

## **UC Merced**

### **Proceedings of the Annual Meeting of the Cognitive Science Society**

#### **Title**

Les Liaisons Dangereuses: Quantifying French liaison-induced homophony

#### **Permalink**

<https://escholarship.org/uc/item/6s55q1n9>

#### **Journal**

Proceedings of the Annual Meeting of the Cognitive Science Society, 44(44)

#### **Authors**

Peperkamp, Sharon  
Antoine, Victor  
Turnbull, Rory

#### **Publication Date**

2022

Peer reviewed

# *Les Liaisons Dangereuses: Quantifying French liaison-induced homophony*

**Sharon Peperkamp (sharon.peperkamp@ens.psl.eu)**

Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL / CNRS / EHESS, 29 rue d'Ulm  
75005 Paris, France

**Victor Antoine (chez.vantoine@gmail.com)**

Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL / CNRS / EHESS, 29 rue d'Ulm  
75005 Paris, France

**Rory Turnbull (Rory.Turnbull@newcastle.ac.uk)**

School of English Literature, Language and Linguistics, Newcastle University  
Newcastle, UK

## **Abstract**

The French phonological rule of liaison, whereby certain underlying word-final consonants surface only when the following word starts with a vowel, sometimes creates homophony. For instance, *un œuf* 'an egg' and *un neuf* 'a nine' are both pronounced [ɛ̃.nœf]. While homophony is cross-linguistically frequent, there is evidence that it is constrained in various ways. Here, we quantify liaison-induced homophony by comparing its occurrence in real French to that in a benchmark consisting of versions of French with modified liaison consonants. We find that liaison induces more homophony in the benchmark than in real French. This is the first evidence that a phonological rule that applies across words is subject to an anti-homophony bias.

**Keywords:** liaison; lexicon; word segmentation; homophony; anti-homophony, French

## **Introduction**

French liaison is a complex phonological rule whereby an underlying word-final consonant, present in a fixed set of words and morphemes, surfaces only when the following word starts with a vowel. For instance, the final /t/ of *petit* 'small', is pronounced in *petit âne* 'small donkey' but not in *petit chat* 'small cat' or in *c'est petit* 'it is small'. It is rare to find two adjacent vowels within a word in French (only 1.8% of the lemmas in the lexical database Lexique (New et al., 2004) contain such a sequence); liaison therefore removes an obvious word boundary cue. In addition, surfacing liaison consonants are resyllabified as onset consonants, thus blurring word boundaries and sometimes creating homophony. For instance, *un œuf* 'an egg' and *un neuf* 'a nine' are both pronounced [ɛ̃.nœf]. Despite subtle acoustic cues that distinguish homophonous sequences like these, the unintended form is activated during online speech processing, although to a lesser degree than the intended one (Spinelli et al., 2003). More generally, multiple lexical activations slow down online processing (Tabossi et al., 1995; Christophe et al., 2004). Thus, liaison-induced homophony makes it harder for the listener to process incoming speech.

Since at least Martinet (1955), linguists have entertained the hypothesis that there are diachronic pressures against

lexical homophony (e.g. English *bear/bare*). In the realm of phonological rules, there is some cross-linguistic evidence that word-level rules are indeed subject to an anti-homophony bias. First, such rules generally induce less homophony than expected by chance (Silverman, 2010; Kaplan, 2011). Second, some rules are blocked or less likely to apply whenever they would result in homophony (Barkai, 1978; Ichimura, 2006; Kaplan & Muratani, 2015). Are phonological rules that apply across words also subject to an anti-homophony bias? Here, we examine this question for the case of French liaison. Specifically, we compare the occurrence of liaison-induced homophony in French to a benchmark consisting of modified versions of French, i.e. in versions in which a given liaison consonant is replaced by another French consonant.

In a first, baseline, analysis, we examine all liaison-induced homophone doublets. In a second analysis, we take into account the finding that different kinds of homophones appear to be more or less tolerable within languages. In particular, there is modeling evidence that cross-linguistically, lexical homophones are avoided especially within syntactic and semantic categories, hence where the potential ambiguity is greatest (Ke, 2006; Dautriche et al., 2018). We might expect that if liaison-induced homophony is avoided, this holds especially for cases where neither of the lexical segmentations is clearly more likely than the other. To investigate this, our second analysis thus focuses on a subset of liaison-induced homophone doublets that are expected to pose real-life segmentation difficulty for the listener.

A few more preliminary remarks on liaison are in order. First, six French consonants can serve as liaison consonant: /z/, /n/, /t/, /r/, /p/, and /g/. The most frequent ones are /z/ and /n/ (Adda-Decker et al. 2012); both occur at the end of many determiners (/z/ as a plural marker) and other high-frequency words. Second, liaison is subject to syntactic and prosodic constraints. For instance, it is obligatory in certain syntactic contexts (e.g., Det-N, pronoun-V, Adj-N) optionally in others (e.g. Aux-PP, Prep-NP), and is forbidden in still others. For the present purposes, we do not distinguish between obligatory and optional cases of liaison (but see footnote 1 below). Third, the correspondence between the liaison consonant and its orthographic representation is not always

straightforward. In particular, liaison /z/ always corresponds to one of the graphemes <s> and <x>, and while liaison /t/ corresponds most often to the grapheme <t> it corresponds to <d> in the words *grand* ‘big’ and *quand* ‘when’.

All scripts used for data extraction, analysis and visualization are available for download at <https://github.com/Vantoine2019/CognitiveScienceMaster-LiaisonProject>.

## Part 1: Baseline analysis

In this part, we perform a baseline analysis, where we consider all cases in which liaison creates a homophone doublet, such as {*un œuf* – *un neuf*}.

### Method

A list of homophone doublets was systematically constructed using Lexique (New et al., 2004), as follows. From the 22,633 words with a frequency of at least one per million in Lexique’s corpus of movie subtitles, we extracted all 4569 words with an underlying liaison consonant that surfaces either obligatorily or optionally in certain syntactic contexts in informal speech.<sup>1</sup> The breakdown of these **liaison words** according to their liaison consonant is shown in Table 1.

Table 1: Liaison words extracted from Lexique.

Liaison consonant	N	examples
/z/	3,217	<i>les</i> ‘the <sub>pl.</sub> ’, <i>elles</i> ‘they <sub>fem.</sub> ’, <i>très</i> ‘very’, <i>leurs</i> ‘their <sub>pl.</sub> ’, <i>petites</i> ‘small <sub>fem-pl.</sub> ’, <i>bons</i> ‘good <sub>masc-pl.</sub> ’, <i>quelques</i> ‘some <sub>pl.</sub> ’, <i>dans</i> ‘in’, <i>pas</i> ‘not’
/t/	1,419	<i>grand</i> ‘big <sub>masc.sg.</sub> ’, <i>tout</i> ‘all, any’, <i>est</i> ‘is’, <i>comment</i> ‘how’, <i>ont</i> ‘have <sub>3rd-pl.</sub> ’
/n/	17	<i>un</i> ‘a <sub>masc.</sub> ’, <i>on</i> ‘we’, <i>mon</i> ‘my’, <i>bon</i> ‘good <sub>masc.sg.</sub> ’, <i>ancien</i> ‘old <sub>masc.sg.</sub> ’
/r/	3	<i>dernier</i> ‘last <sub>masc.sg.</sub> ’, <i>premier</i> ‘first <sub>masc.sg.</sub> ’, <i>léger</i> ‘light <sub>masc.sg.</sub> ’
/p/	2	<i>trop</i> ‘too’, <i>beaucoup</i> ‘many’
/g/	1	<i>long</i> ‘long <sub>masc.sg.</sub> ’

We also extracted all pairs of words that differ only in the presence vs. absence of one of {/z/,/t/,/n/,/r/,/p/,/g/} word-initially (e.g., *an* /ɑ̃/ ‘year’ – *rang* /rɑ̃/ ‘row’, *heureux*<sup>2</sup> /øʁø/ ‘happy’ – *peureux* /pøʁø/ ‘fearful’) (henceforth: minimal pair). The total number of minimal pairs is 2137, (mean: 356). Table 2 shows the breakdown per consonant.

<sup>1</sup> We thus excluded words that can trigger liaison only in formal speech, including conjunctions, infinitives ending in *-er*, modal verb forms, and adverbs ending in *-ment* (Delattre, 1966; Malécot, 1975; Armstrong, 2001).

Table 2: Number of pairs of words differing only in the presence vs. absence of an initial consonant used in liaison (‘minimal pairs’).

Consonant	/z/	/g/	/n/	/t/	/p/	/r/
Number of minimal pairs	6	179	256	367	487	842

We then created all possible doublets. That is, we combined each liaison word with each minimal pair that obeys two constraints: first, the relevant consonant distinguishing its members is identical to the liaison consonant, and second, both members can be preceded by the liaison word. For instance, {*dernier an* – *dernier rang*} ‘last year/row’ and {*trop heureux* – *trop peureux*} ‘too happy/fearful’ are valid doublets, but the first constraint rules out {*un an* – *un rang*} ‘a year/row’ and {*tout heureux* – *tout peureux*} ‘all happy/fearful’. Similarly, {*dernier appel* – *dernier rappel*} ‘last call/reminder’ is a valid doublet, but the second constraint rules out {*dernier appellent* – *dernier rappellent*}, where the masculine singular adjective *dernier* is followed by a 3<sup>rd</sup> person plural verb form, yielding an ungrammatical sequence.<sup>3</sup> We thus obtained a final set of 322 doublets (mean per consonant: 54; range: 0-133).

For the benchmark comparison, we created alternative versions of French by keeping the set of liaison words, but substituting their liaison consonants one by one with each of the other French consonants (including the other liaison consonants). For example, we created a version of French where the word *premier* ends in a liaison /f/, a version where it ends in a liaison /v/, a version where it ends in a liaison /t/, and so on. Non-liaison words were not changed. As French has 16 consonants that can occur word-initially, of which 6 can serve as liaison consonants, we thus created 6x15=90 alternative versions of French, in which homophone doublets were composed of the same liaison words but combined with other minimal pairs. Besides the minimal pairs extracted from Lexique to create doublets in real French, we thus extracted all additional pairs of words differing in the presence vs. absence of an initial consonant (N=3563, mean per consonant: 356). We then used the same two constraints as for real French to construct homophone doublets. For example, in the alternative French where liaison /n/ is replaced with /l/, {*bien imiter* – *bien limiter*} ‘to imitate/limit well’ and {*on imite* – *on limite*} ‘we imitate/limit’ are homophone doublets. Similarly, in the alternative French where liaison /n/ is replaced with (liaison) /r/, {*bien athée* – *bien raté*} ‘really atheist – well missed’ and {*un an* – *un rang*} ‘a year/row’ are homophone doublets. The mean number of homophone doublets in the alternative versions is 343 (range: 0-4047), for a total of 30,843. To illustrate, Table 3 provides the number of minimal pairs, with

<sup>2</sup> In French, <h> is always silent, but in some words it corresponds to ‘h-aspiré’, which behaves like a consonant and hence blocks liaison. This is not the case of *heureux*.

<sup>3</sup> Note that while most valid doublets involve words from the same grammatical category, in some they belong to different ones, e.g. {*quand on* – *quand ton*} ‘when we/your’.

examples, that yield a homophone doublet in real French as well as the benchmark versions for a sample liaison word, i.e. *bien* ‘well’.

Table 3: Number and examples of minimal pairs that yield a doublet with the sample liaison word *bien* ‘well’, in real French and in the benchmark versions.

Liaison consonant	N	examples
/n/ (real French)	3	être-naître, ôter-noter, ôté-noté
/r/	47	athée-raté, ouvrir-rouvrir, appeler-rappeler, emporté-remporté, établi-rétabli, ...
/k/	11	ouvert-couvert, analyser-canaliser, oser-causer, ôté-côté, omis-commis, ...
/s/	10	avant-savant, en-sans, aider-céder, ôter-sauter, eu-su, avoir-savoir, allé-salé, ...
/l/	10	atteint-latin, en-lent, égal-légal, imiter-limiter, eu-lu, ...
/b/	8	anal-banal, au-beau, avare-bavard, ôter-botter, eu-bu, ...
/d/	8	en-dans, écrit-décrit, eu-dû, ôté-doté, étendre-détendre...
/f/	7	aux-faux, utile-futile, été-fêté, ...
/m/	7	oral-moral, hériter-mériter, annuel-manuel, ...
/p/	7	heureux-peureux, oser-poser, eu-pu, uni-puni, ...
/ʃ/	5	au-chaud, aux-chaud, armer-charmer, armé-charmé, eu-chu
/v/	4	ôter-voter, ôté-voté, eu-vu, omis-vomi
/g/	2	athée-gâté, athée-gâtée
/t/	1	eu-tu
/ʒ/, /z/	0	

## Results and discussion

Numbers of homophone doublets for each liaison consonant in real French and in the alternative versions are shown in Figure 1. We performed one-tailed one-sample Wilcoxon signed rank tests with, for each liaison consonant, the number of homophone doublets in real French being used as the value against which the numbers of homophone doublets obtained in each alternative version were compared. The results of these tests are also shown in Figure 1. Note that the median of the number of homophone doublets obtained with the

alternative versions of French is significantly higher than the number of the homophone doublets of real French for four of the six liaison consonants (/z/, /t/, /n/, /g/). By contrast, no such difference was obtained for the liaison consonants /r/ and /p/.

Importantly, /r/ and /p/ account for only five of the total of 4569 liaison words. Hence, liaison-induced homophony does appear to broadly occur less frequently than what would be the case with different, randomly chosen, liaison consonants; hence, this is evidence for an anti-homophony bias.

This result is not difficult to understand when we consider the liaison consonant /z/: Due to its status as a plural marker, /z/ accounts for 70% of the liaison words (Table 1). At the same time, /z/ is by far the least frequent consonant word-initially and accounts for less than 1% of minimal pairs (Table 2). Rather unsurprisingly, then, /z/-liaison words yield no homophone doublets at all, despite being extremely numerous. By contrast, in the benchmark versions of French, these words end in a consonant for which the number of minimal pairs and, ultimately, the number of homophone doublets is on average much higher.

## Part 2: Subset analysis

In our baseline analysis we only considered grammatical doublets, i.e. doublets that can occur within sentences and that hence potentially induce segmentation difficulty for the listener. Yet, to what extent is such segmentation difficulty real? Frequency likely plays a key role here. For instance, given that *acte* [akt] ‘act’ is a high-frequency and *tact* [takt] ‘tact’ a low-frequency word, it should not be difficult to discard *grand tact* upon hearing *grand acte*. (By the same token, depending on context discarding *grand acte* upon hearing *grand tact* can of course be difficult, but this case will be rare.) In contrast, for other doublets it might be more difficult to dismiss one of the two alternatives, e.g. {*est en – est tant*} ‘is in/so’. Indeed, both *en* and *tant* are highly frequent in isolation as well as following *est*.

Intuitively, a homophone doublet poses real-life segmentation difficulty if two conditions are met: First, listeners experience it frequently, i.e. at least one member is frequent. Second, lexical segmentation is not straightforward, i.e. the two members are close in terms of frequency. To investigate whether homophone doublets in French induce few real-life difficulties, we extracted a frequency-based subset of doublets and compared the number of these doublets in real French with their numbers obtained in the alternative versions of French.

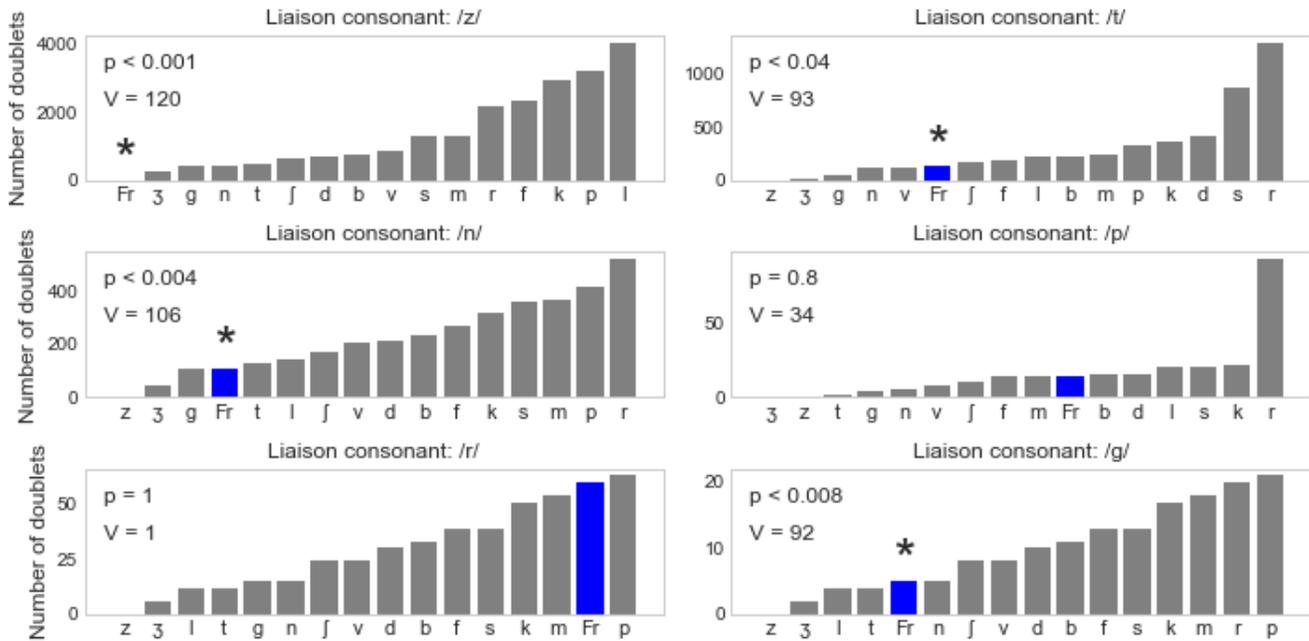


Figure 1: Number of homophone doublets for each liaison consonant (subplots) and each version (individual bars) in the baseline analysis. Subplots are ordered from largest (/z/) to smallest (/g/) y-axis. Blue bars represent data for real French and grey bars represent data for the alternative versions. (There is no blue bar for the subplot for /z/ as there are no homophone doublets.) For each liaison consonant, the V- and p-values from the Wilcoxon tests are displayed in the associated subplot.

## Method

Google Ngram Viewer (Michel et al., 2011; Lin et al., 2012) was used to retrieve frequency data of all pairs of words included in the homophone doublets examined in Analysis 1. For this, 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. We used the grammatical categories (in Google Ngram Viewer terminology: part-of-speech tags) in the queries in order to

obtain the most accurate data, after manually converting the grammatical categories given by Lexique to match those of Google Ngram Viewer as closely as possible.

In order to extract a frequency-based subset of doublets we proceeded as follows. Let  $a$  and  $b$  be the respective frequency values of the members of a given doublet. Then we extracted those doublets for which both the doublet frequency sum  $[a+b]$  and the doublet frequency ratio  $[\min(a, b) / \max(a, b)]$  are above the 50<sup>th</sup> percentile of the respective distributions.<sup>4</sup>

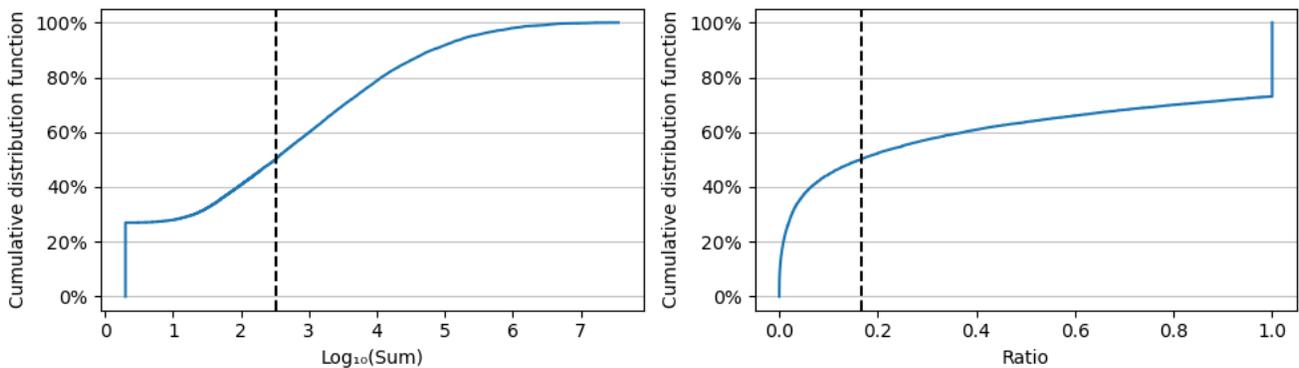


Figure 2: Cumulative distribution function of (i) the decimal logarithm of the sum of the frequency values of the doublets (left), and (ii) the ratio of the frequency values of the doublets (right). The dashed line delineates the 50<sup>th</sup> percentile value.

<sup>4</sup> We performed Laplace smoothing on the frequency values  $a$  and  $b$  (i.e. adding 1) to deal with cases where one or both were zero.

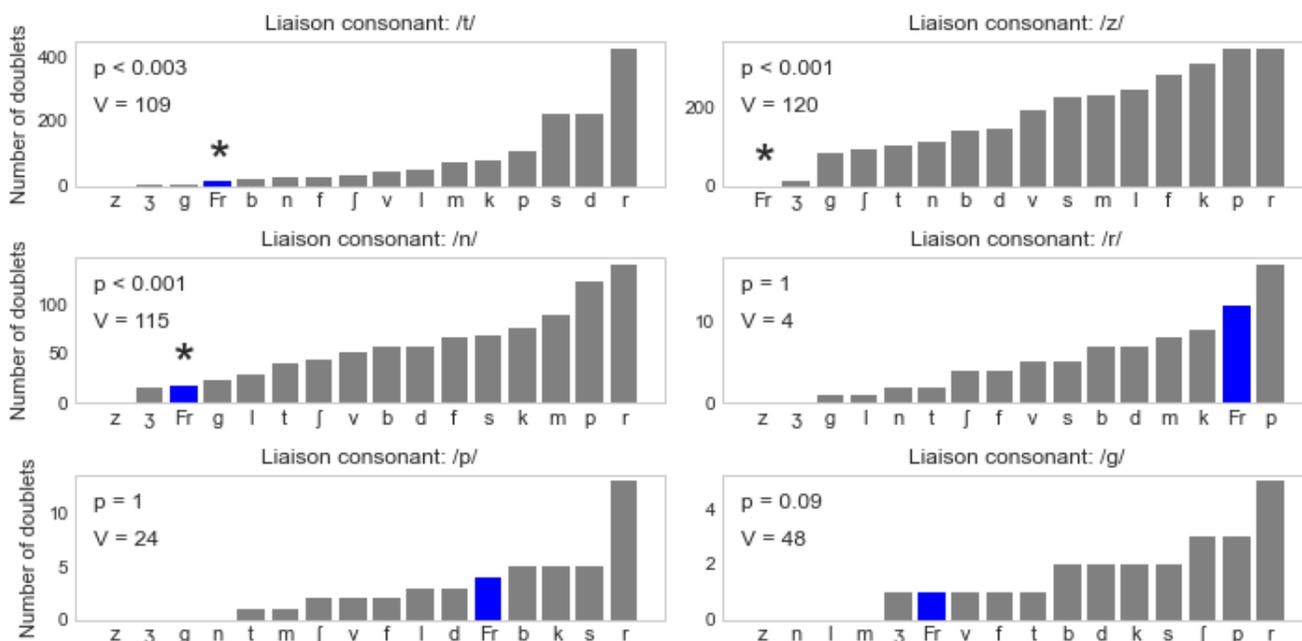


Figure 3: Number of homophone doublets for each liaison consonant (subplots) and each version (individual bars) in the subset analysis. Subplots are ordered from largest (/t/) to smallest (/g/) y-axis. Blue bars represent data for real French and grey bars represent data for the alternative versions. (There is no blue bar for the subplot for /z/ as there are no homophone doublets.) For each liaison consonant, the V- and p-values are displayed in the associated subplot.

Thus, we extracted doublets that are frequently encountered (frequency sum > 323) and whose members have close frequency values (frequency ratio > 0.17) (Figure 2, and note:  $10^{2.51} = 323$ ).

## Results and discussion

Numbers of homophone doublets for each liaison consonant in real French and in the alternative versions are shown in Figure 3. As before, the results of one-tailed one-sample Wilcoxon signed rank tests for the individual liaison consonant are shown in the Figure as well. Recall that for these tests, the number of doublets in real French is used as the value against which the numbers of doublets obtained in each alternative version are compared. We found that the median of the number of doublets obtained with the alternative versions of French is significantly higher than the number of homophone doublets of real French for the liaison consonants /z/, /t/, and /n/. By contrast, no such difference was obtained for the liaison consonants /r/, /p/, and /g/.

These results are very similar to the ones for all homophone doublets, examined in Part 1, where the comparison with the alternative versions of French yielded a significant difference in the same direction for /z/, /t/, /n/, and /g/. Thus, the only difference concerns /g/, for which evidence of a bias against homophony was found in the baseline analysis but not in the present one. As there is only one liaison word with /g/, we can hardly interpret this finding as evidence that – contrary to expectation – the anti-homophony bias is weaker when there

is more real-life segmentation difficulty. Note, though, that we clearly found no evidence either that homophony is especially avoided when neither of the lexical segmentations is more likely than the other.

## General discussion

For both the entire set of homophone doublets and the subset, we found larger amounts of liaison-induced homophony in a benchmark consisting of modified versions of French with alternative liaison consonants than in real French. As far as we know, the present study is the first one examining a phonological rule that applies across words and hence the first one providing evidence for a relatively low amount of homophony induced by such a rule, i.e. for an anti-homophony bias.

Our method of generating alternative versions of French relied on (i) a modification of the lexicon, since we changed the liaison consonant of liaison words, and (ii) a modification of the liaison rule, as it involved a new consonant with each substitution. For example, when the liaison consonant /r/ was replaced by /t/, we modified both all words containing the liaison consonant /r/ and the liaison rule. Our analysis is thus unique not only because it focuses on a rule that applies across words, but also because it relies on an interaction between lexicon and phonological rule.

Whether the bias against homophony reflects an incidental diachronic development that happens to reduce segmentation difficulty, or a cognitive pressure to avoid liaison-induced

homophony that constrained French diachrony, is an open question. The latter option would be consistent with evidence for a diachronic pressure against lexical homophony (Bouchard-Côté et al., 2013; Wedel et al., 2013).

There are several limitations to our study. First, we relied on the intuitions of a single native speaker of French, the second author, to decide which liaison words to exclude because they appear to trigger liaison only in formal speech. Other native speakers may have different intuitions on which words are concerned, given the variability of liaison production (Fougeron et al., 2001; Meinschaefer et al., 2015). Second, due to the large size of the benchmark, we used a set of general rules to discard ungrammatical doublets rather than filter them out by hand. This left a number of ungrammatical ones. For instance, the adverb *beaucoup* ‘a lot’ is exceptional in that it cannot modify any adjectives. Therefore, {*beaucoup heureux* – *beaucoup peureux*} is not a valid doublet in real French (cf. valid {*trop heureux* – *trop peureux*}, with the adverb *trop* ‘too’). Other doublets are questionable or outright impossible for semantic reasons.<sup>5</sup> For instance, *long os* ‘long bone’ is fine, but *long gosse* ‘long kid’ is not. Finally, doublets were included in the subset analysis based on the conjunction of two thresholds, one concerning the sum and the other the ratio of the frequency values of each member of a doublet. Using a single composite threshold would be more accurate, but it remains a modeling challenge given the widely different distributions of the two parameters. Moreover, our analysis did not take into account a possible interaction of ratio and sum for doublets’ real-life segmentation difficulty.

Future research could focus on refining the definition of doublets that pose real-life segmentation difficulty, to further examine whether liaison-induced homophony is more strongly avoided in cases where lexical segmentation is particularly difficult. This could be done in two ways. First, additional grammatical constraints could be implemented. For instance, {*est au* – *est tôt*} ‘is at the – is early’ is a valid doublet since both members are grammatical sequences. Yet, the next word will necessarily disambiguate the sequence, as *au* can be followed only by a noun or an adjective (e.g., *elle est au marché* ‘she is at the market’; *elle est au dernier cours* ‘she is at the last class’), neither of which can follow *tôt*. Second, following what has been done for lexical homophony (Dautriche et al., 2018), Latent Semantic Analysis would allow us to examine the role of semantic similarity. In particular, the less semantically similar the members of a minimal pair are, the easier it is for the listener to discard one of them (cf. dissimilar *or* ‘gold’ – *tort* ‘wrong’ and highly

similar *achat* ‘purchase’ – *rachat* ‘repurchase’). In addition, one might examine doublets involving liaison words that trigger liaison in formal speech only and that were discarded in the present study, such as adverbs in *–ment* (see footnote 1). Would these doublets show less evidence for homophony avoidance, given that they arise less often?

Ultimately, it would be interesting to address the topic of an anti-homophony bias experimentally, and test whether speakers avoid liaison-induced homophone doublets in everyday speech. Kaplan & Muratani (2015) showed that Japanese speakers conversing with each other avoid applying a nasal contraction rule if it introduces a lexical segmentation ambiguity. We would expect that French speakers similarly avoid homophone doublets, especially ones that were included in our subset analysis, either by not applying the rule (in case of optional liaison), or by avoiding the two-word sequence (in case of obligatory liaison).

### Acknowledgments

This research was supported by grants from the ANR (ANR-17-CE28-0007-01 and ANR-17-EURE-0017).

### References

- Adda-Decker, M., Fougeron, C., Gendrot, C., Delais-Roussarie, E., & Lamel, L. (2012). La liaison dans la parole spontanée familière : Une étude sur grand corpus. *Revue française de linguistique appliquée*, 17, 113-128.
- Armstrong, N. (2001). *Social and stylistic variation in spoken French*. Amsterdam: John Benjamins.
- Barkai, M. (1978). Phonological opacity v. semantic transparency: Two cases from Israeli Hebrew. *Lingua*, 44, 363-378.
- Bouchard-Côté, A., Hall, D., Griffiths, T., & Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110, 4224-4229.
- 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, 523-547
- 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). *Studies in French and Comparative Phonetics*. The Hague: De Gruyter Mouton.

<sup>5</sup> Due to the large size of the benchmark it is impossible to check all the doublets for grammaticality and semantic acceptability. However, we did examine the 322 doublets obtained in real French. We found 22 ungrammatical ones, of which 16 are due to mistakes in Lexique (i.e., the plural noun *ados* ‘adolescents’ labeled as singular (three doublets); the feminine noun *aide* ‘help’ having no label for gender (one doublet); and the loanword *nan* ‘nan’ with an incorrect phonetic transcription (12 doublets)). Among the six remaining ones, five concern the incorrect prenominal occurrence of

the adjective *naval* ‘naval’; the last one is {*beaucoup heureux* – *beaucoup peureux*} mentioned in the main text. As to the semantic acceptability of the 300 grammatical doublets, we found 26 of them to be questionable or outright impossible. (Many acceptable ones occur in rare contexts, though.) In order to not introduce a processing difference between real French and the benchmark versions, we did not exclude any of these ungrammatical or semantically problematic doublets.

- Fougeron, C., Goldman, J.-P., Dart, A., Guélat, L., & Jeager, C. (2001). Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. *Actes de la 8<sup>ème</sup> conférence sur le traitement automatique des langues naturelles* (pp 173-182).
- Ichimura, L. (2006). Anti-homophony blocking and its productivity in transparadigmatic relations. Doctoral dissertation, Boston University.
- Kaplan, A. (2011). How much homophony is normal? *Journal of Linguistics*, 47, 631–671.
- Kaplan, A., & Muratani, Y. (2015). Categorical and gradient homophony avoidance: Evidence from Japanese. *Laboratory Phonology*, 6, 167-195.
- Ke, J. (2006). A cross-linguistic quantitative study of homophony. *Journal of Quantitative Linguistics*, 13, 129-159.
- Lin, Y., Michel, J. B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic annotations for the google books ngram corpus. *Proceedings of the 50<sup>th</sup> annual meeting of the association for computational linguistics. Volume 2: Demo Papers*.
- Malécot, A. (1975). French liaison as a function of grammatical, phonetic and paralinguistic variables. *Phonetica*, 32, 161-179.
- Martinet, A. (1955). *Économie des changements phonétiques*. Berne, Éditions A. Francke S. A.
- Meinschaefter, J., Bonifer, S., & Frisch, C. (2015). Variable and invariable liaison in a corpus of spoken French. *Journal of French Language Studies*, 25, 367-396.
- Michel, J.-B., Shen, Y., ..., & Lieberman Aiden, E. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331, 176-182.
- New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36, 516-524.
- Silverman, D. (2010). Neutralization and anti-homophony in Korean. *Journal of Linguistics*, 46, 453-482.
- Spinelli, E., McQueen, J., & Cutler, A. (2003). Processing resyllabified words in French. *Journal of Memory and Language*, 48, 233-254.
- Tabossi, P., Burani, C., & Scott, D. (1995). Word identification in fluent speech. *Journal of Memory and Language*, 34, 440-467.
- Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, 128, 179-186.