language, 48(2), 233-254.

Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. Journal of Memory and Language, 34(4), 440-467.