Master of Cognitive Science

Ecole Normale Supérieure / PSL, Université de Paris, EHESS

2020 - 2021

Quantifying near-homophony induced by French liaison

Victor Antoine

Supervised by Rory Turnbull & Sharon Peperkamp

Laboratoire de Sciences Cognitives et Psycholinguistique

Team Language and its acquisition

Ecole Normale Supérieure, 75005 Paris

Rory Turnbull
(School of English Literature, Language and Linguistics, Newcastle University., United Kingdom)

Sharon Peperkamp
(LSCP, École Normale Supérieure, PSL University, EHESS, CNRS, France)

Acknowledgements

I am extremely grateful to both my supervisors, Dr. Sharon Peperkamp and Dr. Rory Turnbull, for their guidance and valuable advice. I could not have asked for a better internship experience, particularly under these unusual circumstances.

I especially want to express my sincere appreciation to Sharon Peperkamp for the past two years - always helpful, supportive and knowledgeable.

I would like to thank all the people at LSCP who kindly welcomed and helped me, despite the remote working conditions.

Je voudrais remercier mes amis et ma famille qui sont toujours disponibles, attentionnés, et disposés à m'aider malgré le peu de temps libre que j'ai pu leur accorder ces dernières années.

Je voudrais remercier mes parents qui ont brillamment relevé le défi qu'est d'élever trois enfants : je n'ai jamais manqué de rien, sûrement pas d'amour. Merci. Sincèrement.

Je voudrais finalement remercier Anna, pour toutes les raisons qu'elle connaît.

Dziękuję Ci bardzo.

Summary

Liaison in French consists of the pronunciation of an otherwise silent underlying word-final consonant when the following word is vowel-inital e.g. *dernier chat* 'last cat' [dɛʁ.nje.ʃa] but *dernier achat* 'last purchase' [dɛʁ.nje.**ʁ**a.ʃa]. Liaison consonants can lead to near-homophonous sequences e.g. *dernier rachat* 'last repurchase' is pronouced as *dernier achat*. Such sequences raise difficulties as speakers activate both forms - but must be able to determine which one is intended - and thus are processed more slowly than non-homophonous ones (Spinelli et al., 2003). We investigated whether French is structured in such a way as to reduce liaison-induced near-homophonous environments. Using the corpus Lexique, we extracted (i) all words with an underlying liaison consonant e.g. *dernier* and (ii) all pairs of words differing only by the presence vs. absence of an initial consonant e.g. *achat / rachat*. We combined them to create confusing doublets e.g. {*dernier achat / dernier rachat*} – keeping only grammatical ones. As we were interested in whether liaison consonants were the ones that produced the least number of doublets, we devised substitutions of liaison consonants. Words were identical except for their liaison consonant - e.g. *dernier* would have [f] and no longer [ʁ] as liaison consonant – and thus giving new confusing doublets – e.g. {*dernier acteur* 'last actor' */ dernier facteur* 'last factor'}. We found a significantly lower number of confusing doublets for four of the six liaison consonants. Post-hoc frequency analyses were run to test whether these doublets were troublesome in everyday speech. We selected so-called troublesome doublets i.e. (i) those which the speaker experienced frequently and (ii) which were difficult to segment. We found a significantly lower number of troublesome doublets for three of the six liaison consonants. We claim that these results suggest that French is indeed structured such as to minimize liaison-induced near-homophony.

Keywords : Liaison; Homophones; Lexicon; Lexical segmentation; French

Declaration of Novelty

Previous studies have highlighted the inherent difficulties of homophone environments in which two lexical segmentations are available to the speaker. French liaison generates similar situations and shows similar segmentation difficulties. Other studies have noticed that languages are structured in such a way as to reduce the difficulties provided by homophones. However, no study has investigated the lexical organization of French for liaison which creates near-homophonous environments. In this study, we determine whether other liaison consonants would have been better to reduce liaison-induced homophony. We also begin a frequency analysis and suggest future research possibilities.

Declaration of Contribution

Victor Antoine, Rory Turnbull, and Sharon Peperkamp jointly defined the research question and the project design. Rory Turnbull and Sharon Peperkamp regularly provided the relevant literature to address the research question. Victor Antoine coded the scripts needed to extract the data from the electronic dictionary Lexique. Victor Antoine performed the statistical analyses under the supervision of Rory Turnbull and Sharon Peperkamp. Victor Antoine interpreted the results under the guidance of Rory Turnbull and Sharon Peperkamp. Victor Antoine wrote this report and produced the tables and figures, guided by comments from Rory Turnbull and Sharon Peperkamp.

Pre-registration document - submitted on April 22, 2021

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' pronounced [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 pronounced [pə.ti.ta.mi]. In sum, in addition to the usual instances of homophony within[1] and through[2] 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).

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

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).

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.

Contents

1. Introduction

Lexical segmentation is a complex process. Indeed, oral language input constitutes a continuous acoustic flow - in which words are not clearly delimited - unlike written language. Speakers must be able to extract words, and this challenge can be illustrated as follows. [paʁɛgzɑ̃pl] is an oral sequence in French but no delimitation between words is indicated and the possibilities are very numerous – [paʁ]+[ɛg]+[zɑ̃pl] or maybe [paʁɛg]+[zɑ̃pl], etc. Yet this statement simply means *par exemple* 'for example' in French. Speakers are constantly confronted with this phenomenon. This is made even more complex by the fact that resyllabification is frequent in French e.g. the utterance *par exemple* 'for example' is pronounced [pa.**ʁ**ɛg.zɑ̃pl] and not [pa**ʁ**.ɛg.zɑ̃pl]. The syllabic structure perceived by speakers does not therefore provide a crucial aid to segmentation. Moreover, as in other languages, many words are embedded thereby adding difficulties to the segmentation process - *carapace* 'shell' contains *car* 'because', *rapace* 'hawk', *passe* 'pass'. This embedding can also be found within phrases.

French is a special case due to the phonological phenomenon of liaison. It occurs when a latent consonant surfaces between two words (word_1 and word_2 respectively); word_1 ending with an otherwise silent liaison consonant and word_2 beginning with a vowel or, in some cases, a glide. For example, *grand* 'tall' ends with the underlying liaison consonant /t/. This consonant will not be pronounced in an utterance-final position, e.g. *il est grand* 'he is tall' [i.lɛ.gʁɑ̃] or if the following word is consonant-initial, e.g. *le grand chat* 'the tall cat' [lə.gʁɑ̃.ʃa]. When the following word is vowel-initial, [t] will surface e.g. *le grand arbre* 'the big tree' [lə.gʁɑ̃.**t**aʁbʁ]. Liaison consonants are often reflected in the written form (Table 1) but no precise rule defines whether a word contains a liaison consonant. Both *ces* 'these' and *intrus* 'intruder' end with a written *s* – which is a written form of the liaison consonant [z] - but *intrus* does not have an underlying liaison consonant. When adding a vowel-initial word, one pronounces *l'intrus adroit* 'the skillful intruder' [lɛ̃.tʁy.a.dʁwa] without uttering a [z], but *ces amis* 'these friends' [sɛ.**z**a.mi] with liaison of the [z] of *ces*.

Liaison is said to be *enchaînée* 'chained' because it leads to resyllabification. The latent consonant of word_1 becomes the onset of the first syllable of word_2 (*grand arbre* 'big tree' pronounced [gʁɑ̃.**t**aʁbʁ] and not [gʁɑ̃**t**.aʁbʁ]). This particularity brings an additional difficulty. Succession of 2 vowels within a word is rare in French. Out of 12544 lemmas available in the lexical database Lexique 3.8.3 (New et al., 2004; New et al., 2001), only 231 have a succession of 2 vowels and one has a succession of 3 vowels (*brouhaha* 'hubbub' [bʁu.a.a]), i.e. 1.8% of lemmas. Therefore, if two vowels succeed each other, it is very likely that they belong to two different words. By adding this liaison consonant between the final vowel of word_1 and the initial vowel of

word_2, liaison removes the original information of these two successive vowels. Another additional difficulty is the fact that liaison creates syllabically identical homophonous forms - *grand ami* 'tall friend' [gʁɑ̃.tami] could also be *grand tamis* 'large sieve'. Thus, even if the speaker correctly extracts the meaning of *grand* from the acoustic flow, the difficulty of knowing which sequence is the intended one will persist - does [tami] refer to 'sieve' or is it [ami] 'friend' with a liaison consonant [t] by the end of [gʁɑ̃]? The present project is precisely interested in the quantification of this homophony induced by liaison.

The remainder of this introduction will first focus on the phonological, syntactic and lexical characteristics of liaison (section 1.1). Then, experimental data on speech segmentation in unclear contexts will be reviewed (section 1.2) before setting out the details of the present study (section 1.3).

Liaison consonant	Written letter	Examples		
[z]	{s, z, x}	nous sommes	[nu.sɔm]	'we are'
		nou**s** avons	[nu.**z**a.vɔ̃]	'we have'
[n]	{n}	un chien	[ɛ̃.ʃjɛ̃]	'a dog'
		u**n** ami	[ɛ̃.**n**a.mi]	'a friend'
[t]	{d, t}	un grand piano	[ɛ̃.gʁɑ̃.pja.no]	'a big piano'
		un gran**d** arbre	[ɛ̃.gʁɑ̃.**t**aʁbʁ]	'a big tree'
[ʁ]	{r}	le dernier pacte	[lə.dɛʁ.nje.pakt]	'the last pact'
		le dernie**r** acte	[lə.dɛʁ.nje.**ʁ**akt]	'the last act'
[p]	{p}	trop cordial	[tʁo.koʁ.djal]	'too cordial'
		tro**p** aimable	[tʁo.**p**ɛ.mabl]	'too kind'
[g]	{g}	un long dimanche	[ɛ̃.lɔ̃.di.mɑ̃ʃ]	'a long sunday'
		un lon**g** hiver	[ɛ̃.lɔ̃.**g**i.vɛʁ]	'a long winter'

Table 1 : Liaison consonants

1.1 Liaison and its characteristics

Liaison is a variable process that appears to depend on many factors. First, liaison consonants are restricted to a limited number (Table 1). Words containing these liaison consonants are diverse: pre-nominal adjectives (e.g. *ancien* 'old' with the liaison consonant [n]), determiners (e.g. *un* 'a' with [n]), prepositions (e.g. *dans* 'in' with [z]), verbs (e.g. *peut* 'can (3sg)' with [t]), adverbs (e.g. *trop* 'too' with [p]), inflectional morpheme of plural (e.g. *amis* 'friends' with [z]), etc. As there is no precise rule determining whether a word contains a liaison consonant, speakers learn them only through the linguistic input they receive. Liaison is generally applied if word_2 is vowel-

initial, and some glide-initial words also trigger liaison. For example, *le grand ouest* 'the big west' is pronounced [lə.gʁɑ̃.**t**wɛst], with liaison of the [t] of *grand,* but *le grand ouistiti* 'the big marmoset' is pronounced [lə.gʁɑ̃.wis.ti.ti], with no liaison. Some vowel-initial words similarly do not trigger liaison. Many of these words are spelled with a written *h* and are referred to as having an *h aspiré* 'aspirated h', as the *h* « blocks » the liaison from occurring. For example, *le grand hameau* 'the big hamlet' is pronounced as [lə.gʁɑ̃.a.mo], with no liaison. There are also words with an *h-aspiré* which are not spelled with a written *h*, such as *un* 'number one' e.g. *le grand un* 'the big number one' [lə.gʁɑ̃.ɛ̃]. However, many words spelled with *h* do in fact trigger liaison, such as *le grand habit* 'the big suit' [lə.gʁɑ̃.**t**a.bi] and some words have variable *h-aspiré* e.g. *les haricots* 'the beans' that one can hear [le.a.ʁi.ko] or [le.**z**a.ʁi.ko].

The main restriction that applies is probably the one at the phonological level. Liaison seems to apply only within a phonological phrase (Selkirk, 1974). For example, in the sequence *le grand arbre grandit* 'the big tree grows' [lə.gʁɑ̃.**t**aʁbʁ.gʁɑ̃.di], the liaison consonant surfaces - which is no longer the case in *le grand aime le chocolat* 'the tall one likes chocolate' [lə.gʁɑ̃.ɛm.lə.ʃo.ko.la]. The second sequence does not have a liaison consonant because *le grand* and *aime* belong to different phonological phrases - the separation between nominal and verbal phrases inducing a strong phonological boundary.

A second restriction applies at the syntactic level. Some grammatical contexts induce a high frequency of liaison - so-called obligatory cases - while others induce a more inconsistent realization - so-called optional cases (Armstrong, 2001; Durand & Lyche, 2008). For example, a compulsory case is that between a determinant and a noun - *un arbre* 'a tree' [ɛ̃.**n**aʁbʁ] - and an optional case is that between a plural noun and its adjective - *animaux heureux* 'happy animals' [a.ni.mo.**z**ø.rø]. To give an idea, De Mareüil et al. (2003) showed using corpus of read newspaper speech that compulsory liaison rates vary between 70 - 95% and optional liaison rates vary between 14 - 44%. Meinschaefer et al.(2015) recorded in a corpus of spontaneously spoken language an overall optional liaison rate of 27%. It is also possible to find some cases of hypercorrection, i.e. sequences in which the speaker produces a liaison consonant in an unexpected case – e.g. *soldat anglais* 'English soldier' pronounced [sɔl.da.**t**ɑ̃.glɛ] while *soldat* does not contain any liaison consonant. The difference made between these different contexts is important but it seems relevant to insist on the fact that the majority of liaison occurrences are limited to a resticted number of cases. For example, De Mareüil et al (2003) showed in their corpus of read newspaper speech that 35.6% of liaisons observed followed the rule [determiner + word_2] – e.g. *un ami* 'a friend' [ɛ̃.**n**a.mi] - and 13.6% followed the rule [auxiliary + word_2] – e.g. *il est évident* 'it is obvious' [i.le.**t**e.vi.dɑ̃]. These two rules alone therefore explained 50% of the liaisons observed in their

corpus. Meinschaefer et al (2015) found in their oral corpus that 46.9% of liaisons were between a clitic pronoun and a verb – e.g. *peut-on* 'can we' [pø.**t**õ] – and 19.5% were between a prenominal determiner or modifier and a following noun – e.g. *ses amis* 'his/her friends' [se.**z**a.mi].

A third restriction concerns the lexical level. Empirically, corpus analyses show that some syntactic contexts are limited to a few lexical forms. Barreca & Christodoulides (2017) showed that the liaison between an adjective and a noun is mainly produced with *bon* 'good', *autres* 'others', *petit* 'small (masc. sing)', *petit(e)s* 'small (plur)', and *grand* 'big'. A frequency factor is probably involved as some contexts apparently concern only a few very common lexical forms. Adda-Decker et al (1999) found that the 1% most frequent word tokens in their corpus explained more than half of the liaison contexts.

1.2 Lexical segmentation in confusing environments

Liaison can create homophonous sequences due to the resyllabification of word_2 with the liaison consonant. As explained above, *grand ami* 'tall friend' and *grand tamis* 'large sieve' are pronounced identically [gʁɑ̃.ta.mi]. It then becomes necessary for speakers to be able to determine the correct form in such sequences. Previous studies have shown, among other things, that there is indeed a form of uncertainty in such cases.

In lexical segmentation, speakers must be able to extract lexical forms from a continuous flow of sounds. This segmentation is partly done sequentially, i.e. with each new sound heard, speakers will try to determine the most probable form. For example, if speakers hear [ʒa], they could expect *jardin* 'garden' [ʒaʁ.dɛ̃], *j'arrive* 'I am coming' [ʒa.ʁiv] or also *jamais* 'never' [ʒa.mɛ]. If they then hear [ʁ] giving [ʒaʁ], the option *jamais* would no longer be available and there would remain *jardin* or *j'arrive*. The purpose of this example is to show that, as time goes by, some lexical forms will remain in competition while others will be eliminated, and that until only one remains. This type of question was notably addressed in the 1980s with the elaboration of different models (Dahan & Magnuson, 2008; for a review). The segmentation process is getting worse when it is impossible to complete it because several forms remain available - as in our example *grand ami ~ grand tamis*. How is it then possible to determine the right word sequence?

Several studies tend to show that so-called "ambiguous" inputs - i.e. a flow of sounds with several coherent word sequences - induce the activation of close forms and that these are handled more difficultly than control environments. It is possible, for example, to refer to the study by Gow & Gordon (1995) using a lexical priming technique. The principle was to make participants listen to ambiguous sentences and ask them to indicate whether the word that appeared on the screen was a word or a nonword. In this type of experiment, the lexical decision is expected to be faster if the

previous stimulus has "primed" its activation. In their experiment, the authors used ambiguous sequences such as "*She tried to put her two lips on his cheek* [...]" which can also be segmented with 'tulips' and not 'two lips'. Their results showed that lexical decision on FLOWER is enhanced, even when the subject heard the sentence with 'two lips' suggesting that both lexical forms 'lips' and 'tulips' are activated during such sequences. Tabossi et al (1995) showed similar results in their study using the same paradigm. Participants listened to the sequence "*Le circostanze rendevano inevitabili visite* [...]" 'Circumstances made visits inevitable [...]' of which *visite* 'visits' can also be *visi te* 'faces' plus the beginning of another word *tediadi* 'bored'. The results showed an increased lexical decision on PARENTI 'relatives' whether participants hear *visite* or *visi te* suggesting the idea that even embedded sequences are processed by participants. Crucially, the authors showed in a subsequent experiment that lexical decisions on a related word were faster after hearing unambiguous sequences than after hearing ambiguous sequences. The decision on ANNOIATI 'bored' after the unambiguous sequence *è tediati* 'is bored' was faster than the decision on ANNOIATI 'bored' after *visi tediati* 'faces bored', which is ambiguous as it can be also segmented with *visite* 'visits'. The idea is that when sequences are ambiguous, all forms are temporarily activated by the participant and therefore it slows down the process - as these decisions would be faster in unambiguous cases. These data thus tend to suggest that homophonous sequences created by liaison would also be a source of difficulties for lexical segmentation in French.

Christophe et al (2004) are also interested in such sequences and their relationship to phonological structure. The design is different: a word to be searched is first displayed on the screen - CHAT 'cat' - and then a sentence is read. Participants are then asked to indicate every time they hear the word they are looking for. The authors used ambiguous sequences such as [ɛ̃.ʃa.gʁɛ̃ ...] which can be segmented as *un chat grincheux* 'a grumpy cat' [ɛ̃.ʃa.gʁɛ̃.ʃø] or *un chagrin* 'a grief' [ɛ ɛ̃.ʃa.gʁɛ̃]. Conversely, the sequence [ɛ̃.ʃa.dʁo ...] can only be segmented as *un chat drogué* 'a drugged cat' [ɛ̃.ʃa.dʁo.gé]. The results showed that ambiguous sequences - [ɛ̃.ʃa.gʁɛ̃ ...] - are processed more slowly than non-ambiguous sequences - [ɛ̃.ʃa.dʁo …] - and this despite the existence of fine acoustic cues. An important fact is that this difference is no longer found when the sequence appears across two different phonological phrases. So, when the sequence [ʃa.gʁɛ̃ ...] is included in *son grand chat grimpait* 'his big cat was climbing' [sɔ̃.gʁɑ̃.ʃa.gʁɛ̃.pɛ], this slowing down is no longer found suggesting that lexical research is constrained by phonological factors. The authors identified subtle acoustic features that could explain these lexical access constraints. In terms of duration, they observed a lengthening of phrase-final vowels ([a] in *son grand chat grimpait* 'his big cat was climbing') and in terms of pitch, they noticed an elevation across word boundaries of a given phonological phrase along with a decrease across a phonological phrase

boundary.

Therefore, speakers seem to be sensitive to fine acoustic differences that could help them distinguish possible segmentations. For cases implying liaison, however, clear findings are difficult to draw from studies that specifically addressed this issue. In a phoneme detection task - the principle being for participants to indicate whenever they hear a particular phoneme within sentences - Nguyen et al. (2007) found that participants have more difficulties to detect the liaison consonants [z] and [n] - e.g. *j'ai remis des écrous* 'I put some nuts back' [ʒɛ.ʁə.mi.de.**z**e.kʁu] - than [z] and [n] as initial word_2-consonants - *il y a des zéros* 'there are zeros' [i.lja.de.ze.ʁo]. Despite this processing difference, the authors did not find a significant difference in duration for target consonants ([z] in the above example) or preceding vowels ([e] in the above example) in the liaison cases compared to those with initial word_2-consonants. This result was not consistent from previous findings (Spinelli et al., 2003). They found acoustic differences between sequences with vs. without liaison, with liaison consonants being shorter than regular consonants - but the analysed liaison consonants were not the same, namely. [p], [ʁ], [t] and [n]. Yet, in their cross-modal priming experiments - participants listened to a sentence which can be segmented in two ways due to liaison, and then had to make a lexical decision on a word displayed on the screen – for both forms, response times were shorter than those obtained with an unambiguous sequence, indicating activation of both simultaneously. For example, after hearing the sentence which has a liaison *c'est le dernier oignon* 'this is the last onion' [sɛ.lə.dɛʁ.nje.**ʁ**o.ɲɔ̃], participants were faster to make lexical decisions on OIGNON 'onion' but also on ROGNON 'kidney' which corresponds to the near-homophonous sequence without liaison *c'est le dernier rognon* 'this is the last kidney' [sɛ.lə.dɛʁ.nje.ʁo.ɲɔ̃]. Specifically, results for the unintended form constituted an intermediate case between those for the intended form and those for the unambiguous baseline form i.e. the processing advantage was greater for OIGNON than ROGNON in *c'est le dernier oignon*. Difficulties thus seem to be induced by liaison but precise acoustic cues might explain why listeners can correctly segment such sequences.

Overall, by inducing environments that can be segmented in two ways, liaison introduces additional lexical segmentation difficulties in French to those typically found in languages. Several studies showed that such sequences are more difficult to process than non-homophonous ones. However, there exist fine acoustic cues that help listeners to retrieve the form intended by the speaker. So, rather than ambiguous, we will refer to such sequences as confusing ones.

1.3 The present study

Liaison in French - through creating confusing environments by generating near-homophonous cases such as *grand ami ~ grand tamis* - constitutes an additional difficulty that speakers must manage. Dautriche et al. (2018) showed through the study of corpora from several languages including French that homophones were distributed in a way that minimized their processing difficulties - particularly useful for their acquisition by children. Through comparison to randomly generated lexicons, the authors showed that homophones in natural languages tend to belong to different syntactic and semantic categories. Due to the similarity of the confusing environments induced by liaison in French to those obtained with regular homophones, it is possible to hypothesize that French is also structured to reduce liaison-induced homophony. The present project is precisely focused on quantifying this homophony by investigating whether French liaison consonants are those that induce a minimal number of confusing environments.

Let us refer to cases like *grand ami ~ grand tamis* as a confusing doublet - given that there are two possible segmentation alternatives for the speaker. We were interested in investigating whether French liaison consonants ([g], [n], [p], [t], [z], [ʁ]) are those that provide less confusing doublets than the other French consonants. To quantify the number of confusing doublets due to liaison, we needed (i) all the words with an underlying liaison consonant (e.g. *grand* 'tall') - further reported as word_1 - and (ii) all pairs of words that differ by the presence of a liaison consonant in onset position (e.g. *ami* 'friend' / t*amis* 'sieve') - further reported as minimal pairs of words_2. Doublets were created by combining words_1 and minimal pairs of words_2 - e.g. {*grand ami* 'little friend' / *grand tamis* 'little sieve'}. These doublets were sorted to remove those that are ungrammatical. For example, *dernier* 'last' (adjective) can trigger a liaison with [ʁ]. The doublet {*dernier achat* 'last purchase' / *dernier rachat* 'last repurchase'} can be confusing but not {**dernier achètent* 'last purchase (3pl)'/ **dernier rachètent* 'last repurchase (3pl)'} which is simply impossible in French. Constraints to identify which minimal pair can follow a word_1 were set up – e.g. *dernier* (adjective) can be followed by nouns but not verbs. This provided a number of confusing doublets for 'real' French.

As we aimed to compare this number to the numbers of confusing doublets obtained if the liaison consonants were different, we created alternative versions of French in which the words with liaison consonants were the same but their consonant was substituted. For example, we generated the alternative version of French where the liaison consonant [ʁ] was replaced by [f]. The word *dernier* could still trigger liaison but its liaison consonant would be pronounced [f] and not [ʁ]. We then applied the same constraints than those of real French to select confusing doublets - in our case that *dernier* can only be followed by nouns. Therefore, in this alternative version of French, the

doublet {*dernier acteur* 'last actor' / *dernier facteur* 'last postman'} was confusing which was not the case in real French. The overall aim was to imagine all possible substitutions: for each liaison consonant we created all the possible alternative versions of French. We then compared the results obtained in real French to those obtained with these substitutions - with the hypothesis that liaison consonants in (real) French provide fewer confusing doublets than other consonants.

Frequency post-hoc analyses were performed to investigate whether confusing doublets were marginally detrimental in everyday speech because of low frequencies. We assumed that a doublet is 'troublesome' if it occurs frequently i.e. if one or both members have a high number of occurrences. This reflects the listener's difficulty: if at least one member is frequent, then the problem is frequently encountered. For example, there exists the doublet {*ont eu* 'have (3pl) had' / *ont tu* 'have (3pl) concealed'} for which *ont eu* is highly frequent; hence the listener will often be confronted with this segmentation problem. In addition, we assumed that a doublet is 'troublesome' if the two members have similar frequencies even if these are not particularly high. This reflects the listener's choice between the two possible alternatives: if one is much more frequent than the other, then the former is very likely to be the right one - and thus the segmentation process is quite easy. For example, the frequency value of *petit ami* 'boyfriend' is much larger than that of *petit tamis* 'little sieve'; hence [pə.ti.ta.mi] is more likely to be *petit ami* than *petit tamis*. After retrieving these troublesome doublets, we compared the number of troublesome confusing doublets in French to those obtained when liaison consonants were substituted; with the hypothesis that liaison consonants in (real) French provide fewer troublesome confusing doublets than other consonants.

2. Pre-registered analysis - Number of confusing doublets

2.1 Material

The electronic dictionary Lexique (New et al., 2004; New et al., 2001) was used to extract doublets that may be confusing due to liaison. It contains 142,694 French entries and provides frequency data extracted from literature and film subtitles. We decided not to use frequency data from literature as our study is focused on oral language. A screening process of this dictionary was performed in order to avoid ungrammatical doublets. The following steps were performed:

1. A threshold of one was set for the frequency in film subtitles. It corresponds to keeping words found at least one time per million of occurrences in the subtitles corpus - for a total of 22,633 words.
2. Letter names and Roman numerals were removed as they do not contain any liaison consonant and cannot induce liaison as word_2.
3. Onomatopoeias were removed e.g. *beurk* 'yuck'

4. The words *au* 'at the', *aux* 'at the (plural)', *de* 'of the' and *du* 'from', initially coded as articles, were recoded into prepositions[1].

2.2 Words_1

Words with an underlying liaison consonant were extracted from Lexique. Decisions were made manually – thus subjectively - among words ending with a liaison consonant in its written form (i) but with a vowel in its phonological form e.g. *petit* [pə.ti] 'small' or (ii) with a consonant different from that liaison consonant e.g. *pleurent* [plœʁ] 'cry (3pl)'. Judgments were made with the help of two dictionaries (Littré, 1873-1874; TLFi, 2004) and one book (Dubroca, 1824) which provide information on earlier pronunciations. The following choices were made when retrieving words_1 :

- Words that can only trigger liaison in highly specific contexts were not retained.

  e.g. *accent* 'accent' triggers liaison in the fixed expression *accent aigu* [ak.sɑ̃.**t**e.gy] 'acute accent' whereas liaison will not be made elsewhere – *accent abject* [ak.sɑ̃.ab.ʒɛkt] 'abject accent'.

- Liaison production seems to depend on sociolinguistic factors and in particular on the context of occurrence with more liaisons produced in a formal context (Armstrong, 2001; Delattre, 1966; Malécot, 1975). To avoid considering rare and highly formal utterances:

  - Words that can only trigger liaison in highly formal contexts were not retained.

  e.g. *court* 'short' could trigger liaison in a formal speech *un court espace* [ɛ̃.kuʁ.**t**ɛs.pas] 'a small space' but not in a standard speech – [ɛ̃.ku.ʁɛs.pas].

  - Infinitives ending with *r* in the written form were not retained.

  e.g. *ériger* 'build', although one could pronounce a liaison in *ériger un palais* [e.ʁi.ʒe.**ʁ**ɛ̃.pa.lɛ] 'build a palace' for example.

  - Adverbs ending with *-ment* in the written form were not retained.

  e.g. *suffisamment* 'enough', although one could pronounce a liaison in *il est suffisamment aiguisé* [i.lɛ.sy.fi.za.mɑ̃.**t**ɛ.gi.ze] 'it is sharp enough' for example.

1 We will define which liaison patterns are possible by using grammatical constraints e.g. by stating that *grand*, an adjective that contains the liaison consonant [t], can be followed by nouns as in *grand ami*. *Au*, *de* and *du* do not contain any liaison consonant and are hence only of interest as words_2. Therefore, they must be coded accordingly to what may precede them. For example, for *au* which is the combination of *à* plus *le*, we were interested in knowing what can precede *à* which is a preposition. Following the same rationale for *de* and *du,* we coded them as prepositions. In order to propose a coherent analysis, we coded *aux* as a preposition although this word contains the liaison consonant [z]. It was not an issue as it was the only preposition with the liaison consonant [z]. Hence, we were able to establish what could follow *aux* (i.e. what could follow the determiner *les* since *aux* is *à* plus *les)* without conflicting other words_1.

Liaison consonant	Number of words_1	Examples
[g]	1	*long* 'long'
[p]	2	*trop* 'too', *beaucoup* 'a lot'
[ʁ]	3	*dernier* 'last', *premier* 'first', *léger* 'light'
[n]	17	*bon* 'good', *mon* 'my', *son* 'its', *un* 'a', ...
[t]	1419	*petit* 'small', *tout* 'all', *quand* 'when', *est* 'is (3sg)', ...
[z]	3217	*des* 'some', *anciens* 'old (masc. pl)', *jolies* 'pretty(fem. pl)', ...

Table 2 : Number of words_1 for each liaison consonant

2.3 Minimal pairs of words_2

Minimal pairs were pairs of words that differ by the presence vs. absence of a consonant in onset position (e.g. *ami* [a.mi] 'friend' / t*amis* [ta.mi] 'sieve'). Minimal pairs were extracted automatically from Lexique. When retrieving them, words beginning with an *h aspiré* were removed as they prevent liaison. For example, *un heaume* 'a helmet' is pronounced [ɛ̃.om] and not [ɛ̃.nom]. A screening process was then performed on these minimal pairs in order to control their acceptability. Several words were problematic due to the reasons listed below. Minimal pairs containing one of these words were removed.

- The coder does not know the word or no definition of the word was found (n = 52).
  e.g. unknown word *ante* (fem. noun - mostly in plural form)
  « prominent pillar built on the face of a wall »
  e.g. no definition *pars* as a noun
- The context of use of the word is narrow (n = 18).
  e.g. *for* (masc. noun) mostly used in *for intérieur* 'inner self'
- The word is a combination of other words (n = 5).
  e.g. *l'un* which is *le* 'the' plus *un* 'one'
- The word is an abbreviation (n = 2).
  e.g. *min* (fem. noun) for *minute* 'minute'

Consonant	Number of minimal pairs of words_2	Example[1]
[z]	6	*aile* 'wing' / *zèle* 'zeal'
[ʒ]	96	*elle* 'she' / *gel* 'frost'
[g]	179	*ardent* 'ardent' / *gardant* 'keeping'

1 The examples given are various to show the diversity of the cases, but all these minimal pairs will not be retained to form future doublets like {*petit ami* / *petit tamis*}, as some of them do not induce liaison cases. For example, *emporte* / *remporte* differs by the liaison consonant [ʁ] found only in the adjectives *dernier*, *léger* and *premier*, but no verb can follow these adjectives to create a liaison - a sequence combining {*léger* or *premier* or *dernier}* plus *{emporte* or *remporte}* is simply impossible in French. Hence this type of minimal pairs will not be retained for future confusing doublets.

Consonant	Number of minimal pairs of words_2	Example
[ʃ]	213	*armé* 'armed' / *charmé* 'charmed'
[n]	256	*avait* 'had (3sg)' / *navet* 'turnip'
[b]	318	*ail* 'garlic' / *bail* 'lease'
[l]	320	*arme* 'weapon' / *larme* 'tear'
[d]	353	*émission* 'program' / *démission* 'resignation'
[k]	364	*âne* 'donkey' / *canne* 'cane'
[t]	367	*entend* 'hears (3sg)' / *tentant* 'tempting'
[v]	398	*écus* 'heraldic shields' / *vécu* 'lived'
[f]	437	*orge* 'barley' / *forge* 'forge'
[m]	439	*aile* 'wing' /*mêlent* 'mingle (3pl)'
[p]	487	*aimant* 'loving' / *paiement* 'payment'
[s]	625	*antenne* 'antenna' / *centaine* 'one hundred'
[ʁ]	842	*emporte* 'takes away (3sg)' / *remporte* 'wins (3sg)'

Table 3 : Number of minimal pairs of words_2 for each consonant.

2.4 Creation and sorting of doublets

Doublets were created by combining words_1 with minimal pairs of words_2 – e.g. combining *petit* with *ami/tamis* giving {*petit ami* / *petit tamis*}. Most confusing doublets involve words from the same grammatical categories e.g. {*petit ami* / *petit tamis*} but it is possible that they belong to different ones e.g. {*quand on* 'when we' / *quand ton* 'when your'}. However, a given minimal pair cannot always induce a confusing doublet with a word_1 due to grammatical constraints. For example, words_1 for the liaison consonant **[ʁ]** - i.e. three adjectives *premier* 'first', *dernier* 'last', *léger* 'light' - can only be followed by masculine singular nouns. The minimal pair of nouns *achat* / *rachat* will therefore create confusing doublets but not the minimal pair of verbs *achètent* / *rachètent*. Grammatical constraints were thus set up to identify the minimal pairs that could follow a given word_1. Applying this procedure provided a number of confusing doublets for 'real' French - e.g. for the liaison consonant **[ʁ]** by identifying all confusing doublets made up of *premier*, *dernier* or *léger* along with a minimal pair of masculine singular nouns differing by **[ʁ]** on onset position.

To investigate whether these liaison consonants in French are those that induce the least number of confusing doublets, we created alternative versions of French by substituting a given liaison consonant with another consonant from French. We imagined that words_1 were the same except for their underlying liaison consonant. For **[ʁ]**, the words *premier, dernier* and *léger* would no longer have an **[ʁ]** but another French consonant – e.g. [f]. Confusing doublets were thus composed of the same words_1 but other minimal pairs of words_2. We applied the same selection

constraints to obtain a number of confusing doublets for these 'alternative' versions of French. When **[ʁ]** became [f], we only selected the pairs of masculine singular nouns as in 'real' French. Confusing doublets were thus still composed of *premier, dernier* and *léger* but with minimal pairs of masculine singular nouns differing by [f] in onset position. The doublet {*dernier acteur* 'last actor' / *dernier facteur* 'last postman'} was therefore confusing.

All selection constraints used are available in Appendix A. The following choices were made when establishing them:

- Liaison with conjunctions was not examined as it is restricted to highly formal speech.

  e.g. *Le premie**r** et le dernier auront des récompenses.*

  [lə.pʁə.mje.**ʁ**e.lə.dɛʁ.nje.o.ʁɔ̃.de.ʁe.kɔ̃.pɑ̃s]

  'The first and the last ones will have rewards.'

- All past participles were masculine singular.[1]

2.5 Results

Numbers of confusing doublets for each liaison consonant and each substition are shown in Figure 1. One-sample Wilcoxon signed rank tests were performed. For each liaison consonant, we used the number of confusing doublets for that liaison consonant as a theoretical value - i.e. the number of confusing doublets in 'real' French. The number of confusing doublets obtained for each substitution – i.e. in 'alternative' versions of French - constituted the dataset to be compared to this theoretical value.

For example, we found 5 confusing doublets for the liaison consonant [g] and {4, 21, 11, ...} ambiguous doublets when [ʁ] was subtituted to {[t], [p], [b] ....}. We thus compared the set {4, 21, 11, ...} to the 'real' French value 5. We performed this analysis for all the liaison consonants in French i.e. [p], [g], [z], [n], [t] and [ʁ]. As we expected to find fewer confusing doublets in 'real'

1 In French, past participles agree in gender and number only in specific cases: the passive voice (*Les souris ont été mang**ées** par le chat*. 'The mice were eaten by the cat.') ; with the verbs that use the verb *être* 'to be' (*Mes grands-mères sont mort**es** ce matin*. 'My grandmothers died this morning.') ; with *verbes pronominaux* 'reflexive verbs' (*Elles se sont regard**ées***. 'They looked at each other.') and when the *complément d'objet* 'object complement' precedes the verb. (*Les souris, je les ai mang**ées***. 'The mice, I ate them.'). In other cases, the past participles are masculine singular.
Lexique differentiates past participles e.g. *habillé* 'dressed' in *il a habillé* 'he has dressed' from an adjective derived from a past participle e.g. *habillé* in *il est habillé* 'he is dressed'. For example, there is the minimal pair *habillé* 'dressed' / *rhabillé* 'redressed'. Consider the case where *rhabillé* is a past participle: if *habillé* is also a past participle, then both alternatives can be in the same syntactic position (*il a habillé* / *il a rhabillé*) but if *habillé* is an adjective, this will not be possible due to the use of different auxiliaries (*il **est** habillé* / *il **a** rhabillé*). The main exceptions are the few past participles working with the verb *être* 'to be' which agree in gender and number. These past participles can be in the same syntactic position as an adjective e.g. the past participle *allées* 'gone (f.pl)' in the same position as the adjective *tâlées* 'bruised (f.pl)' in {*elles sont allées* 'they went' / *elles sont tâlées* 'they are bruised'}. In order to avoid manual sorting of this type of doublets, we decided to consider only masculine singular past participles allowing one to avoid many impossible sequences such as (*il **est** habillé* / *il **a** rhabillé*) at the expense of a few sequences such as {*elles sont allées* / *elles sont tâlées*}.

French than in 'alternative' versions of French, we performed one-tailed tests with the following hypotheses:

- H_0: Median of confusing doublets 'alternative' French = N of confusing doublets 'real' French
- H_1: Median of confusing doublets 'alternative' French > N of confusing doublets 'real' French

Results of these tests are given in the subplots of Figure 1.

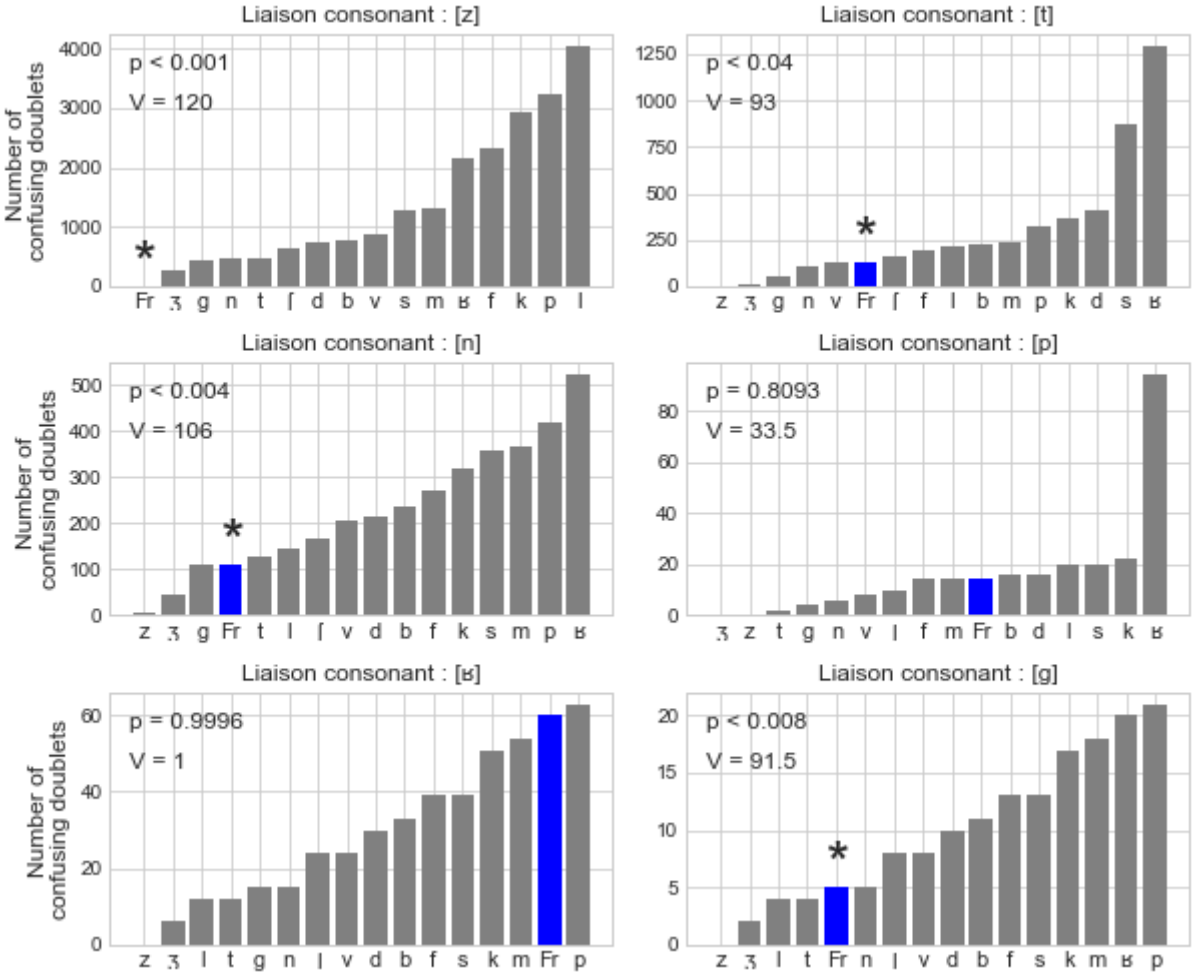

Figure 1: Number of confusing doublets for each liaison consonant (subplots) and each substitution (individual bars). Subplots are ordered from largest ([z]) to smallest ([g]) y-axis. Blue bars represent data for French – labelled 'Fr' - and grey bars represent data for each substitution. There is no blue bar for the subplot for [z] as there are no confusing doublets. For each liaison consonant, the V-values and p-values are displayed in the associated subplot. V is the sum of positive ranks, obtained from the one-sample Wilcoxon signed rank test. An asterisk indicates a significant result (α = 0.05).

As can be seen from the figure, significant results were obtained for four of the six liaison consonants ([z], [t], [n], [g]). By rejection of the null hypothesis, this means that for these four liaison consonants, the median of the number of confusing doublets obtained with the alternative versions of French, when the liaison consonant was replaced by another consonant, is significantly higher than the number of confusing doublets of (real) French . This is not the case for the liaison

consonants [p] and [ʁ] whose median obtained by the substitutions does not differ from that of (real) French.

3. Exploratory analysis - Frequency data

Given the novelty of this project, anticipating the precise design we were going to follow was a challenging process. We had described the main protocol regarding the raw number of confusing doublets in the preregistration document available at the beginning of this thesis; but had no clear idea of how long it would take to complete. We had also registered that the inclusion of other parameters such as the frequencies of the members of the doublets and their semantic similarity would provide a more accurate and complete analysis. Unfortunately, due to time constraints, we could not include all the parameters we would have liked to. We provide in this section an analysis of the frequencies of the members of the doublets extracted in the previous section.

The frequency values of the members of a doublet could be a key factor. It is conceivable that potentially confusing doublets bring only few difficulties since some members could simply be very rare. For example, for the doublet {*petit ami* / *petit tamis*}, a French speaker will activate both forms during word segmentation (Spinelli et al., 2002; Spinelli et al., 2003) but the fact that *tamis* is rare and hence that *petit tamis* is also rare reduces the difficulty in many cases. In contrast, other doublets are much more problematic such as {*est en* 'is in' / *est tant* 'is so ...'} for which both members are frequent - and it is thus more difficult to dismiss one of the two alternatives. As the aim was to retrieve doublets that could trouble listeners in everyday speech, we decided to assume that such doublets would be (i) those that listeners frequently experience and (ii) where lexical segmentation is not straightforward. We will refer to these 'real-life annoying' doublets as 'troublesome' doublets. To investigate whether confusing doublets in French induce few real-life difficulties due to the frequencies of their members, we extracted so-called troublesome doublets and compared the number of these doublets with the number of troublesome doublets obtained with the substitutions.

3.1 Material

We used the doublets collected in the previous analysis (part 2) - thus using the Lexique dictionary following the same steps. Google Ngram Viewer (Lin et al., 2012; Michel et al., 2011) was used to retrieve frequency data. It provides the number of occurrences of a Ngram - e.g. *petit ami* is a bigram (two words) - over the years. We used the latest corpus in French, including books up to 2019 and ran our queries for data from 1950 onwards - thus obtaining frequency data over the

period 1950 – 2019. Queries were case-insensitive and if the first letter of the word_1 had an accent, an additional query was performed without this diacritic as it is frequently removed in books when the letter is uppercased at the beginning of a sentence. We used the grammatical categories in the queries in order to obtain the most accurate data (in Google Ngram Viewer terminology: part-of-speech tags). The grammatical categories given by Lexique for each word were manually converted to match those of Google Ngram Viewer as closely as possible. Nevertheless, some doublets could not be analyzed for the following reasons:

- Google Ngram Viewer does not differentiate French *un* 'one', the numeral, from *un* 'a', the determiner. We used the part-of-speech tag 'determiner'.
- Google Ngram Viewer does not differentiate auxiliaries from verbs. We used the part-of-speech tag 'verb'.
- Google Ngram Viewer does not allow one to run queries with part-of-speech tags if a word contains a hyphen[1]. We ran these queries without part-of-speech tags.

These cases ultimately represent a very small proportion as we obtained frequency data for 30689 doublets out of the 31165 extracted in the previous section, i.e. a loss of 0.015%.

3.2 Extraction of troublesome doublets

Let *x y* and *x z* be members of a doublet {*x y* / *x z*} and *a* and *b* their respective frequency values. For each confusing doublet {*x y* / *x z*}, we computed two parameters: the doublet ratio [*min(a, b) / max(a, b)*] and the doublet sum [*a* + *b*]. We performed Laplace smoothing on the frequency values *a* and *b* - i.e. adding 1 to *a* and *b* - to deal with cases where one or both members of the doublet had zero frequency values. We assumed that the 'real-life' difficulty of a doublet increases (i) if at least one member is frequent, as it makes encountering a confusing sequence more likely and (ii) the closer the two members are in terms of frequency, as it makes that neither of the alternatives can be easily dismissed. To select the troublesome doublets, we decided to retain only doublets that had a doublet ratio and a doublet sum above the 50th percentile (from the doublet ratio and doublet sum distributions of our 30689 doublets). It means that so-called troublesome doublets are those that are frequently encountered - doublet sum in the top 50% (sum of more than 323) - and whose members have close frequency values - doublet ratio in the top 50% (ratio of more than

1 The character '-' is an operator in Google Ngram Viewer i.e. the query *x-y* will basically ask Google Ngram to display the frequency of x minus the frequency of y. It is possible to force Google to consider '-' as a hyphen by adding square brackets to the query i.e. [*x-y*]. Unfortunately it is impossible to add part of speech tags in a query with square brackets.

0.17)). This process allowed us to obtain a number of troublesome doublets for each liaison consonant and each substitution.

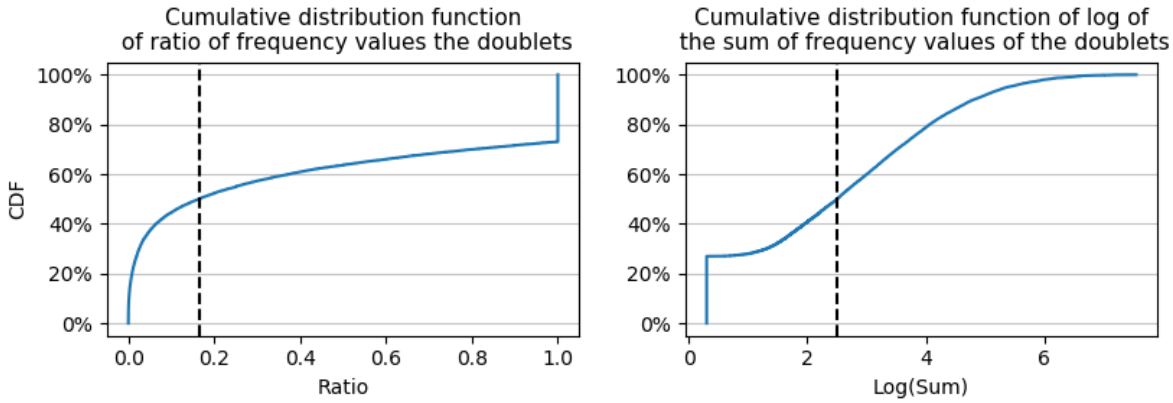

Figure 2: Cumulative distribution function of (i) the ratio of the frequency values of the doublets; and (ii) the decimal logarithm of the sum of the frequency values of the doublets. The black dashed line delineates the 50th percentile value (0.17 for the ratio and 2.51 for the log(sum) - which is equal to 323 for the raw sum). The plot for the sum uses the decimal logarithm to easily visualize the distribution of the data; which is highly skewed when using raw data. The difficult doublets are those with ratio and log(sum) values to the right of theses dashed lines.

3.3 Results

Numbers of troublesome doublets for each liaison consonant and each substition are shown in Figure 3. The statistical procedure was the same as for the previous part regarding the raw number of confusing doublets. One-sample Wilcoxon signed rank tests were performed. For each liaison consonant, we used the number of troublesome doublets for that liaison consonant as a theoretical value - i.e. the number of troublesome doublets in 'real' French. The number of troublesome doublets obtained for each substitution – i.e. in 'alternative' versions of French - constituted the dataset to be compared to this theoretical value. As we expected to find fewer troublesome doublets in 'real' French than in 'alternative' versions of French, we performed one-tailed tests with the following hypotheses:

- H_0: Median of troublesome doublets 'alternative' French = N of troublesome doublets 'real' French
- H_1: Median of troublesome doublets 'alternative' French > N of troublesome doublets 'real' French

Results of these tests are given in the subplots of Figure 3.

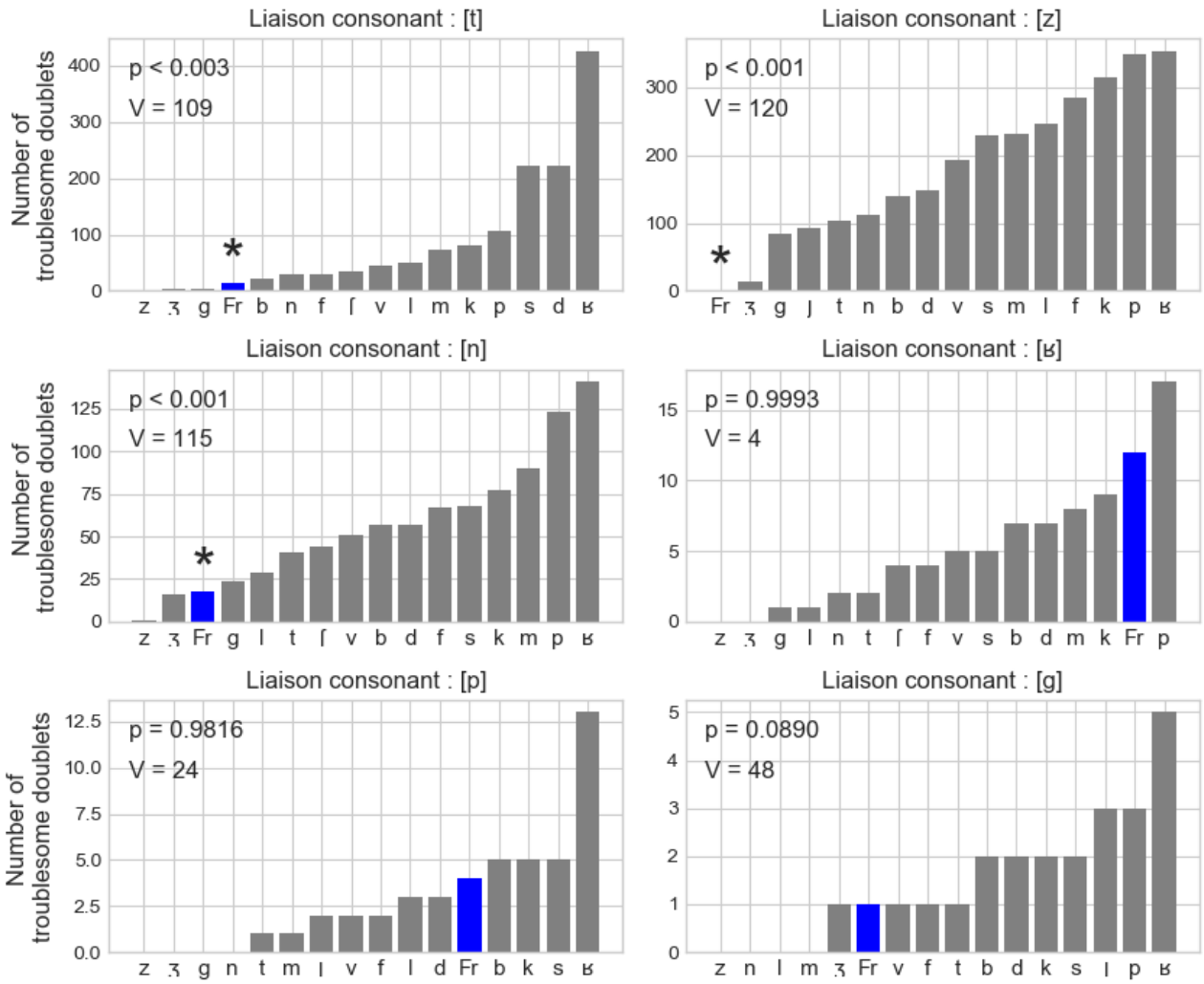

Figure 3: Number of troublesome doublets for each liaison consonant (subplots) and each substitution (individual bars). Subplots are ordered from largest ([t]) to smallest ([g]) y-axis. Blue bars represent data for French – labelled 'Fr' - and grey bars represent data for each substitution. There is no blue bar for the subplot for [z] as there are no troublesome doublets. For each liaison consonant, the V-values and p-values are displayed in the associated subplot. V is the sum of positive ranks, obtained from the one-sample Wilcoxon signed rank test. An asterisk indicates a significant result (α = 0.05).

As can be seen from the figure, significant results were obtained for three of the six liaison consonants ([z], [t], [n]). By rejection of the null hypothesis, this means that for these three liaison consonants, the median of the number of troublesome confusing doublets obtained with the alternative versions of French, when the liaison consonant was replaced by another consonant, is significantly higher than the number of troublesome doublets of (real) French. This is not the case for the liaison consonants [p], [ʁ] and [g] whose median obtained by the substitutions does not differ from that of (real) French.

4. Discussion

4.1 Main results

Regarding the number of confusing doublets, we obtained significant results for four of the six liaison consonants ([g], [n], [t], [z]). By rejection of the null hypothesis, this means that for these four liaison consonants, the median number of confusing doublets with substitutions is significantly higher than the number of confusing doublets in French. This suggests that the use of these four consonants limits excessive homophony induced by liaison. Regarding the exploratory analysis of the number of so-called troublesome doublets due to their frequency values, we obtained significant results for three of the six liaison consonants ([n], [t], [z]). By rejection of the null hypothesis, this means that for these three liaison consonants, the median number of troublesome doublets with substitutions is significantly higher than the number of troublesome doublets in French. The use of these three consonants limits the number of confusing environments that are difficult to process.

Two key points should be highlighted from these results. First, significant results were obtained in our two analyses for the three liaison consonants [n], [t], and [z]. These consonants cover the vast majority of words_1; namely 4652 tokens of words_1 for only 6 tokens of words_1 for the remaining liaison consonants ([ʁ], [p], [g]). Hence, controlling the occurrences of confusing sequences seems to be broadly implemented in the liaison system. Second, except for the liaison consonant [g], for which we obtained significant results only in the anlaysis of the raw number of confusing doublets, the same liaison consonants yield significant results. It is relevant as other methods to limit liaison-induced near-homophony could have been possible. Some consonants could have been controlled based on the number of homophonous doublets they generate while others could have been controlled via a specific frequency organization. It seems here that the liaison system is organized in such a way that some consonants are controlled, but not necessarily others as they initially lead to very few confusing environments.

On a broader level, it is worth recalling that our method of generating alternative versions of French relied on (i) a modification of the lexicon since words_1 no longer ended with their native liaison consonant and (ii) a modification of the liaison rule as it involved a new consonant with each substitution. For example, when the liaison consonant [ʁ] was replaced by [f], we modified (i) words_1 containing the liaison consonant [ʁ] by substituting [ʁ] with [f] and (ii) modified the liaison rule by accepting its realization with the new consonant [f] and removing the one with [ʁ]. The liaison, and consequently our analysis, is thus a special case in the literature as it depends both on the French lexicon (words_1 and words_2) and on the grammar of French (liaison rules). We have therefore provided evidence through this project that - at least for consonants with the most words_1 - the interaction between the French lexicon and French grammar is such that it avoids

lexical segmentation difficulties due to liaison-induced near-homophony. However, these results are not entirely clear-cut, and it is not easy to provide a precise interpretation. The fact is that for four (analysis of the raw number of doublets) and three (frequency analysis) of the six liaison consonants, we obtained results suggesting an avoidance of the homophony induced by liaison. Since this avoidance does not concern the entire set of liaison consonants, it remains hard to claim that French is fully structured to avoid liaison-induced homophony.

4.2 Limitations

There are several limitations to our study. First, the retrieval of words_1 containing liaison consonants relied on the intuitions of a single native speaker of French. This process was subjective, and some words were excluded because they implied liaisons supposedly made in narrow or formal contexts. Other speakers might have different intuitions given the variability of liaison production (Fougeron et al., 2001; Mallet, 2008; Meinschaefer et al., 2015). The inclusion of multiple speakers in such decisions would be relevant. Second, the establishment of constraints to select grammatical doublets from the electronic dictionary Lexique is mainly based on grammatical criteria (Appendix A) - e.g. removing the doublet *dernier achète* 'last purchases' / *dernier rachète* 'last repurchases' as verbs will not trigger liaison if they follow the adjective *dernier* 'last' due to a phonological boundary. This procedure is not perfectly accurate due to distributional factors at the lexical level - *léger ennui* 'slight boredom' is fully acceptable in French but *léger habit* 'light clothing' is less so, whereas both *ennui* and *habit* are nouns and can theoretically follow the adjective *léger*. Hence, on one hand, the analysis includes doublets that are not fully acceptable; and on the other hand, the analysis excludes doublets that should have been included. A solution to this limitation remains complicated to address since offering a fully accurate analysis appears unfeasible in a limited time. Finally, the post-hoc frequency analysis is built on the selection of so-called troublesome doublets based on the ratio and the sum of their frequency values. Building a composite frequency difficulty measure would be better - instead of splitting ratio and sum - but it remains a modeling challenge given the widely different distributions of those parameters. This analysis also relies on the assumption that 'real-life' difficulty of a doublet increases (i) the more frequently the doublet is encountered and (ii) the closer the frequency values of the members of the doublet. However, it may not be the frequency parameters that best capture the actual difficulty for speakers - an interaction between the two would for instance be possible.

4.3 Broader interpreation

The results of this project provide crucial new data and paths for further research. French

seems to be structured in such a way as to limit liaison-induced homophony, but not completely. These results are consistent with previous studies that support our difficulty of interpretation. Kaplan (2011) was interested in whether eight neutralizing rules of Korean induce a limited number of homophones. A neutralizing rule is a rule that triggers alterations such that differences between usually distinct segments disappear. The author gives the example of pre-consonantal neutralization of obstruents in Korean when followed by the morpheme /-k'wa/ 'and'. For instance, the forms /nɑtʃ-k'wɑ/ 'day and' and /nɑtʃʰ-k'wɑ/ 'face and' become identical; both being pronounced [nɑt.k'wɑ]. The difference in the underlying forms is no longer found in surface utterances; this type of rule thus generates homophonous sequences. By randomly generating rules close to those in Korean, she showed that most Korean rules do indeed generate few homophones compared to these randomly generated similar rules. However, one exception was the resyllabification rule in which a single consonant coda is resyllabified as the onset of the next syllable if it starts with a vowel or a [h]. For this rule - which is close to what we observe in French - the homophony in Korean is much higher than what is generated by similar rules. It is possible to assume that lexicons of languages are indeed structured to avoid homophony but that some phonological rules are still difficult to restrain.

Dautriche et al. (2018) supported this idea of homophony avoidance by finding that homophones in four Indo-European languages - including French - tend to belong to different syntactic and semantic categories - which would make learning and using homophones easier. However, members of our doublets frequently belong to the same syntactic categories as the words_2 of a minimal pair must be able to follow the same word_1. This observation goes against that of Dautriche et al. which suggests the existence of a dimension beyond a lexical one in the liaison-induced homophony, or even that our doublets do not correspond to regular homophones. As we did not perform a semantic analysis on the doublets for this thesis due to time constraints, it thus seems relevant to examine this dimension. It could be the case that the members of our doublets designate concepts very distant from each other compared to those of the alternative versions of French, hence following the observation of Dautriche et al. (2018) - or they could refer to relatively close concepts. In both cases, this semantic analysis appears relevant to gain more insights about the status of liaison-induced near-homophony in French. Spinelli et al (2003) showed through cross-modal priming experiments - the participant heard a short sentence and had to indicate whether the word displayed on the screen was a French word or not - that the two possible forms were indeed activated with liaison. However, the authors also reported subtle acoustic differences with shorter (liaison) consonants when the intended word is vowel-initial ([ʁ] shorter in *dernier oignon* [dɛʁ.nje.**ʁ**o.ɲɔ̃] 'last onion' than *dernier rognon* [dɛʁ.nje.ʁo.ɲɔ̃] 'last kidney'). Homophony induced by liaison in French is perhaps less problematic than regular homophones. This could explain the

fact that not all liaison consonants lead to a significantly lower number of doublets than other consonants.

4.3 Further research

Many paths for further research emerge from this project. First, it would be possible to complete our analysis with a selection of confusing doublets using grammatical constraints on what follows words_2. Currently, the selection constraints are based on what precedes the doublet. The doublet {*est au* 'is at the' / *est tôt* is early} is considered confusing as listeners can, for example, hear it in the sequence *il est au*/*tôt* [i.lɜ.to] - which is confusing as long as they do not hear the remainder of the acoustic sequence. Indeed, segmentation is quickly straightforward as *au* will be followed by a noun or adjective (*il est au marché* 'he is at the market'; *il est au dernier cours* 'he is at the last class'), which is unacceptable with *tôt* (**il est tôt marché*; **il est au tôt dernier cours*). Adding such grammatical constraints would provide more realistic data regarding the amount of liaison-induced near-homophony. Then, it seems relevant to use Latent Semantic Analysis to examine whether the members of French doublets have a lower semantic similarity than those of doublets obtained with substitutions. If so, liaison-induced near-homophony in French would be closer to 'regular' homophones as observed by Dautriche et al. (2018). Moreover, it is worth recalling the significant results in both analyses (raw number and frequency) for the liaison consonants [z], [t] and [n] which correspond to highly common utterances - plural of nouns and adjectives; 3rd person singular and plural of verbs; determiner *un* 'a' respectively. Restraining homophony in these cases seems largely relevant. Ultimately, the three remaining liaison consonants constitute only six tokens of word_1 (*trop* 'too', *beaucoup* 'much', *dernier* 'last', *premier* 'first', *léger* 'light' and *long* 'long'). An analysis of a possible trade-off between reducing the number of confusing doublets and the number or frequency of words_1 might be of interest. The construction of a composite difficulty score combining numbers and individual frequencies of words_1 and words_2, numbers and frequencies of doublets and semantic similarity is a further logical step in this project. Lastly, following these theoretical data obtained from the French lexicon, it would be worthwhile to obtain experimental data to examine whether speakers avoid these confusing doublets in everyday speech. Kaplan & Muratani (2015) indeed showed that Japanese speakers conversing with each other avoided applying a nasal contraction rule if it resulted in a confusing environment with two possible lexical segmentations. If liaison-induced near-homophony is not fully controlled inside the French lexicon, it may merely be avoided by French speakers.

Closing note

The length of this thesis does not fall within the range of 10,000 - 15,000 words stated in the writing guidelines.The objective was to provide a readable thesis that would be accessible to non-experts in the field - and thus not to overload the reader with technical information.

In an open science approach, a github repository is accessible, providing the scripts used for data extraction, analysis and visualization. These scripts allow one to obtain all the data and figures presented in the thesis, starting from the Lexique database without any modification. The different decisions that have been taken are thus totally transparent.
Github repository: https://github.com/Vantoine2019/CognitiveScienceMaster-LiaisonProject

References

Adda-Decker, M., Boula de Mareüil, P., & Lamel, L. (1999). Pronunciation variants in French: schwa & liaison. In *Proceedings of the XIVth International Congress of Phonetic Sciences* (pp. 2239-2242).

Armstrong, N. (2001). *Social and stylistic variation in spoken French*. John Benjamins.

Barreca, G., & Christodoulides, G. (2017). Analyse fréquentielle de la liaison variable dans un corpus de français parlé. *Journal of French Language Studies*, *27*(1), 27.

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.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*. Mit Press.

Dahan, D., & Magnuson, J. S. (2006). Spoken word recognition. In *Handbook of psycholinguistics* (pp. 249-283). Academic Press.

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.

Delattre, P. (1966). Les facteurs de la liaison facultative en français. In *Studies in French and Comparative Phonetics* (pp. 55-62). De Gruyter.

De Mareüil, P. B., Adda-Decker, M., & Gendner, V. (2003). Liaisons in French: a corpus-based study using morpho-syntactic information. In *Proc. of the 15th International Congress of Phonetic Sciences*.

Dubroca, L. (1824). *Traité de la prononciation des consonnes et des voyelles finales des mots français*. Delaunay et A. Johanneau.

Durand, J., & Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, *18*(1), 33.

Gow, 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.

Lin, Y., Michel, J. B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus.

Litté, E. (1873-1874). *Dictionnaire de la langue française* (online ed.). Hachette. https://www.littre.org/

Malécot, A. (1975). French liaison as a function of grammatical, phonetic and paralinguistic variables. *Phonetica*, *32*(3), 161-179.

Mallet, G. (2008). La liaison en français: descriptions et analyses dans le corpus PFC. *Unpublished PhD dissertation. Université Paris Ouest, France*.

Meinschaefer, J., Bonifer, S., & Frisch, C. (2015). Variable and invariable liaison in a corpus of spoken French. *Journal of French Language Studies*, *25*(3), 367-396.

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.

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.

Nguyen, N., Wauquier-Gravelines, S., Lancia, L., & Tuller, B. (2007). Detection of liaison consonants in speech processing in French: Experimental data and theoretical implications.

Selkirk, E. (1974). French liaison and the X notation. *Linguistic inquiry*, 573-590.

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

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.

TLFi (2004). *Trésor de la langue française informatisé*. CNRS Editions. http://atilf.atilf.fr/

Appendix A: grammatical constraints established to select grammatical doublets

This appendix provides the grammatical constraints that were used to select the grammatical doublets. For each liaison consonant, we provide, (i) the grammatical categories of the words 1 with examples, (ii) a precise list of grammatical categories that can follow the given words 1 with examples of liaison contexts and (iii) the grammatical doublets.

For example, there are only two words containing the liaison consonant [p] (*trop* and *beaucoup),* which are adverbs. Then, we list the four grammatical categories of words that can trigger liaison after *trop* and *beaucoup* (adjective, preposition, infinitive verb and past participle verb). Finally, the grammatical doublets are noted with each type of doublet occupying one row. With *trop* and *beaucoup,* three types of doublets were considered grammatical: those with (i) a minimal pair involving adjectives and/or prepositions (i.e. an adjective/adjective, adjective/preposition or preposition/preposition pair), (ii) a minimal pair of infinitive verbs and (iii) a minimal pair of past participles.

Following the notation used in Lexique, we use the following abbreviations:

- ADJ adjective
- ADJ:ind indefinite adjective
- ADJ:num numeral adjective
- ADJ:pos possessive adjective
- ART:def definite article
- ART:ind indefinite article
- ADV adverb
- AUX auxiliary
- CON conjunction
- NOM noun
- PRE preposition
- PRO:ind indefinite pronoun
- PRO:per personal pronoun
- PRO:rel relative pronoun
- VER verb
- modals modals (*devoir* 'should', *falloir* 'must', *pouvoir* 'can', *vouloir* 'to want')

- imp imperative
- ind indicative
- inf infinitive
- cnd conditional
- subj subjunctive
- par:pas past participle
- par:pre present participle

- s, sing singular
- p, plur plural
- m masculine
- 3p 3rd person singular

- agree gender agreement in gender (masculine / feminine)
- agree nbr agreement in number (singular / plural)
- agree pers agreement in person (for verbs)

- pre-nom pre-nominal
- post-nom post-nominal

These two last criteria were recorded for adjectives but not implemented in the final project as no French lexical database indicating whether an adjective is pre- or post-nominal was found.

Liaison consonant **[p]**

Words 1

- Adverbs: trop, beaucoup.

Examples

Word 1	Word 2	Examples
ADV	ADJ	Des enfants trop_attentifs
ADV	PRE	Il est trop_à l'école
ADV	VER(inf)	Trop_utiliser le four est dangereux
ADV	VER(par:pas) m.s.	Ils ont trop_ouvert le four

Constraints

Word 1	Minimal pairs of Words 2
ADV	ADJ, PRE
	VER(inf)
	VER(par:pas)

Liaison consonant **[ʁ]**

Words 1

- Adjectives: premier, dernier, léger.

Examples

Word 1	Word 2	Examples
ADJ	NOM m.s.	Mon premier_ami

Constraints

Word 1	Minimal pairs of Words 2
ADJ	NOM m.s.

Liaison consonant **[g]**

Words 1

- Adjective: long.

Examples

Word 1	Word 2	Examples
ADJ	NOM m.s.	Le long_hiver

Constraints

Word 1	Minimal pairs of Words 2
ADJ	NOM m.s.

Liaison consonant **[n]**

Words 1

- Pre-nominal adjectives: ancien, bon, certain, divin, prochain
- Indefinite adjective: aucun
- Numeral adjective + indefinite article: un
- Possessive adjectives: mon, ton, son
- Adverbs: bien, non
- Preposition: en
- Indefinite pronoun: rien
- Personal pronouns: en, on

Examples

Word 1	Word 2		Examples
ADJ	NOM m.s.		Ancien_ami
ADJ:ind	NOM m.s.		Aucun_ami ne m'a aidé
	ADJ m.s. (pre-nom)		Aucun_ancien ami ne m'a aidé
ADJ:num	NOM m.s.		Un_ami et deux ennemis
	ADJ m.s. (pre-nom)		Un_ancien ami et deux ennemis
ADJ:pos	NOM m.s.		Mon_ami
	ADJ m.s. (pre-nom)		Mon_ancien ami
ART:ind	NOM m.s.		Un_ami
	ADJ m.s. (pre-nom)		Un_bon ami
ADV	bien	ADJ	Elles sont bien_admiratives
		PRE	Elles sont bien_à l'école
		PRO:per	C'est bien_elle
		VER (inf)	Bien_arriver à l'heure est important
		VER (par:pas) m.s.	Elles ont bien_accepté la décision
	non	ADJ	Des œuvres non_illustrées
PRE	NOM		Une chaise en_ébène
	PRO:per		Croire en_elle …
	ADJ:ind		En_aucune manière
	ART:ind		En_un instant, …
	VER (par:pre)		En_arrivant ce matin, …
PRO:ind	PRE		Rien_à se reprocher
	VER (inf)		Tu ne dois rien_aimer …
	VER (par:pas) m.s.		Elles n'ont rien_aimé
PRO:per	on	VER (≠ par, ≠ inf) (3pers, sing)	On_arrive
		AUX (3p, sing)	On_a mangé
	en	VER (≠ par, ≠ inf ≠ imp)	Il en_achète tous les jours
		AUX (≠ par, ≠ inf, ≠imp)	Elles en_ont

Constraints

Word 1	Minimal pairs of Words 2
ADJ	NOM m.s.
ADJ:ind	NOM m.s./ ADJ m.s. (pre-nom)
ADJ:num	NOM m.s./ADJ m.s. (pre-nom)
ADJ:pos	NOM m.s./ADJ m.s. (pre-nom)
ART:ind	NOM m.s./ADJ m.s. (pre-nom)
ADV (bien)	ADJ / PRE
	PRO:per
	VER (inf)
	VER (par:pas) m.s.
ADV (non)	ADJ
PRE	PRO:per / ADJ:ind / ART:ind / VER (par:pre)
	NOM
PRO:ind	PRE
	VER (inf)
	VER (par:pas) m.s.
PRO:per (on)	VER (≠ par, ≠ inf) (3p, sing) / AUX (≠ par, ≠ inf, ≠imp) (3p, sing)
PRO:per (en)	VER (≠ par, ≠ inf) / AUX (≠ par, ≠ inf, ≠imp)

Liaison consonant **[t]**

Words 1

- Pre-nominal adjectives: grand, profond, excellent, petit
- Indefinite adjectives: maint, tout
- Numeral adjectives: cent, vingt
- Adverbs: quand, tant, tout
- Auxiliaries: aient, ait, est, ont, serait, …
- Conjunction: quand
- Relative pronoun: dont
- Verbs: tapent, exigent, sautait, ...

Examples

Word 1	Word 2	Examples
ADJ	NOM m.s.	Petit_ami
ADJ:ind (maint)	NOM m.s.	Maint_homme
ADJ:ind (tout)	NOM m.s.	Tout_homme est prié de ...
	ART:ind m.s.	Tout_un art de ….
	ADJ m.s. (pre-nom)	Tout_ancien combattant doit ...
ADJ:num	NOM p.	Cent_hommes
ADV (quand)	AUX (ind, cnd)	Quand_est-ce que ...
ADV (tant)	ADJ	Elles sont tant_affaiblies
	VER (par:pas) m.s.	Elles ont tant_attendu
ADV (tout)	ADJ	Il est tout_admiratif
	ADV	C'est tout_admirablement qu'il ...
	PRE	Tout_à-coup, tout_en marchant
	VER (par:pas) m.s.	Elles ont tout_abîmé
PRO:rel (dont)	ADJ:ind	Dont_aucun homme
	ADJ:num	Dont_un homme et deux femmes
	ART:ind	Dont_un ami
	PRO:ind	Dont_aucuns ne sortiront indemnes
	PRO:per	L'homme dont_elle tient ses yeux est son grand-père
	AUX (ind, cnd) (3p)	dont_aurait été victime
CON (quand)	ADJ:ind	Quand_aucun homme n'est capable ...
	ADJ:num	Quand_un homme et deux femmes viendront ...
	ART:ind	Quand_un ami ...
	PRO:ind	Quand_aucunes ne peuvent venir, ….
	PRO:per	Quand_elle part à la chasse
VER (≠par ≠inf ≠imp)	PRO:per (agree pers)	Descend_il (written : descend-t-il)
VER (≠ par) (modals)	PRO:per (agree pers)	Veut-il du beurre ?
	VER (inf)	Il veut_aimer son prochain.

Examples (continued)

Word 1	Word 2	Examples
VER (par:pre)	ADJ:ind	Ne prenant_aucune considération pour ...
	ADJ:num	Oubliant_un gâteau et deux boissons, ...
	ART:ind	Oubliant_un ami, ….
	PRE	Arrivant_à l'école
VER (par:pre) (modals)	ADJ:ind	Ne voulant_aucune erreur ...
	ADJ:num	L'homme voulant_une cerise et deux citrons
	ART:ind	L'homme voulant_un gâteau
	VER (inf)	L'homme pouvant_aimer ...
AUX (être)	ADJ (agree nbr)	Il est_évident
	ADV	Il est_agréablement surpris
	PRE	Il est_à l'école
	ART:ind	C'est_un homme joyeux
AUX (avoir)	ADV	Ils ont_admirablement surmonté l'épreuve
	ART:ind	Ils ont_un problème
	VER (par:pas) m.s.	Ils ont_aimé le spectacle

Contraints

Word 1	Minimal pairs of Words 2
ADJ	NOM m.s
ADJ:ind (maint)	NOM m.s
ADJ:ind (tout)	NOM m.s. / ART:ind m.s. / ADJ m.s. (pre-nom)
ADV (quand)	AUX (ind, cnd)
ADV (tant)	ADJ
	VER (par:pas) m.s.
ADV (tout)	ADJ / ADV / PRE
	VER (par:pas) m.s.
PRO:rel (dont)	ADJ:ind / ADJ:num / ART:ind / PRO:ind / PRO:per / AUX (ind, cnd) (3p)
CON (quand)	ADJ:ind / ADJ:num / ART:ind / PRO:ind / PRO:per
VER (≠par ≠inf ≠imp)	PRO:per (agree pers)
VER (≠par ≠inf≠ imp) (modals)	PRO:per (agree pers)
	VER (inf)
VER (par:pre)	ADJ:ind / ADJ:num / ART:ind / PRE
VER (par:pre) (modals)	ADJ:ind / ADJ:num / ART:ind / VER (inf)
AUX (être)	ADJ (agree nbr) / ADV / PRE / ART:ind
AUX (avoir)	ADV / ART:ind / VER (par:pas) m.s.

Liaison consonant **[z]**

Words 1

- Pre-nominal adjectives — anciens, bons, jolis, jolies, meilleurs, …
- Demonstrative adjective — ces
- Indefinite adjectives — aucuns, certains, certaines, plusieurs, …
- Interrogative adjectives — quelles, quels
- Numeral adjectives — cents, trois, deux, dix, six, ...
- Possessive adjectives — mes, nos, ses, tes, vos, leurs
- Adverbs — moins, pas plus, très, mieux
- Definite article — les
- Indefinite article — des
- Auxiliaries — auras, aurions, seras, êtres, suis, avez, ...
- Conjunction — mais
- Nouns — tribunaux, sondages, temples, ...
- Preposition — chez, après, dans, depuis, sans, sous, aux
- Personal pronouns — nous, vous, ils, elles
- Verbs (modals) — pouvions, pourrez, voudras, ...

Examples

Word 1	Word 2	Examples
ADJ	NOM p. (agree gender)	Derniers_amis
ADJ:ind	ADJ p. (pre-nom) (agree gender)	Certains_anciens amis
	NOM p. (agree gender)	Certaines_amies
ADJ:dem (ces)	ADJ p. (pre-nom)	Ces_anciens amis
	NOM p.	Ces_amis
ADJ:num	ADJ p. (pre-nom)	Deux_anciens amis
	NOM p.	Six_amis
ADJ:pos	ADJ p. (pre-nom)	Mes_anciens amis
	NOM p.	Mes_amies
ART:def	ADJ p. (pre-nom)	Les_anciennes amies
	NOM p.	Les_amies
ART:ind	ADJ p. (pre-nom)	Des_anciennes amies
	NOM p.	Des_amies
ADJ:int	ADJ p. (pre-nom) (agree gender)	Quels_anciens amis ...
	NOM p. (agree gender)	Quelles_amies ...
	AUX p. (ind,cnd) (3p)	Quelles_ont été ...
ADV (pas)	ADJ	Il n'est pas_attentif
	VER (par:pas) m.s.	Il n'a pas_aimé
	ART:ind	Ce n'est pas_un idiot
ADV (others)	ADJ	Il est très_attentif
	VER (par:pas) m.s.	Il l'a moins_aimé
AUX (être)	ADJ (agree nbr)	Vous êtes_attentifs
	ADV	Vous êtes_admirablement performants
	PRE	Tu es_à l'école le jeudi
	ART:ind	Tu es_un génie

Examples (continued)

Word 1	Word 2	Examples
AUX (avoir)	ADV	Vous avez_admirablement réussi
	ART:ind	Tu as_un château
	VER (par:pas) m.s.	Tu as_aimé le spectacle
CON (mais)	ADJ:ind	Mais_aucune erreur restera impunie !
	ADJ:num	Mais_une erreur sera suffisante pour ...
	ART:ind	Mais_un ami viendra te sauver.
	PRE	Il n'est pas avec Gérard mais_avec Jean.
	VER (inf)	C'est beau d'espérer mais_avoir un but est ...
	PRO:per	Il est fort mais_il ne tiendra pas le coup
	PRO:ind	Certains hommes sont beaux mais_aucuns ne ….
NOM	ADJ p. (post-nom)(agree gender)	Les tribunaux_urbains
PRE (chez, dans, sous, depuis, après)	ADJ:ind	Chez_aucun homme
	ADJ:num	Dans_une chapelle et deux églises
	ART:ind	Sous_une table
PRE (sans)	Same as PRE (chez, dans, sous, depuis, après) with, in addition :	
	VER (inf)	Il est venu sans_accepter notre offre
PRE (aux)	ADJ p. (pre-nom)	Il a rendu visite aux_anciens amis de Jean
	NOM p.	Il est allé aux_étangs.
	PRO:ind	Il prend aux_uns et aux_autres
PRO:per (nous, vous, ils, elles)	AUX p. (agree pers)	Elles_ont été ...
	VER p. (≠ par, ≠ inf) (agree pers)	Nous_aimons ...
VER (pouvoir, devoir, falloir, vouloir)	VER (inf)	Tu peux_aller aux toilettes

Contraints

Word 1	Minimal pairs of Words 2
ADJ	NOM p. (agree gender)
ADJ:ind	ADJ p. (pre-nom) (agree gender) / NOM p. (agree gender)
ADJ:dem (ces)	ADJ p. (pre-nom) / NOM p.
ADJ:num	ADJ p. (pre-nom) / NOM p.
ADJ:pos	ADJ p. (pre-nom) / NOM p.
ART:def	ADJ p. (pre-nom) / NOM p.
ART:ind	ADJ p. (pre-nom) / NOM p.
ADJ:int	ADJ p. (pre-nom) (agree gender) / NOM p. (agree gender) / AUX p. (ind,cnd) (3pers)
ADV (pas)	ADJ / ART:ind
	VER (par:pas) m.s. / ART:ind *(= pairs with ADJ and VER (par:pas) m.s. are impossible)*

Contraints (continued)

Word 1	Minimal pairs of Words 2
ADV (others)	ADJ
	VER (par:pas) m.s.
AUX (être)	ADJ (agree nbr) / ADV / PRE / ART:ind
AUX (avoir)	ADV / ART:ind / VER (par:pas) m.s.
CON (mais)	ADJ:ind / ADJ:num / ART:ind / PRE / VER (inf) / PRO:per / PRO:ind
NOM	ADJ p. (post-nom)(agree gender)
PRE (chez, dans, sous, depuis, après)	ADJ:ind / ADJ:num / ART:ind
PRE (sans)	ADJ:ind / ADJ:num / ART:ind / VER (inf)
PRE (aux)	ADJ p. (pre-nom) / NOM p. / PRO:ind
PRO:per (nous, vous, ils, elles)	AUX p. (agree pers) / VER p. (≠ par, ≠ inf) (agree pers)
VER (modals)	VER (inf)