Studii de lingvistică 11, 2021, 39 - 53

Distribution and deletion of /ʁ/ in fluent speech

Yaru Wu[1]
Martine Adda-Decker[2]

Abstract: This study provides a detailed overview of the distribution of French /ʁ/ in different word positions, and investigates factors conditioning its potential deletion in fluent speech. Three manually transcribed speech corpora totaling ~200 hours of spoken French were used: the ESTER corpus of formal journalistic speech, the ETAPE corpus of informal journalistic speech, and the NCCFr corpus of casual speech. Forced alignment of /ʁ/ and schwa variants was used to produce phonetic transcriptions from the orthographic. The distribution of /ʁ/ was measured in 20 segmental contexts, of which the seven most frequent account for over 90% of all occurrences (word-types 93%; word-tokens 91%). Word-internal positions are shown to influence /ʁ/ deletion: /ʁ/ in word-final position or in word-final consonant clusters is more susceptible to deletion than /ʁ/ in other positions (i.e. word-initial and -internal). Post-lexical contexts also affect /ʁ/ deletion, triggering deletion more frequently in post-lexical consonantal contexts than in vocalic. As to speech style, the less formal the speech style, the more /ʁ/ is deleted.

Key words: French, /ʁ/-deletion, variation, large corpora, fluent speech, post-lexical context, speech style.

1. Introduction

In French, /ʁ/ happens to be the most frequent consonant (Wioland 1985, Adda-Decker 2006). The /ʁ/ consonant may appear in almost any position of French words including a wide range of consonant clusters. We can find it as the unique consonant in syllable onset or in syllable coda position, at the beginning, middle or end of a word. /ʁ/ can also be found in consonant clusters such as Cʁ either as a syllable onset (e.g. *très* /**tʁ**ɛ/ 'very') or as a syllable coda, (e.g. *autre* /o**tʁ** 'other'). When /ʁ/appears in an ʁC cluster, it is either in

[1] CRISCO/EA4255, Université de Caen Normandie; Laboratoire de Phonétique et Phonologie, UMR7018, CNRS-Sorbonne Nouvelle; yaru.wu@sorbonne-nouvelle.fr.
[2] Laboratoire de Phonétique et Phonologie, UMR7018, CNRS-Sorbonne Nouvelle; LISN/ CNRS, UMR 9015, Université Paris-Saclay; madda@limsi.fr.

word-internal position or word-finally in a syllable coda. More complex clusters containing /ʁ/ can be found.

Previous studies mainly focused on its production variation in word-final position, potentially word-final consonant clusters (Nyrop 1914, Grammont 1933, Laks 1977, Cornulier 1978, Brand & Ernestus 2015, Gendrot 2017). Up to now, only few studies have addressed the question of /ʁ/'s general distribution and realization, respectively deletion. This study aims to increase our knowledge about the general distribution of /ʁ/ and its potential deletion in various segmental contexts by using large speech corpora and tools from automatic speech recognition system.

More precisely, we will give a general description of the distribution of /ʁ/ in spoken French, as derived from the corpus' orthographic transcriptions. We use automatic forced alignment results to measure the realization and deletion rates of /ʁ/. We are investigating the influence of both segmental context and position in words on the realization or deletion of /ʁ/. We hypothesize that /ʁ/ is deleted more often in word-final position than in word-initial and word-internal positions (Morin, 1986). Furthermore, in the case of consonant clusters, the number of consonants in a row is expected to play a role: the longer the consonant cluster, the more /ʁ/ tends to be deleted, following a similar but opposite tendency as the well-known "law of three consonants" for schwa in French (Grammont 1894). The proposed approach makes use of specific pronunciation variants with and without /ʁ/ in the proposed surface forms during forced alignment. This allows us to quantify /ʁ/ deletion tendencies with respect to a large variety of most representative contexts in French.

2. Method

Three large corpora of French continuous speech (≈ 190 hours) were analyzed in this study: formal journalistic corpus ESTER (Galliano *et al.* 2006), informal journalistic corpus ETAPE (Gravier *et al.* 2012) and the Nijmegen Corpus of Casual French (NCCFr, Torreira *et al.* 2010). The ESTER corpus (≈ 100 hours) mainly contains broadcast news, whereas ETAPE (≈ 50 hours) is mostly composed of public conversations and debates. NCCFr (≈ 40 hours) comprises casual conversations between friends.

We used the automatic speech recognition system at LISN/CNRS (former LIMSI/CNRS, Gauvain *et al.* 2002, Adda-Decker & Lamel 2000), and more precisely forced alignment with schwa and /ʁ/ variants. Thus, when the ASR system is in forced alignment mode, the system indicates whether the schwa segment or the /ʁ/ is present according to the acoustic information. Manual verifications were carried out on a subset of the data (~6000 occurrences). Cohen's kappa coefficient

used to evaluate the agreement between the automatic and the manual alignments shows almost perfect agreement (kappa = 0.832).

For instance, the word *quatre* /katʁ/ 'four' could be aligned with either [katʁ], [kat], [katʁə] or [katə] depending on the speakers' productions. Figure 1 shows an example of the forced alignment including schwa and /ʁ/ variants. It is worth mentioning that the minimum duration of a segment is 30ms using the LISN speech transcription system. This method allows us to identify /ʁ/ reduction cases, including deletion of /ʁ/ and extremely short /ʁ/ segments.

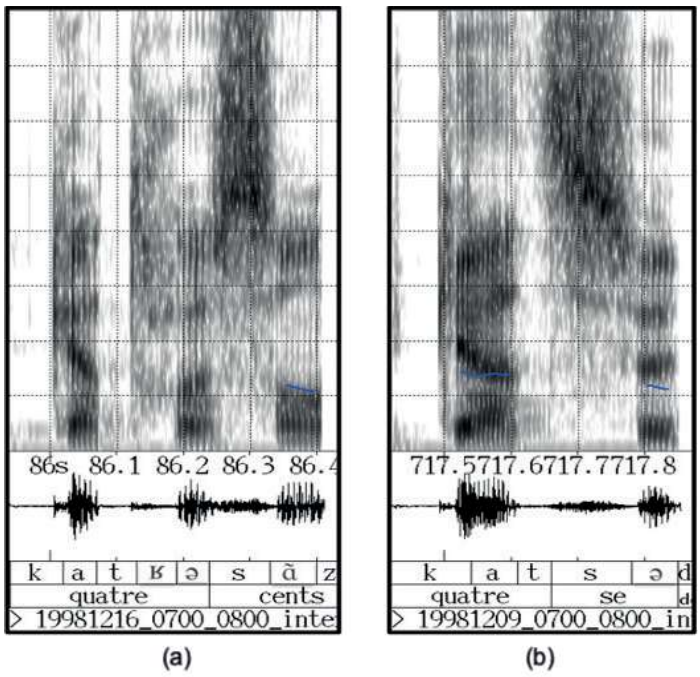

Figure 1: The word *quatre* /katʁ/ 'four') aligned by the LISN automatic speech recognition system (figure from Wu *et al.* 2019) using forced alignment with /ʁ/ and schwa variants. (a) shows /katʁ/ aligned with [katʁə] and (b) shows /katʁ/ aligned with [kat].

We used the pronunciation dictionary *Lexique 380* as a reference (New *et al.* 2007) to quantify the achievement or not of /ʁ/. Each word-token in our corpus will thus be accompanied by a reference pronunciation (*Lexique 380*, canonical pronunciation) and a surface pronunciation (aligned pronunciation).

As far as quantization of words containing /ʁ/ is concerned, note that if a word contains more than one /ʁ/ in different positions, it will be counted once for each position. For example, the word *directrice* /diʁɛktʁis/ 'director' contains two /ʁ/: the first is of the form "ʁ" and

the second is of the form "CCʁ". This word is therefore counted once for the form "ʁ" and once for the form "CCʁ". This concerns 15% (2992 occurrences) of the categorized word-types. As for word-tokens, if a word has two /ʁ/ in different positions, it is counted once in the first position and once in the second. We thus have 594283 occurrences (instead of 502759 occurrences, not counting the repetition of /ʁ/ in the same word) for the quantification of /ʁ/ in the three corpora. We excluded Radio Télévision Marocaine (RTM) from the ESTER corpus (73794 occurrences out of 594283 – 12%) in order to rule out possible variability in pronunciation from a French-speaking country other than France. We thus have 520489 occurrences for this quantification study after this removal. We have a quite large number of occurrences (5855) of the word *parce* /paʁs/ in the category "Cʁ#", which is almost always used in the expression *parce que* /paʁs#kə/ 'because'). We have therefore decided not to consider this word in this study. Thus, the rate of word-tokens in the category "Cʁ#" presented in this study will not be influenced by a specific frequent word which itself has an ambiguous lexical status. After removing *parce* from the "Cʁ#" category, we have a total of 514634 occurrences available for this study.

In this study, we are particularly interested in the distribution and the deletion of /ʁ/ in French. With regard to the distribution of /ʁ/, word-types and word-tokens containing /ʁ/ are categorized into 20 segmental contexts: #ʁ, #Cʁ, #ʁC, #CCʁ, #CʁC, ʁ, ʁC, Cʁ, ʁCC, CCʁ, CʁC, ʁCʁ, CCCʁ, CCʁC, ʁ#, ʁC#, Cʁ#, ʁCC#, CCʁ#, ʁCʁ#. Concerning analyses on /ʁ/ deletion, we investigated the impact of within-word position (initial, internal and final), post-lexical contexts (consonant and vowel segments of the preceding or the following word)[3] and speech style (formal journalistic speech, informal journalistic speech and casual speech). The three within word positions were obtained from a recategorization of the 20 segmental contexts according to their position in the word. It needs to be noted that in this study, when we talk about positions in the word for /ʁ/, we refer to /ʁ/ in a specific position either as a single consonant or as part of a consonantal cluster:

- **initial**: /ʁ/ in word-initial position or in word-initial onset consonantal cluster,
- **internal**: /ʁ/ in word-internal position or consonant cluster containing /ʁ/ surrounded by two vowels,
- **final**: /ʁ/ in word-final position or in word-final consonant cluster.

Analyses on the distribution of words with /ʁ/ concern all words containing at least one /ʁ/. In order to better understand the influence of within-word position on /ʁ/ deletion, we included segmental contexts

[3] In this article, we refer to "post-lexical" as beyond word boundaries (opposed to "within word").

containing more than 2k occurrences in the canonical form of /ʁ/ words in all three corpora. Analyses on the preceding post-lexical contexts were based on word-initial /ʁ/ or word-initial consonant clusters containing /ʁ/ of segmental contexts that have more than 300 occurrences for both vocalic and consonantal contexts. Similarly, analyses on the following post-lexical contexts were based on word-final /ʁ/ or word-final consonant clusters containing /ʁ/ of segmental contexts that have more than 300 occurrences for both vocalic and consonantal contexts.

Generalized linear mixed models (GLMM) were used for the statistical analyses of this study (McCulloch & Neuhaus 2005). Three GLMMs were carried out given that analyses on post-lexical contexts concern two different subsets of the data and the post lexical contexts concern preceding or following context of the word in question according to different segmental contexts. The first GLMM (A) was used to test the influence of within-word position, number of consonant(s) in the reference form and speech style on the realization or not of French /ʁ/. Within-word position (initial, internal and final; reference: initial), number of consonant(s) (1C, 2C and 3C; reference 1C) and speech style (formal journalistic speech ESTER, informal journalistic speech ETAPE, casual speech NCCFr; reference: formal journalistic speech ESTER) were included as fixed effects. Intercepts for subjects and items were considered as random effects. The second GLMM (B) examined the influence of preceding post-lexical context for word-initial /ʁ/ or word-initial consonant clusters containing /ʁ/. The data set is much smaller than that used in the first model, since only word-initial /ʁ/ or word-initial consonant clusters containing /ʁ/ were concerned. Preceding post-lexical context (consonant and vowel; reference: consonant) was included as fixed effect. Number of consonant(s) and speech style were included as control variables and intercepts for subjects and items were considered as random effects. The third model (C) investigates the following post-lexical context for word-final /ʁ/ or word-final consonant clusters containing /ʁ/. Therefore, only word-final /ʁ/ or word-final consonant clusters containing /ʁ/ were concerned in this model. Following post-lexical context (consonant and vowel; reference: consonant) was included as fixed effect. Number of consonant(s) and speech style were included as control variables. Intercepts for subjects and items were considered as random effects. Post-hoc tests based on each model were carried out to obtain information on each level of the fixed effects.

3. Results

In this section, we first present the distribution of both word-types as well as word-tokens containing /ʁ/ according to different segmental contexts and their positions in the words (Section 3.1).

The subsequent section present /ʁ/ deletion rates, which are computed based on the comparison between a word's reference pronunciation and its surface pronunciation as selected during forced alignment. In Section 3.2, quantitative results of /ʁ/ deletion are presented by considering different segmental contexts restricted to the word without the neighbouring post-lexical context. In Section 3.3, we take a closer look at the influence of the post-lexical context (here the immediately preceding or following segment of the word containing /ʁ/). In section 3.4, we examine the influence of speech style on /ʁ/ deletion. As mentioned before, we have used journalistic speech (the more formal ESTER, and less formal ETAPE corpora) and casual face-to-face conversations between friends (NCCFr corpus) and quantified the deletion of /ʁ/ for different segmental contexts.

1.1. Distribution of words containing /ʁ/

Frequency distributions of both word-types and word-tokens containing /ʁ/ in the reference pronunciation form are presented in this section.

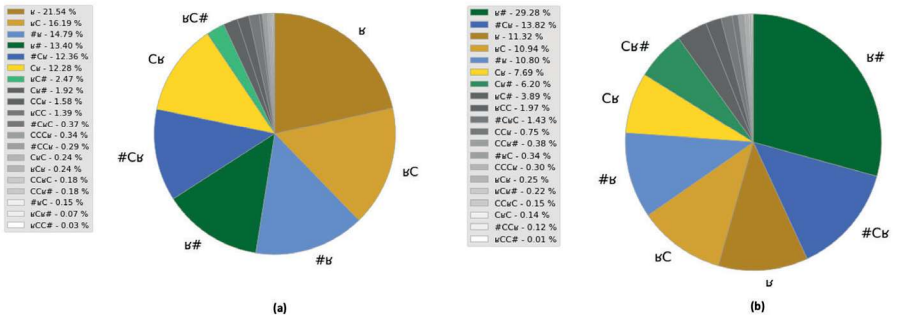

Figure 2: Frequency distribution of word-types (a) and word-tokens (b) containing /ʁ/ in different segmental contexts. Colors code for word-position: blue, yellow and green for word-initial, -internal and -final positions respectively. Low rates are kept in grey.

Figure 2 shows the rates of word-types (a) as found in a lexicon, and word-tokens (b) corresponding to usage in speech, containing /ʁ/ in different segmental contexts. Blue, yellow and green are used for word-initial, -internal and -final positions respectively. By taking a general look at the colours in Figure 2, it can be noticed that word-final /ʁ/ (in green) is more present in the word-token chart (right) than in the word-type chart (left).

In the French lexicon, most of the words containing /ʁ/ are of the forms "ʁ" (≈ 22%, 4.4k occurrences), "ʁC" (≈ 16%, 3.3k occurrences)

and "Cʁ" (≈ 12%, 2.5k occurrences) with /ʁ/ in word-internal position, either as a unique C embedded between two Vs or in a 2C consonant clusters. The rate of word-types in word-initial position "#ʁ" (≈ 15%, 3.0k occurrences) is similar to the rate of word-types in word-final position "ʁ#" (≈ 13%, 2.7k occurrences). /ʁ/ is also often in C_2 position of word-initial consonant clusters "#Cʁ" (≈ 12%, 2.5k occurrences). The remaining 14 forms represent less than 10% of the word-types containing /ʁ/.

The frequency of words containing /ʁ/ in fluent speech is rather different from the picture described above for the French lexicon. With respect to word tokens, the forms with a rate higher than 10% are "ʁ#" (≈30%, 152.4k occurrences), "#Cʁ" (≈ 14%, 71.9k occurrences), "ʁ" (≈ 11%, 58.9k occurrences), "ʁC" (≈ 11%, 56.9k occurrences) and "#ʁ" (≈ 11%, 56.2k occurrences). It is interesting to highlight that the most frequent word-type containing /ʁ/ ("ʁ"$_{type}$ ≈ 22%, see Fig. 1a) ranks third in the word-token figure ("ʁ"$_{token}$ ≈ 11%, see Fig. 1b) and most frequently observed word-token containing /ʁ/ ("ʁ#"$_{token}$ ≈30%, see Figure 1b) concerns only about 13% of the word-types with /ʁ/ (see Fig. 1a).

Form	Example		Word-type			Word-token			
			ESTER	ETAPE	NCCFr	Total	ESTER	ETAPE	NCCFr	Total
Word-initial position/cluster:										
#ʁ	*rat* 'rat'	/ʁa/	2115	1989	1009	3002	25037	23432	7775
#Cʁ	*cri* 'scream'	/kʁi/	1828	1621	823	2508	29997	28699	13245
#ʁC	*roi* 'king'	/ʁwa/	21	18	11	30	586	578	618
#CCʁ	*stressant* 'stressful'	/stʁɛsɑ̃/	36	38	27	58	289	255	75
#CʁC	*croix* 'cross'	/kʁwa/	60	50	39	76	3201	2671	1552
Word-internal position/cluster:										
ʁ	*Paris* 'Paris'	/paʁi/	2976	2675	1199	4371	27540	22335	9040
ʁC	*permet* 'allows'	/pɛʁmɛ/	2390	2056	898	3282	25984	22885	8055
Cʁ	*fondra* 'will melt'	/fɔ̃dʁa/	1767	1557	676	2491	19509	15293	5213
ʁCC	*pourquoi* 'why'	/puʁkwa/	213	172	79	281	4839	4130	1261
CCʁ	*électrique* 'electric'	/elɛktʁik/	227	209	99	320	1962	1423	506
CʁC	*détruit* 'destroyed'	/detʁɥi/	43	37	7	49	302	323	93
ʁCʁ	*surprise* 'surprise'	/syʁpʁiz/	42	29	16	48	715	476	128
CCCʁ	*extreme* 'extreme'	/ɛkstʁɛm/	55	46	20	68	711	751	120
CCʁC	*construit* 'built'	/kɔ̃stʁɥi/	24	22	13	37	467	255	59
Word-final position/cluster:										
ʁ#	*par* 'by'	/paʁ/	2087	1844	1011	2718	66567	59794	26027
ʁC#	*porte* 'door'	/pɔʁt/	389	325	232	500	6157	5537	2678
Cʁ#	*quatre* 'four'	/katʁ/	331	283	168	390	13477	13342	5453
ʁCC#	*cercle* 'circle'	/sɛʁkl/	5	3	3	6	34	20	9
CCʁ#	*semester* 'semester'	/səmɛstʁ/	30	22	12	37	1038	852	94
ʁCʁ#	*perdre* 'lose'	/pɛʁdʁ/	12	11	8	14	648	420	102

Table 1: Number of word-types and word-tokens containing /ʁ/ in different segmental contexts according to the corpora (speech style) studied

The details of the occurrences are shown in Table 1. It is worth mentioning that the total number of word-types for each form is not necessarily the sum of the occurrences observed for the three corpora, since a word could appear in several corpora. For example, we could

have 331 occurrences for the word-types of the form "Cʁ#" for the ESTER corpus, 283 for the ETAPE corpus and 168 for the NCCFr corpus and only 390 for the sum of the three corpora. In addition to what we saw in Figure 2, we notice that the vocabulary is much richer for ESTER and ETAPE than for the NCCFr corpus in words containing the consonant /ʁ/ in the reference pronunciation form. This is not due to the size of each corpus, but to the choices of the speakers because even in the case of forms for which we have more word-tokens in NCCFr (e.g. #ʁC), we still have more type-words in ESTER and ETAPE corpora than in NCCFr corpus.

1.2. Within-word positions

In the following, /ʁ/ deletion rates are presented as a function of different segmental contexts according to within word positions (initial *vs* internal *vs* final).

According to Grammont's "law of three consonants" for schwa in French (Grammont, 1894), the realization of schwa becomes mandatory when the surface form resulting from schwa deletion would result in three or more consonants in a row. Similarly, if the consonant clusters containing /ʁ/ have three or more consonants in a row, we would expect /ʁ/ deletion rates to be higher for these long consonant clusters (as compared to two-consonant clusters or simple /ʁ/ segments).

Based on the coda effects (Harris 1994, Blevins 1995, Ségéral & Scheer 2008), /ʁ/ is expected to be deleted more in word-final position than in word-initial position. It would be interesting to test whether /ʁ/ deletion rate in word-internal position would be closer to that in word-initial position or that in word-final position in our data.

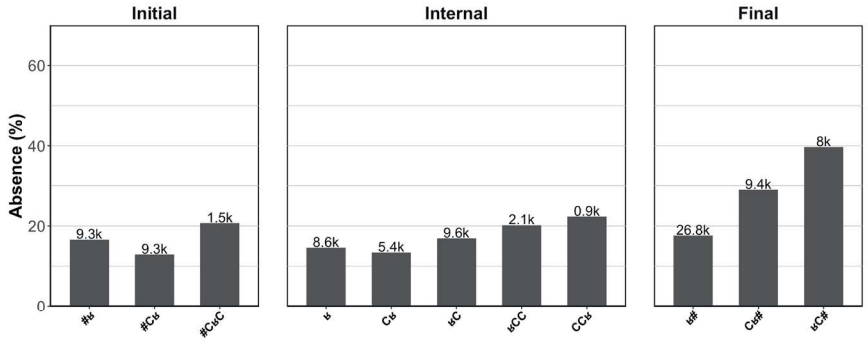

Figure 3: /ʁ/ deletion rates and corresponding number of tokens according to different segmental contexts grouped by word positions (initial *vs* internal *vs* final). Only segmental contexts containing more than 2k occurrences in the canonical form of /ʁ/ words in all three corpora are shown in this figure.

Figure 3 illustrates the deletion rate of /ʁ/ as a function of different segmental contexts and word position (all corpora pooled). Segmental contexts containing more than 2k occurrences in the canonical form of /ʁ/ words in all three corpora are shown in this figure so that we could focus on the relatively frequent segmental contexts. In general, longer consonant clusters tend to have a higher deletion rate than shorter consonant clusters or a single consonant, which is in line with what we expected based on similar logic from Grammont's "law of three consonants" for schwa in French. Results show that /ʁ/ is deleted more in in word-final clusters and it tends to have similar deletion rates in word-initial and -internal positions. The results concerning word-final and word-initial positions are in line with our predictions: /ʁ/ deletion is observed more in word-final position than in word-initial position. It is interesting to observe that Cʁ achieves the lowest deletion rates (lower than simple /ʁ/ contexts) in word-initial and -internal positions. However, this is not the case for word-final Cʁ# (which has a very high deletion rate). This could be related to the fact that word-final Cʁ# violates the sonority sequencing principle (Clements 1990). Results of the GLMM model (A) confirm that /ʁ/ deletion rate is significantly higher in word-final position than that observed in word-initial position [log odds ratio = 0.038282, $|Z| = 2.063$, $p < 0.05$]. No significant difference is found between word-initial position and word-internal position. Concerning the effect of number of consonant(s) in the investigated segmental contexts, higher /ʁ/ deletion rate is found for 3C than for 1C [log odds ratio = 0.227187, $|Z| = 4.823$, $p < 0.001$] and no significant difference is found between 2C and 1C.

1.3. Post-lexical contexts

/ʁ/ deletion rates are presented as a function of post-lexical contexts of the word containing /ʁ/ in this section. Due to lack of space, we only present words containing /ʁ/ in word-initial position or in word-initial cluster as a function of their preceding contexts and words containing /ʁ/ in word-final position or in word-final cluster as a function of their following contexts. These two analyses could suggest the deletion pattern of /ʁ/ according to surrounding consonantal or vocalic contexts. Post-lexical pausal contexts are not presented in the following analyses due to lack of occurrences in this context. As mentioned in Section 2, we include categories that have more than 300 occurrences for both vocalic and consonantal contexts in the analyses of this section on post-lexical contexts. According to the same logic on /ʁ/ deletion in long consonant clusters mentioned in section 3.2[4], we

[4] When the preceding post-lexical context is a consonant, it adds an extra consonant to the existing consonant (here /ʁ/) or to the consonant cluster containing /ʁ/.

expect /ʁ/ deletion rate to be higher when the post-lexical context is a consonant (C) than when the post-lexical context is a vowel (V).

1.3.1. Preceding word ending in C or V

In the following, /ʁ/ deletion rates are discussed for words containing /ʁ/ in word-initial position or in word-initial cluster as a function of their preceding contexts: consonantal context (C) *vs* vocalic context (V). As defined earlier, we include #ʁ, #Cʁ and #ʁC only since they contain more than 300 occurrences for C and V post-lexical contexts.

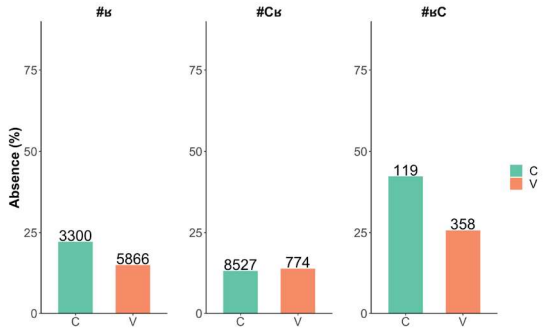

Figure 4: /ʁ/ deletion rates and corresponding number of tokens in word-initial position or in word-initial consonant cluster according to the preceding context of the word (C stands for consonant; V stands for vowel)

Figure 4 illustrates the deletion rate for /ʁ/ in word-initial position or word-initial cluster according to the preceding context of the word. Results show that apart from #Cʁ, which show similar /ʁ/ deletion rate for both the consonantal and the vocalic contexts, /ʁ/ tends to be deleted more when the preceding segment is a consonant, comparing to that observed in the preceding vocalic context. Similar to that observed on within-word positions in Section 3.2, these results also confirmed our hypotheses based on a logic similar to Grammont's well-known "law of three consonants" for schwa in French (Grammont 1894). For schwa, the more consonant we have in a row in the canonical form, the more schwa is realized; however, for /ʁ/, the more consonant we have in a row in the canonical form, the more /ʁ/ is deleted. Moreover, these results suggest that the impact of the number of consonants in a row could influence /ʁ/ both on a within word scenario and on a post-lexical level. Results of the GLMM model (B) suggest that /ʁ/ deletion rate is significantly lower when the preceding post-lexical context is a vowel (V), comparing to that observed in the preceding post-lexical consonantal context [log odds ratio = -0.40808, $|Z| = 15.340$, $p < 0.001$].

1.3.2. Following word starting in C or V

/ʁ/ deletion rates are also analyzed for words containing /ʁ/ in word-final position or in word-final consonant cluster as a function of their following contexts: consonantal context (C) *vs* vocalic context (V). As defined earlier, we include ʁ#, Cʁ#, ʁC# and CCʁ# only since they contain more than 300 occurrences for C and V post-lexical contexts.

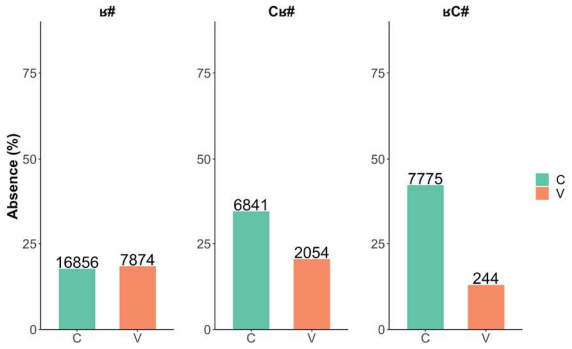

Figure 5: /ʁ/ deletion rates and corresponding number of tokens in word-final position or in word-final consonant cluster according to the following context of the word

Figure 5 illustrates the deletion rate for /ʁ/ in word-final position or in word-final consonant cluster according to the preceding context of the word. Similar pattern for the influence of the following context is found for Cʁ# and ʁC#: when /ʁ/ is in Cʁ# or ʁC# in word-final position, /ʁ/ is more likely to drop if the following context is a consonant than if it is a vowel. It is interesting to note that for ʁ#, /ʁ/ has deletion rates which are quite similar for both post-lexical C and V. This might be due to the fact that ʁ# concerns only one consonant and adding an extra consonant if followed by a consonant (#C) to the existing from (i.e. ʁ#C) would not generate a sequence that violates the sonority sequencing principle. Results of the GLMM model (C) show that /ʁ/ deletion rate is significantly lower when the following post-lexical context is a vowel (V), comparing to that observed in the following post-lexical consonantal context [log odds ratio = -0.23433, |Z|= 15.234, p< 0.001].

1.4. Speech styles

In this section, we present /ʁ/ deletion rates as a function of different segmental contexts according to within word positions (initial *vs* internal *vs* final) and speech styles.

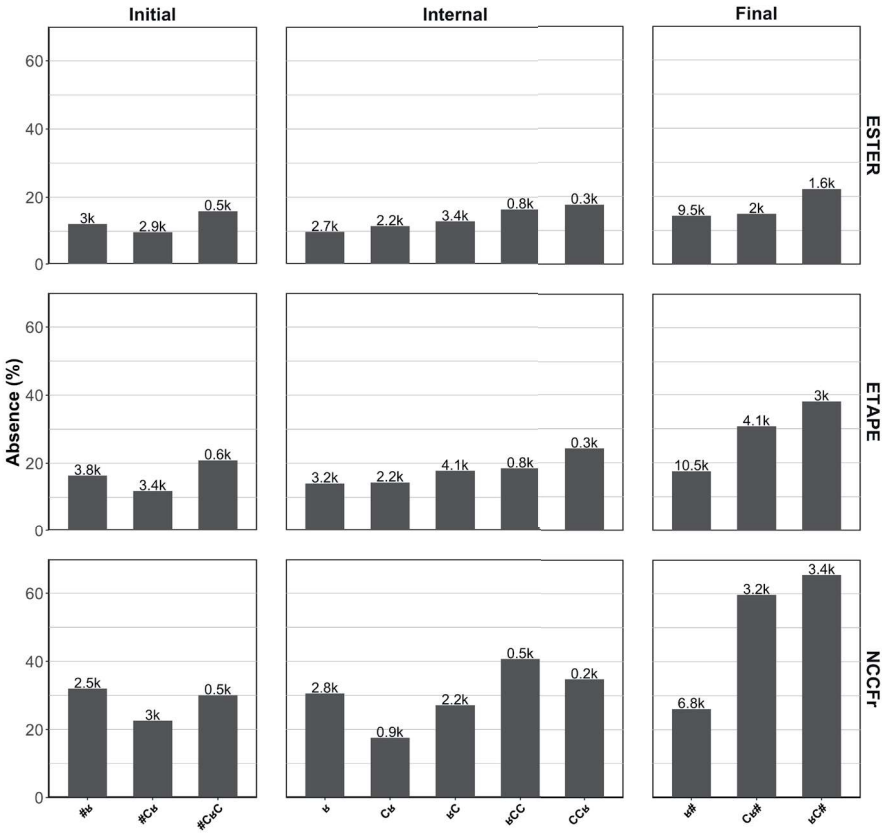

Figure 6: /ʁ/ deletion rates and corresponding number of tokens according to different segmental contexts grouped by word positions (from left to right: initial *vs* internal *vs* final) and speech styles (from top to bottom: ESTER-formal journalistic speech *vs* ETAPE-conversational journalistic speech *vs* NCCFr-casual speech)

Figure 6 shows the percentage of absent /ʁ/ for different segmental contexts as a function of word position and speech style. Similar to Figure 4, segmental contexts containing more than 2k occurrences in the canonical form of /ʁ/ words in all three corpora are shown in this figure. Results show that /ʁ/ is deleted the most in word-final clusters for each of the three corpora. As observed for Figure 4, where all three corpora are pulled, longer consonant clusters tend to have a higher deletion rate than shorter consonant clusters or a single consonant for ESTER (formal journalistic speech) and ETAPE (conversational journalistic speech) corpora. Interestingly, single consonant is observed to have relatively high deletion rate in NCCFr (casual speech). This might be due to the fact that massive reduction

zones are quite frequent in the casual speech corpora. /ʁ/ tends to have similar deletion rates in word-initial and -internal positions. It is interesting to notice that Cʁ achieves the lowest deletion rates (lower than simple /ʁ/ contexts) in word-initial and -internal positions for each of the three corpora. Again, this is not the case for word-final Cʁ#, possibly due to the sonority sequencing principle, as mentioned in the previous section. Based on GLMM model (A), Informal journalistic speech ETAPE [log odds ratio = 0.302215, |Z|= 6.386, $p < 0.001$] and casual speech NCCFr [log odds ratio = 0.968717, |Z|= 14.546, $p < 0.001$] show higher /ʁ/ deletion rate than formal journalistic speech ESTER. Post-hoc results based on the GLMM model show that all pairwise-comparisons of the three corpora are significant ($p < 0.001$).

3. Conclusion

Distribution and deletion of /ʁ/ in different segmental contexts were analyzed in this study using large speech corpora and automatic speech alignment.

Analyses on the word-types and word-tokens of /ʁ/ in different word contexts show that the top seven most frequent forms occupy over 90% (word-types: 93%; word-tokens 91%) of the occurrences of all 20 segmental contexts. The word-types have a fairly balanced distribution (i.e. the frequency for these forms does not vary much from one form to another). Unlike what was observed for the frequency of the word-tokens of the first six forms, the difference between the frequencies varies widely from the most frequent form to the least frequent form. The "ʁ" form is the most frequent form with respect to the word-types, and "ʁ#" is the most frequent form with respect to the word-tokens.

Analyses on the deletion of /ʁ/ according to different segmental contexts and corpora allowed us to identify forms that facilitate /ʁ/ deletion. Interesting patterns can be observed according to within-word positions and post-lexical contexts. /ʁ/ tends to be deleted the most in word-final consonant clusters containing /ʁ/, comparing to what is observed for word-initial and word-internal consonant clusters containing /ʁ/. The number of consonants in the consonantal cluster containing /ʁ/ also has an impact on /ʁ/ deletion. Longer consonant clusters tend to have a higher deletion rate than shorter consonant clusters or a single consonant, in line with what we drew from Grammont's "law of three consonants" for schwa in French. This suggests a similar but opposite tendency expected for schwa: the more consonants there are in the canonical form, the more /ʁ/ is deleted. It is interesting to observe that "Cʁ" achieves the lowest deletion rates (lower than simple "ʁ" contexts) in word-initial and -internal positions. However, this is not the case for word-final "Cʁ#" (which has a very

high deletion rate). This could be related to the fact that word-final "Cʁ#" violates the sonority sequencing principle (Clements 1990) and /ʁ/ is more prone to deletion in this case. The overall patterns found here are also observed for each of the corpora: /ʁ/ is deleted more in longer consonant clusters than in shorter ones and it is more likely to observe /ʁ/ deletion in word-final position/clusters, comparing to that observed in word-initial or -internal position/cluster. It is worth mentioning that the influence of segmental contexts is more salient for the corpora ETAPE and NCCFr than for the corpus ESTER.

Our results on deletion rate for /ʁ/ in word-initial position/cluster according to the preceding context of the word and in word-final position/cluster according to the following context of the word further suggest the influence of the surrounding consonantal or vocalic contexts. /ʁ/ tends to be deleted the most when the preceding/following segment is a consonant, comparing to that observed in post-lexical vocalic context. These results suggest that Grammont's well-known "law of three consonants" for schwa in French (Grammont 1894) could also be generated for /ʁ/ following similar logic, both word-internally and on a post-lexical level: the more consonants there are in the consonantal sequence, the more /ʁ/ is deleted.

In this study, we did not take into account the presence or absence of the epenthetic schwa since some forms require a specific treatment regarding the realization or not of the epenthetic schwa and others do not. However, it would be interesting to select one or two consonant clusters and investigate at the same time the absence/presence of both /ʁ/ and schwa in a future study.

Acknowledgment

This work was supported by the French *Investissements d'Avenir* – Labex EFL program (ANR-10-LABX-0083).

References

Adda-Decker, M. (2006), « De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux », *XXVIes Journées d'Étude sur la Parole*, p. 389-400.

Adda-Decker, M., Lamel, L. (2000), "The use of lexica in automatic speech recognition", in Van Eynde, F., Gibbon, D. (eds), *Lexicon Development for Speech and Language Processing. Text, Speech and Language Technology*, Springer, Dordrecht, p. 235-266, https://doi.org/10.1007/978-94-010-9458-0_8.

Blevins, J. (1995), "The syllable in phonological theory", in Goldsmith, J. (ed.), *Handbook of phonological theory*, Blackwell, p. 206-244.

Brand, S., Ernestus, M. (2015), "Reduction of obstruent-liquid-schwa clusters in casual French", in Wolters, M. *et al.* (eds), *Proceedings of the 18th*

International Congress of Phonetic Sciences (ICPhS 2015), University of Glasgow.

Clements, G. N. (1990), "The role of the sonority cycle in core syllabification", *Papers in laboratory phonology*, 1, p. 283-333.

Cornulier, B. de (1978), « Syllabe et suite de phonèmes en phonologie du français », in Cornulier, B. de, Dell, F. (eds), *Études de phonologie française*, CNRS Editions, Paris, p. 31-69.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J. F., Mostefa, D., Choukri, K. (2006), "Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news", in *Proceedings of LREC 2006*, vol. 6, p. 315-320.

Gauvain, J. L., Lamel, L., Adda, G. (2002), "The LIMSI broadcast news transcription system", *Speech communication*, 37/1, p. 89-108.

Gendrot, C. (2017), "Perception and Production of Word-Final /ʁ/ in French", in *Proceedings of Interspeech*, p. 3926-3930.

Grammont, M. (1894), « Le patois de la Franche-Montagne et en particulier de Damprichard (Franche-Comté). IV: La loi des trois consonnes », *Mémoires de la Société de linguistique de Paris*, 8, p. 53-90.

Grammont, M. (1933), *Traité de phonétique générale*, Delagrave, Paris.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., Galibert, O. (2012), "The ETAPE corpus for the evaluation of speech-based TV content processing in the French language", in *LREC-Eighth international conference on Language Resources and Evaluation*.

Harris, J. (1994), *English sound structure*, Blackwell, Oxford.

Laks, B. (1977), « Contribution empirique à l'analyse socio-différentielle de la chute de /r/ dans les groupes consonantiques finals », *Langue française*, 34, p. 109-125.

McCulloch, C. E., Neuhaus, J. M. (2005), "Generalized linear mixed models", *Encyclopedia of biostatistics*, 4.

Morin, Y. C. (1986), "On the morphologization of word-final consonant deletion in French", in Andersen, H. (ed.), *Sandhi Phenomena in the Languages of Europe*, Mouton de Gruyter, p. 167-210.

New, B., Brysbaert, M., Veronis, J., Pallier, C. (2007), "The use of film subtitles to estimate word frequencies", *Applied psycholinguistics*, 28/4, p. 661-677.

Nyrop, K. (1914), *Manuel phonétique du français parlé*, Gyldendal, Copenhague.

Scheer, T., Ségéral, P. (2008), "Positional factors in Lenition and Fortition", Brandao de Carvalho, J., Scheer, T., Ségéral, P. (eds), *Lenition & fortition*, De Gruyter Mouton, p. 131-172.

Torreira, F., Adda-Decker, M., Ernestus, M. (2010), "The Nijmegen corpus of casual French", *Speech Communication*, 52/3, p. 201-212.

Wioland, F. (1985), *Les structures syllabiques du français: fréquence et distribution des phonèmes consonantiques, contraintes idiomatiques dans les séquences consonantiques*, Editions Slatkine, Genève.

Wu, Y., Gendrot, C., Adda-Decker, M., Fougeron, C. (2019), "Post-consonantal word-final /r/ realization in French: contributions of large corpora", in *19th International Congress of Phonetic Sciences*.