# Liaison and schwa deletion in French : an effect of lexical frequency and competition?

Cécile Fougeron, J P Goldman, U H Frauenfelder

# Liaison and schwa deletion in French : an effect of lexical frequency and competition?

*C. Fougeron°, J.P. Goldman*, U. H. Frauenfelder°*

(*) Lab. de Psycholinguistique Experimentale, Univ. of Geneva, Switzerland
Cecile.Fougeron/ulrich.frauenfelder@pse.unige.ch
(°) Lab. d'Analyse et de Technologie du Langage, Univ. of Geneva, Switzerland
Jean-Philippe.Goldman@lettres.unige.ch

## Abstract

This study aims to determine whether the production of the lexical variants created by the phonological processes of liaison and schwa deletion in French are conditioned by factors linked to lexical recognition. We hypothesise that the realisation of these variants would be favoured for words which are lexically "salient" in term of frequency and in their lexical neighbourhoods. This claim was tested by examining a speech corpus for the effects of lexical frequency, neighbourhood density and neighbourhood frequency on the production of liaison (both in linking and linked words and their co-occurrence) and elision. Overall the results do not support our hypothesis: lexical frequency and competition do not appear to influence strongly whether liaison and elision are realised or not.

## 1. Introduction

Phonetic and phonological variability has been of central interest for psycholinguists trying to explain the mapping and segmentation processes that underlie word recognition. In this study, we are concerned with production variants that are characteristic of French phonology: schwa deletion (henceforth €) and liaison (henceforth £). € refers to the process by which a word containing schwa in its canonical form can be produced without the schwa, for e.g. "samedi" (*Saturday)* can be produced as [saməd**i**] or [samd**i**]. £, on the other hand, is a process by which a final latent consonant may appear in surface for some French words only when the following word begin with a vowel (eg. "grand ami" [grɑ̃tami], *big friend* vs. "grand garçon" [grɑ̃garsɔ̃], *big boy*).

€ and £ have been extensively studied and are known to be conditioned by both linguistic and paralinguistic factors. For example, € occurs more frequently at fast rates, in a casual style, when preceded by a single consonant or by a cluster with increased sonority, and in a late position in the rhythmic group; £ is realised more often between words that are closely related either syntactically or prosodically, in formal speech style and for educated speakers (among others see [1,2,3,4,5]). However, even if these factors can explain some variation in the production of these phenomena, it is still impossible to predict with certitude when € o £ will be realised or not. Indeed, their realisation seem to vary from word to word. In this study we will contribute to this question, by studying how these phenomena are conditioned by lexical factors related to word recognition.

The speech production system has been shown to be conditioned by output-oriented constraints that put pressure on it to accommodate the listener's needs and to maintain sufficient information in the signal for the message to be recovered (e.g. [6]). As a matter of fact, pronunciation variants involving assimilation or vowel reduction have been found to be less frequent when the words are rare, have many competitors, carry a focus or new information, or when the communication situation is impaired by noise [e.g. 7,8].

Since £ or € create pronunciation variants in the surface, it is probable that the realisation of these variants is also constrained to make lexical processing possible. With €, the surface forms differ from the underlying lexical entries by the loss of a segment, the syllabic restructuring that follows and possible assimilation. This resyllabification can occur word internally at different positions (eg. "revenir", *to comeback*: [rə.və.nir][rvə.nir][rə.vnir]) or between words in the case of € in a monosyllabic word (eg. "de famille", *of family*: [dfa.mij]). Thus, the recognition of elided forms implies the restitution of a segment and the recovery of the underlying syllabic structure. In the case of £, word recognition also implies the recovery of lexical boundaries since the liaison consonant is resyllabified with the following vowel initial word (e.g. "grand ami" [grɑ̃.ta.mi]), as well as the processing of the epenthesized liaison consonant. The affiliation of this consonant to the linking word (1st word) may be demanding since this word does not always appear in this form ([grɑ̃] or [grɑ̃t]). This processing relies on several analyses: the identification of the 1st word as a potential linking word, the identification of the 2nd word as a vowel-initial word, and thus the identification of the consonant as a potential liaison consonant and not a word-initial one.

However, while the modelling of lexical access in the cases of € and £ is challenging for current psycholinguistic models, the recognition of elided or linked words does not appear, intuitively, to be particularly difficult.

In this paper we will not address directly this question. Our approach is related but different. We start from the idea that the ease of word recognition is less a function of invariance of surface forms than that of lexical distinctiveness from competing lexical hypotheses. Thus, production variants with elided schwa or liaison may not be recognised with more difficulty if their realisation remains highly distinctive with respect to the other lexical entries. In order to test this hypothesis, we examined in two different speech corpora of read and spontaneous speech whether the realisation of € and £ is function of the word frequency of the items involved and of lexical competition.

## 2. Method

### 2.1. Corpora : design, recording and extraction

Our work is based on the production of 10 native French speakers from Switzerland (6 females and 4 males, aged from 20 to 30). Each speaker was recorded in two conditions : (1) a spontaneous conversation (Spont corpus) of 10 to 17 minutes containing 1570 to 3700 words depending on the speaker; (2) two oral readings of a 1860-word text, first at normal, then at fast rate (Read corpus). Study of rate effect will not be reported here and both readings are combined.

After a manual transcription of the spontaneous speech corpus, possible and realised £ and € were annotated for the 2 corpora. The possible liaison contexts were defined as sequences of words with a 1st word ending with one of the French liaison consonants /t,n,z,r,p/ directly followed by a word starting with a vowel or mute "h". No other linguistic, prosodic or syntactic criteria were considered. Possible € contexts were defined as all words containing a schwa in their phonetic transcription. In the 4th column of table I is given the number of types of words with possible £ and €. In the case of £ in the Read corpus, 229 types of potential linked sequences (e.g. "grand ami") were found and they included 147 types of potential linking words (e.g. "grand") and 112 types of potential linked words (e.g. "ami"). The rate of £ or € realisation was then computed as the % of cases with £ and € for each type of words relative to its number of occurrence in the speech data. As shown in the 5th column of table 1 (range of the number of occurrences for the word types), the number of occurrence greatly varied from item to item (e.g. "le", *the*, appeared 32 times in the text read twice by the 10 speakers (= 640 occurrences), while "petit", *small*, appeared only 2 times in the whole spontaneous corpus). Since the rates of £ and € for each item are usually computed on a larger number of occurrence in the Read corpus than in the Spont corpus, the data for the 2 corpora will be presented separately. In the Spont corpus, we considered only the items appearing at least twice.

*Table 1*: Description of the items found with possible £ and €. (a) refers to the linked sequence (word1+word2), (b) to the linking words (word1), (c) to the linked words (word2).

| | corpus | | # of word types | <range> of occurrences (median) | Rate of realisation |
|---|---|---|---|---|---|
| £ | Read | a | 229 | <15-80> (20) | 40 % |
| | | b | 147 | <15-219> (20) | 23 % |
| | | c | 112 | <17-564> (20) | 60 % |
| | Spont | a | 240 | <2-116> (3) | 31 % |
| | | b | 161 | <2-258> (3) | 19 % |
| | | c | 113 | <2-155> (4) | 48 % |
| € | Read | | 79 | <19-1978> (20) | 27 % |
| | Spont | | 109 | <2-543> (3) | 64 % |

### 2.2. Factors studied

Lexical frequency factor. An estimation of *lexical frequency* of the items involved in £ (linking and linked words) and € (elided word) was calculated as the frequency of occurrence of these words in an external database of 12Mo words (composed mainly of press articles). A measure of frequency of co-occurrence of the linked sequence was calculated on the same database as the number of occurrence of the linking and linked word in context without intervening punctuation marks. These frequency measures have been transformed on a logarithmic scale for linearization.

Lexical competition factors. Different factors have been used to characterise each of the items involved in £ or €, in term of salience among its nearest competitors. For this, two types of competitor neighbourhoods have been defined.

- First, using the single phoneme substitution method we considered as nearest neighbours for the potential linking, linked or elided word, the words differing from the target word by a single phoneme (whatever its position). We call these "similar neighbours".
- Second, in the case of the linked words another type of competitors has to be considered. As said before, recognition of the 2nd word in a linked sequence could be seriously impaired if the liaison consonant is misinterpreted as a word initial consonant For example, the sequence [grɑ̃tami] could be segmented as [grɑ̃t # ami] ("grand ami" *big friend*) or [grɑ̃ # tami] ("grand tamis" *big sieve*). Such a complete ambiguity between the two lexical hypotheses for the 2nd word do not occur very often, but one have to consider the case of partial ambiguity when the linked word competes with words beginning with a consonant similar to the liaison consonant. This competitors will be referred to as "overlapping neighbours".

Thus, 4 factors were used as a measure of lexical competition:

(a) the *similarity neighbourhood density* gives for each item the number of phonologically similar words in the language.

(b) the target *relative frequency* gives the frequency of the target word relative to the summed frequency of its nearest phonological neighbours.

(c) the *overlapping neighbourhood density* calculated for each linked words gives the number of competitors beginning with a sequence of phoneme similar to the "liaison consonant+1st vowel of word2".

(d) the *relative frequency* of the linked word compared to the summed frequency of its overlapping neighbours.

In order to compute these four factors, we used Brulex, a lexical database of 36000 French words including their lexical frequency computed from a sample of texts of 23.5Mo words [9]. Since this database contains only lexemes forms, all the inflected forms of our corpora have been manually lemmatised. The drawback of this procedure is that the search for competitors is made on lexical roots that sometimes differ greatly from the inflected form (e.g. for conjugated verbs transformed to infinitive forms).

Salient words relative to their neighbours are considered to be words with low density similarity neighbourhoods and with high relative frequency (for e.g. [8]).

## 3. Results

### 3.1. Lexical frequency factor

#### *3.1.1. Frequency of linking word*

There is a significant positive correlation between the frequency of the linking word and the rate of £ for both the read corpus (r=.61, p<.0001) and the spontaneous corpus (.44, p<.0001) : there are more cases of £ with more frequent linking words. However, as previously shown in the literature (e.g. [2,10]), £ is known to be realised more often with short linking words and function words, both of which are highly frequent. In order to analyse this in greater depth, we computed the rate of £ for intervals of increasing frequency containing each 10% of the total amount of linking words, as presented in figure 1. Examination of the items included in these intervals shows that monosyllabic function words make up most (about 80%) of the frequent linking words (0-20% range or A&B intervals in the figure) in both corpora, thereby introducing length and category confounds there, where the strongest frequency effects are found. Thus a better measure of the effect of word frequency upon £ is provided by the less frequent intervals, where there is more variety in word length and syntactic category. In this comparison, the effects of frequency are relatively small: the least frequent words are almost never involved in £, while words of intermediate frequency are slightly more often (but not exceeding 40 % of realisation). In sum, it appears that

observed lexical frequency effect is mainly due to other lexical characteristics of the linking words such as word length or grammatical category.

*3.1.2. Frequency of linked word*

A different trend appears for the frequency of the linked word. The rate of £ seems to decrease with increased frequency of the linked word. This negative correlation is significant in both corpora but is relatively weak (r= -.2, p=.03 Read, r= -.2, p=.02 Spont). Again, it has to be noted that the ranges including the 20% most frequent words (intervals A & B) on figure 1 (circles), are mostly composed of short function words (~75% of the items). For the other frequency ranges, no lexical frequency effect is found.

*Figure 1-2-3 :*Rate of £ (%) by intervals containing 10% of the items, ordered by lexical frequency (fig 1), by frequency of co-occurrence (fig.2), and by similarity neighbourhood density (fig. 3). Linking words are shown with squares, linked words with circles, Read corpus in bold line and Spont corpus in dotted line. Intervals not presented contain items that do not appear in Brulex and for which calculation of the factors could not be made.

***Fig. 1 : % of liaison according to lexical frequency***

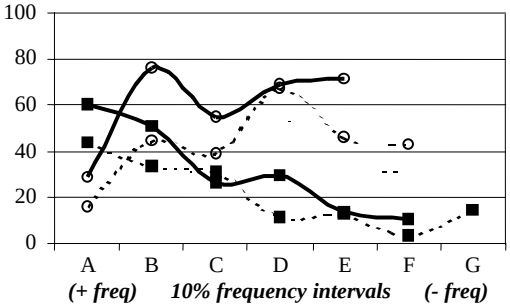



***Fig. 2 : % of liaison according to frequency of co-occurence***

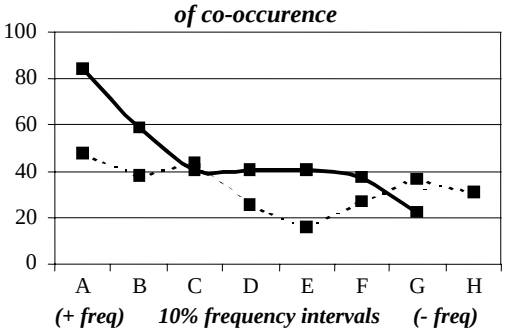
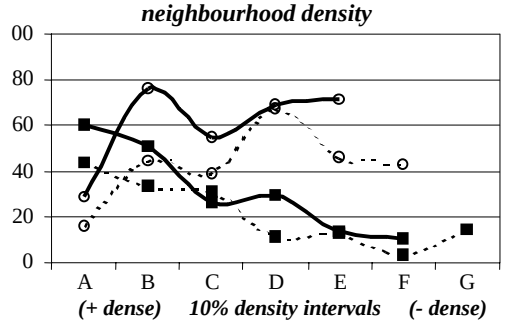



***Fig. 3 : % of liaison according to similarity neighbourhood density***




*3.1.3. Frequency of co-occurrence of linking & linked words*

In the Read corpus a significant positive correlation is found between the rate of £ and the frequency of occurrence of the sequence linking+linked words (r=.39, p<.0001). Compared to the average rate of £ over the whole list of co-occurrences (40%), the items included in the first 10% most frequent co-occurrence undergo 2 times more £ (84%), as illustrated in figure 2 (A, bold line). These sequences are short (2-3 syllables), mostly made up of function words (e.g. "dans un" *in a*, "on a" *one has*), and not only obligatory liaison (60% facultative liaisons, e.g "mais aussi" *but also*, "sont en" *are in*). Interestingly, only a few determinant+noun sequences (15%) occur in this range or in the next (A-B). In the Spont. corpus (dotted line in figure 2), the relation between the rate of £ and the co-occurrence frequency is also significant, but much less striking (r=.16, p=.02).

*3.1.4. Frequency of elided word*

No effect of lexical frequency is found on the rate of € (r=.18, ns Read, r=.001, ns Spont). Frequent words such as clitics are very often elided in spontaneous speech, while they are not in read speech. For the frequent content words, rate of € appears to vary from item to item. When looking at words with schwa in the first syllable as in [3,5], no effect of lexical frequency is found either (r=.21, ns Read, r=.07, ns Spont).

**3.2. Lexical competition factors**

According to most psycholinguistic models, the recognition of a word should be easier when it has few competitors and when it is more frequent than these competitors. Similarly, we hypothesise that the pronunciation variants of these words should also be more easy to recognise. Thus, words in sparse neighbourhood and with large relative frequency should allow a greater rate of £ and €.

*3.2.1. Similarity neighbourhood density and relative frequency of linking and linked words*

Surprisingly, as shown in figure 3 (squares), linking words in denser neighbourhood tend to have a higher rate of £ than such words in sparse neighbourhoods. Correlation between the number of neighbours and rate of £ is relatively strong and significant, particularly in the Read corpus (r=.48, p<.0001 Read; r=.32, p<.0001 Spont). This finding goes against our hypothesis, but it is not surprising given the properties of the lexicon. Indeed, the items included in the ranges A and B (items with the densest neighbourhoods) are monosyllabic or even mono-phonemic words. As previously said, these words are also the most frequent words and words that have the highest rate of £ (see 3.1.1).

When looking at relative frequency, the results do not confirm our hypothesis either: linking words with high relative frequency (salient words) do not have a higher rate of £. We even find an opposite tendency when looking at the words with the smallest relative frequency (50% least frequent, not illustrated here) which have the highest rate of £ (~50%). These items include words like "dans, défis, gros" (*in, challenge, thick*). Globally, similarity neighbourhood density is not correlated with rate of £ in both corpora (r=-.05, ns Read, r=.07, ns Spont).

Concerning the linked words, no correlation is found between rate of £ and neighbourhood density (r=-.13, ns Read, r=.02, ns Spont), or relative frequency (r=-.02, ns Read, r=.04, ns Spont).

*3.2.2. Overlapping neighbourhood density and relative frequency for linked word*

Overlapping neighbours are the words competing with the linked word in the case of a wrong segmentation of the liaison consonant as a word initial consonant. We hypothesised that £ creating ambiguities between several competing hypotheses for the recognition of the 2nd word (i.e. when the linked word has several overlapping competitors starting with a sequence similar to the "liaison consonant+1stvowel"), should be less frequent. However, the results do not confirm this hypothesis.

Overlapping neighbourhood density is not correlated with rate of £ in both corpora (r=-.06, ns for Read, r=.2, p=.02 for Spont). Relative frequency of the linked word compared to its overlapping neighbours do not affect the rate of £ in the expected direction either (r=.23, p=.01 for Read, r=.16, ns for Spont). In fact, when looking at some of the linked words with small relative frequency in the 70-80% interval (not illustrated), our hypothesis is contradicted: they show the highest rate of £ (~70%). These items include words like "arriver, océan, introuvable, envie" (*arrive, ocean, unfoundable, envy*).

*3.2.3. Similarity neighbourhood density and relative frequency of elided word*

As for lexical frequency, the rate of € does not seem to be conditioned by competition factors. € is as frequent for words in sparse neighbourhood than dense neighbourhood (r=.22, p=.05 Read, r=.02, ns Spont), and is independent of relative frequency (r=.07, ns Read, r=.17, ns Spont).

## 4. Discussion and conclusion

The extent to which € and £ are realised in speech is known to be highly variable from token to token and from context to context. This study took as its point of departure the idea that, as for other phonetic variants, the production of phonological variants created by € and £ processes in French is a function of lexical factors affecting word recognition. We hypothesised that by producing € and £ in certain cases but not in others, speakers respect listeners' communicative needs by avoiding the production of variants for words with low frequency and in dense neighbourhoods. Overall, this hypothesis was not confirmed in our data.

As for €, we found no clear effect of lexical frequency or factors linked to lexical competition. This result goes against previous findings [3,5] where words with syllable-initial schwa (e.g. "fenêtre" *window*) were found to be more often elided when frequent. In our corpus, the effect of lexical frequency was not replicated on this kind of words. Thus our results, like others [10], suggest that when a large variety of potentially elided words is considered, lexical frequency is not a main determinant of the production of € .

The realisation of £ cannot be explained by factors affecting the ease of recognition of the linked word (word2) either. Indeed, £ is found to be frequently realised with "non salient" linked words, i.e. words that are rare, are in a dense neighbourhood and do not stand out in this neighbourhood. Furthermore, the existence of strong lexical competitors resulting from the incorrect segmentation of the liaison consonant as a word initial, does not appear to hinder the realisation of £.

The only case where an effect of lexical frequency is found is for linking words. However, since word frequency is highly correlated with other morpho-syntactic and rhythmic properties, it is not possible to interpret this effect in terms of the ease of word recognition. On the other hand, £ occurs more often with words that co-occur frequently. £ realisation may not hinder lexical recognition if these word sequences are lexicalized in their linked form, or if their frequency of co-occurrence makes the presence of £ highly predictable.

In conclusion, based on an analysis of the factors studied here, we cannot affirm that £ or € are produced more frequently on salient words for which recognition of variants should be easier. Compared with other production variants, the lack of an influence of lexical frequency and competition factors may be due to the fact that linked and elided surface forms of words may be lexicalized, and that their recognition can be facilitated by several predictive factors, such as style, syntactic, prosodic and semantic context. Moreover, these results raise the question of whether these variants really lead to a greater cost in word processing. First, since £ between words is known to reinforce syntactic and prosodic integration; it is surprising that this function would be done at the cost of word recognition. Second, € creates words with unusual (and often phonotactically incorrect, e.g. [pti] *small*) clusters, that would makes this form (if lexicalized) totally salient. Another possibility is that even if there were a processing cost associated with these phonological variants, the production system does not take this fact into account. Future research combining experimentation with corpus-based analyses should help clarify the question of the relation between production and perception.

## 5. References


[1] Lucci,V. *Etude phonétique du français contemporain à travers la variation situationnelle*. U. de Grenoble Publisher, 1983.

[2] Agren, J. "Etude sur quelques liaisons facultatives dans le français de conversation radiophonique : fréquence et facteurs". *Acta Universitatis Upsaliensis*, 10, 1973.

[3] Hansen, A.B. "Etude du E caduc…". *J. of French Language Studies*,4 ,25-54, 1994.

[4] Dell, F. *Les règles et les sons*, Paris:Hermann, 1973.

[5] Racine, I & Grosjean F. "La fenêtre ou la f'nêtre: le E caduc facultatif l'est-il réellement?", submitted.

[6] Lindblom, B "Explaining phonetic variation: A sketch of the H&H theory". Hardcastle et al. *Speech Production and Speech Modelling*, 403-439, Kluwer, 1990

[7] Cooper, W.E. & Paccia Cooper J. *Syntax and speech*. Cambridge Mass.: Harvard U. Press, 1980

[8] Wright, R. "Factors of lexical competition in vowel articulation", to appear in Labphon6, 1998

[9] Content A., Mousty P., Radeau M. "Brulex…" L'Année Psycholgique, 90, 551-566, 1990.

[10] Adda-Decker M & al. "Pronunciation Variants in French: schwa & liaison", *ICPhS 99*, 2239-2242, 1999


## Acknowledgements:


We wish to thank J. Golsin, A. Dard, C. Jeager, L. Guélat for their help. The 1st author is supported by grant 1114-059532 from the Fond National de Recherche Suisse, and the 2nd author by the CTI project n°4607.1 and the Projet Plurifacultaire "Prosodie" of the University of Geneva.