Liaisons in French: a corpus-based study using morpho-syntactic information

Philippe Boula de Mareüil◊, Martine Adda-Decker◊ & Véronique Gendner◊+
◊ *LIMSI-CNRS, Orsay, France*
+ *LATTICE, Université Paris VII, France*

ABSTRACT

French liaison consists in producing a normally mute consonant before a word starting with a vowel. Whereas the general context for liaison is relatively straightforward to describe, actual occurrences are difficult to predict. In the present work, we quantitatively examine the productivity of 20 liaison rules or so described in the literature, such as the rule which states that after a determiner, liaison is compulsory. To do so, we used the French BREF corpus of read newspaper speech (100 hours), automatically tagged with morpho-syntactic information. The speech corpus has been automatically aligned using a pronunciation dictionary which includes liaisons. There are 90k liaison contexts in the corpus, about half of which (45%) are realised. A better knowledge of liaison production is of particular interest for pronunciation modelling in automatic speech recognition, for text-to-speech synthesis, descriptive phonetics and second language acquisition.

1 INTRODUCTION

French liaison consists in producing a normally mute consonant before a word starting with a vowel, a mute *h* or some glides. However, this general description of the context of liaison does not allow us to predict actual occurrences. Traditional accounts of liaison in French, mainly found in orthoepic textbooks, distinguish between liaisons that are termed obligatory or compulsory, those that are referred to as optional or variable, and those that are described as forbidden, erratic or impossible (*cuirs, velours, pataquès*). It is noteworthy that a number of set phrases belong to one class or another. But these domains depend on a range of stylistic, socio-linguistic and situational factors. Possible liaisons may be mandatory in a poetic diction or in a theatrical style, whereas in a colloquial use they sound shocking [5]. Furthermore, this is liable to change over time [10]. For example, in his revision of Delattre's [4] classification, Encrevé [6] ranks monosyllabic adverbs and prepositions in the variable category. Although this is not clear-cut, the linguistic resources and tools we dispose of may contribute to establish a new classification. They enable us to obtain a more precise picture of current practice, by producing the percentage realisation of liaisons in a large speech corpus, broken down by syntactic contexts.

In an earlier study [2], we described the occurrences of liaison with respect to word frequency: an important correlation could be found between liaison production and lexical frequency, but we underlined the importance of relating liaison to morpho-syntactic information.

Liaisons are of particular importance for pronunciation modelling in automatic speech processing. They allow for a variable number of phonemes and can thus be considered as *sequential variants*. As far as speech recognition is concerned, if liaison is not properly accounted for, then recognition errors are likely to occur. Liaisons can be represented either directly in the lexicon, as phonological rules, or implicitly in the acoustic models. The first option is generally adopted, but a straightforward solution which consists of adding optional liaison phonemes to all applicable words has proven ineffective. Recognition error rates did not reduce: the large number of variants introduced additional homophone sequences and introduced new errors.

The aim of next sections is to increase our knowledge of actually observed liaisons in a large corpus. In the present work, we quantitatively examine the productivity of roughly 20 liaison rules described in the literature.

2 DESCRIPTION OF LIAISON RULES

As mentioned earlier, the *liaison* phenomenon consists in the realisation of a normally mute final consonant in the context of a following word which begins with a vowel. A simple example is the word sequence *les enfants* ("the children", pronounced in isolation as /le/ and /ɑ̃fɑ̃/), which has to be pronounced /le**z**ɑ̃fɑ̃/: /z/ is the liaison consonant, which is used as the onset of the following syllable. Liaison should not be confounded with chaining and elision phenomena. The former concern normally pronounced consonants [7]. Liaisons without chaining can even be heard, particularly in political debates [6]. Liaison should also be distinguished from elision, which suppresses a vowel. Moreover, a limited number of consonants are used for liaison: /z/, /t/, /n/, /ʀ/, /p/ – the rank order is from highest to lowest for the frequency of occurrences. Yet, in the three cases, the consonant which terminates the first element generally belongs to the initial syllable of the following word, which may make the word boundary recognition more difficult.

How and when is liaison made? We are here in a delicate field, and there is no consensus to answer this question. French liaisons have been studied in [4, 7, 6]. Rather than assessing the accuracy of one of these contributions,

ISBN 1-876346-48-5 © 2003 UAB

we have chosen to compile them, in order to investigate rules describing so-called compulsory (see Tab. 1), forbidden (Tab. 2) or optional liaisons (Tab. 3) In the liaison rules we are going to examine the set of words likely to produce a liaison consonant is expressed, together with a specification of the right context words; the "+" sign delimits the two. One of the contexts may be empty if there is no condition on the preceding or following words. Patterns thus may vary from open to one single word, but correspond most often to part-of-speech (PoS) tags with morpho-syntactic information. In the first 7 rules, for instance, there is a close grammatical link between the words or parts-of-speech: within a noun phrase (see rules 1 and 2) or within a verb phrase especially. As for rule 8, it is an example of pattern limited to one word with unspecified right context. In sum, liaison should be made in the contexts displayed in Tab. 1.

Liaison rules

#	**Rule pattern**	**Example**
1	determiner +	*les‿uns*
2	adjective + noun	*un gros‿arbre*
3	monosyllabic adverb other than *pas* ("not") +	*tant‿en ville qu'à* *la campagne*
4	verb + pronoun	*sort‿-il*
5	clitic pronoun +	*ce dont‿on parle*
6	aux. verb, 3rd person +	*il est‿évident*
7	monosyllabic preposition +	*en‿avance*
8	*quand* ("when") +	*quand‿il vient*

Table 1: Morpho-syntactic patterns with compulsory liaison.

Liaison is generally avoided in the contexts displayed in Tab. 2. At last, liaison is typically described as optional in the contexts shown in Tab. 3.

No liaison rules

<table>
<tr><th>#</th><th>Rule pattern</th><th>Example</th></tr>
<tr><td>9</td><td>non clitic pronoun +</td><td><i>où sont-ils</i> | <i>allés</i></td></tr>
<tr><td>10</td><td>main verb +</td><td><i>tu perds</i> | <i>un temps</i></td></tr>
<tr><td>11</td><td>sing. common noun +</td><td><i>un soldat</i> | <i>anglais</i></td></tr>
<tr><td>12</td><td>polysyll. adv./conj./prep. +</td><td><i>tantôt</i> | <i>ici</i></td></tr>
<tr><td>13</td><td><i>et</i> ("and") +</td><td><i>vingt et</i> | <i>un</i></td></tr>
<tr><td>14</td><td>adjective + ¬ noun</td><td><i>bon</i> | <i>à rien</i></td></tr>
</table>

Table 2: Morpho-syntactic patterns with prohibited liaison (¬noun denotes a word other than a noun).

Optional liaison rules

#	**Rule pattern**	**Example**
15	plural noun + plural adjective	*jours_heureux*
16	*pas* ("not") +	*pas_encore*
17	participle +	*faisant_ainsi*
18	*mais* ("but") +	*mais_enfin*

Table 3: Morpho-syntactic patterns with optional liaison.

3 METHODOLOGY & EXPERIMENTAL CONDITIONS

This study makes use of the BREF corpus [11] of read speech. The data contain 66,500 sentences read by 120 speakers. In the corresponding 26,000 word list, over 25% have possible liaisons, which gives an idea of the phenomenon magnitude.

We define the $\mathcal{L}_p$ corpus (potential liaison corpus) as the set of word sequences of BREF with a potential liaison: a word with a liaison consonant followed by a word starting with a vowel or a glide – therefore, the term "potential" in this sense has nothing to do with "optional" liaisons defined in Section 2. These word sequences are also referred to as liaison contexts: we measured 91,126 occurrences of liaison contexts in the BREF corpus.

We then define the $\mathcal{L}_o$ corpus (observed liaison corpus) as the set of word sequences (liaison contexts) where a liaison is effectively observed. This corpus contains 40,940 liaison occurrences, which gives a global liaison rate of 45%.

In the following, we will measure the relative weight of each rule in both the $\mathcal{L}_p$ and the $\mathcal{L}_o$ corpora. The first percentage $\%\mathcal{L}_p$ indicates whether the rule under consideration is followed or not. The second percentage $\%\mathcal{L}_o$ reflects the contribution or in other terms the relative importance of a given rule with respect to the general liaison phenomenon.

3.1 Acoustic-phonetic alignment

The acoustic phone models are sets of continuous density hidden Markov models (HMMs) with Gaussian mixture. Context-dependent phone models are used to account for allophonic variation observed in different contexts. In order to determine the sequence of realised phones in a given utterance, a Markov chain is formed by concatenating the phone pronunciations associated with the words in the corresponding orthographic transcription. This is then used to constrain the search space for the decoder, aligning the phones with the speech signal. If pronunciation variants are represented in the lexicon or added by phonological rules, a phone graph is constructed and aligned with the signal. In this case, the decoder will produce the most likely sequence of phones along with the time alignment. The LIMSI system was used, the accuracy of which was demonstrated by a series of evaluations [9].

3.2 Pronunciation lexicon

The pronunciations and their variants were generated by a grapheme-to-phoneme converter, into which pronunciation variants were introduced. The program GRAPHON+ [3] was used, whose word error rate on a 30,000 word running text is less than 1%. For the purpose of this study, liaisons are allowed for all words ending with *-s, -x, -z, -n, -d, -t, -r, -p*. In our broad phonetic description, there is no mark for the "disjunctive *h*" which is assumed to prohibit rigth-to-left liaisons as in *les* | *héros* ("the heroes") vs *les‿hommes* ("the men"). We will discuss this further in 4.2.

3.3 Morpho-syntactic tagging

The morpho-syntactic information was produced by the CORDIAL tagger retailed by Synapse Développement (Toulouse). It had to be aligned with the word tokens as used by the speech recognition (alignment) system. Now, though there is a straightforward correspondence for most items, tokenisation problems arise especially for numerals

PoS bigram		#occurr.	%data	part of
Det	Noun	14077	15.3	rule 1
Verb	Prep	6242	6.8	rules 6, 10
Verb	Det	5797	6.3	rules 6, 10
Verb	Verb	5225	5.7	rules 6, 10
Noun	Adj	4627	5.0	rule 11
Noun	Verb	4609	5.0	rule 11
Noun	Prep	4266	4.7	rule 11
Noun	Conj	4217	4.6	rule 11
Pron	Verb	3430	3.7	rule 5
Verb	Adv	3101	3.4	rules 6, 10
Prep	Noun	2383	2.6	rules 7, 13
Num	Noun	1915	2.1	rule 2
Adv	Adj	1864	2.0	rule 3
Adv	Verb	1771	1.9	rule 3
Adj	Num	1733	1.9	rule 2 (entirely)

Table 4: 15 most frequent PoS sequences in the BREF corpus (accounting for 65% of the corpus).

(e.g. a date like *1984*), acronyms and some compounds or idioms. Therefore, tokenisation had to be fitted to match the output of the speech processing system.

The CORDIAL tagset is very close to the one used in a series of evaluation campaigns, inspired by MULTEXT, GRACE [1]. In the current state-of-the-art, the error rate on words is about 3%, which enables quite a reliable analysis.

Looking at the PoS bigrams corresponding to the $\mathcal{L}_p$ corpus (potential liaisons), the 50 most frequent PoS bigrams account for about 98.6% of the liaison corpus – in accord with Zipf's law. The 15 most frequent PoS sequences and their links with the liaison rules defined above are given in Tab. 4. Some results on the occurrences of liaisons are also provided. But it is in next section that rule-specific experimental results will be presented.

4 EXPERIMENTAL RESULTS

4.1 Compulsory, forbidden and optional liaisons

For each of the rules described above, Tables 5, 6 and 7 give the number of occurrences in a potential liaison context, the percentage of occurrences with an observed liaison for the given rule (%$\mathcal{L}_p$), and the percentage these liaisons represent in the set of all observed liaisons (%$\mathcal{L}_o$).

We can observe that rules 1-8 (liaison rules) have all %$\mathcal{L}_p$ rates over 70%. Concerning rule 3 (which applies to 18 different monosyllabic adverbs), this rate is observed only after excluding the negation *pas* ("not") from the rule-specific liaison contexts. Indeed (see Tables 3 and 7), it appears that liaison after *pas* is rather optional (40%). The same 40% rate can be observed for *moins* ("less"); and if 94% of liaisons are realised with *très* ("very"), a 0% liaison can be measured for *loin* ("far"). Therefore, liaison seems to be far from compulsory. This strengthens Encrevé's [6] classification which proposed this liaison as optional.

Rules 4 and 5, which apply to pronouns such as *en, on, ils, elles* ("we, they") – preceded by a dash in rule 4 – are more respected than the previous one. So is rule 6, in which forms of *être* ("to be") and *avoir* ("to have") were included even if they were tagged as main verbs.

Liaison rules				
#	**Pattern**	**#occ.**	**%$\mathcal{L}_p$**	**%$\mathcal{L}_o$**
1	**det. +**	**15272**	**95.3**	**35.6**
2	adj. + noun	1733	72.5	3.1
3	monosyll. adv. $\neg$*pas* +	2570	70.7	4.4
4	verb + pronoun	1081	99.2	2.6
5	clitic pronoun +	4534	83.3	9.2
6	aux. verb +	6997	81.0	13.6
7	monosyll. prep. +	4269	90.5	9.4
8	*quand* +	168	94.6	0.4

Table 5: Liaison rules with their number of liaison contexts in the $\mathcal{C}_p$ corpus. The last two columns indicate the liaison ($\mathcal{L}_p$) rate for each rule and the percentage these liaisons represent in the effectively observed liaison corpus (%$\mathcal{L}_o$).

No liaison rules				
#	**Pattern**	**#occ.**	**%$\mathcal{L}_p$**	**%$\mathcal{L}_o$**
9	non clitic pronoun +	133	1.5	0.0
10	main verb + $\neg$ clit. pron.	254	5.1	0.0
11	sing. common noun	11882	10.4	3.0
12	polysyll. FW +	4336	5.4	10.6
13	*et* +	4004	1.1	9.8
14	adj. + $\neg$noun	4599	5.9	11.2

Table 6: Morpho-syntactic patterns with prohibited liaison.

Likewise in rule 10, these forms and semi-auxiliary verbs were excluded from main verbs. A constraint was also added to the right context, to avoid an intersection with rule 4. The status of the latter is moreover questionable, since an orthographic clue imposes liaison (here observed in over 99% of instances). Still, the rule remains violated whether the verb is in the singular or in the plural, in the following cases: *font irruption* ("burst in"), *commencent aussi* ("begin too"), *vient alors* ("comes then").

Another no liaison rule is rule 12, whose liaison rate can be compared to those of rules 3 and 7. Rule 12 states that liaison is prohibited after polysyllabic adverbs, conjunctions and prepositions. Although the overall liaison rate is very low (5.4%), liaison may be relatively frequent with some word sequences. Interstingly, the word sequence *après avoir* ("after having") occurs with 40% of actual liaison; in *devant eux* ("before them"), liaison is twice as frequent as no liaison. For *devant elle* ("before her"), liaison is even always observed. Putative forbidden liaisons are all realised

Optional liaison rules				
#	**Pattern**	**#occ.**	**%$\mathcal{L}_p$**	**%$\mathcal{L}_o$**
15	plur. noun + plur. adj.	3384	28.7	2.3
16	*pas* + +	1595	41.0	1.6
17	participle +	786	14.1	0.3
18	*mais* +	634	44.0	0.7

Table 7: Morpho-syntactic patterns with optional liaison.

in 10% of occurrences or less. But again, this highlights that, beyond morpho-syntax, word identity exerts a strong influence. This is exemplified by the case of *mais* ("but") in Tab. 7 (rule 18 vs 13) and will be all the more obvious in next subsection.

4.2 Focus on rule 1 (determiner +)

The first rule which states that, after a determiner, liaison is compulsory has been observed in 95%, and these liaisons represent more than a third of all observed liaisons. This is the reason why we focus on this rule in Tables 8 and 9: these tables provide breakdowns by PoS and subpatterns where liaison is generally omitted. Not surprisingly in Tab. 8 (see also Tab. 4), the sequence determiner + noun accounts for a large number of liaisons.

Rule 1	#occ.	$\%\mathcal{L}_p$	**Example**
det. +	15272	95.3	
det. + noun	14077	96.6	*son image*
det. + pron.	376	98.7	*les uns*
det. + adj.	713	95.5	*un obscur*
det. + verb*	33	60.6	*des élus*
det. + num.	57	4.2	*les un virgule*
det. + other	16	-	

Table 8: Rule 1 breakdowns by PoS. (* generally substantivised forms of verbs in the past participle).

When looking for the rule 1 pattern in the $\mathcal{L}_p$ corpus, we observe that most of the liaisons which are not realised stem from words starting with a disjunctive *h*: e.g. *hasard* ("chance"), *hautes* ("high"), *Hongrois* ("Hungarians"). Also, liaison is often avoided with numerals (see Tab. 8), loan words (especially those which begin with a glide) and acronyms – in particular those which start with a graphemic consonant (e.g. *HLM, SVT*), even though their spelled pronunciation starts with a phonemic vowel.

Subpattern	**Example**	**#occ.**	$\%\mathcal{L}_o$
det. + *h*-start	*l/des hasards*	48	0
det. + acronym	*les RPR*	12	0

Table 9: Examples of rule 1 patterns where liaison is generally avoided.

5 DISCUSSION

The yielded results would merit checking at two levels: that of PoS tagging to verify if they fulfil the rules we requested, and that of speech, which requires hours of listening. Only subsets of the data were listened to, so as to test the alignment. The figures presented in the paper were not changed, since we cannot control whether correcting some errors would not introduce new ones. Nonetheless, the presented study of French liaisons is to our knowledge the first automatic investigation of this phenomenon in a large spoken corpus. The measures obtained confirm most of *a priori* linguistic predictions, and allow a ranking of the rules proposed in the literature. Most important is the rule concerning liaisons with determiners, which contributes to more than one third of all observed occurrences. Other important rules are the one for clitic pronouns, the one concerning *être* and *avoir* irrespective of their auxiliary or main verb status, and the rule for monosyllabic prepositions – each of these 3 rules accounts for about 10% of all observed liaisons. The obtained results favour the classification of monosyllabic adverbs in the optional liaison category, partially in accordance with Encrevé's viewpoint [6].

Liaison rates vary significantly depending on word and word bigram identities. This confirms that liaison is strongly linked to both syntactic and lexical levels [8].

A better knowledge of liaison production is of particular interest for descriptive phonetics, second language acquisition and speech processing. An accurate modelling of liaison phenomena may also contribute to a better structuring of the speech flow. Words which are connected by a liaison are acoustically marked as belonging to a larger scale unit: like prosody, liaison is an indicator of between-word juncture. This may open new fields of investigation relating phonetic and semantic structures for speech understanding.

REFERENCES

[1] Adda, G., Mariani, J., Paroubek, P., Rajman, M., Lecomte, J., "Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français", *TALN*, Cargèse (pp. 15-24), 1999.

[2] Adda-Decker, M., Boula de Mareüil, P., Lamel, L., "Pronunciation variants in French: schwa & liaison", *ICPhS*, San Francisco (pp. 2239-2242), 1999.

[3] Boula de Mareüil, P., *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, PhD Thesis, Université de Paris XI, 1997.

[4] Delattre, P., *Principes de phonétique française à l'usage des étudiants anglo-américains*, Middlebury College, 1951.

[5] Eggs, E. & Mordellet, I., *Phonétique et phonologie du français*, Niemeyer Verlag, Tübingen, 1990.

[6] Encrevé, P., *La liaison avec et sans enchaînement. Phonologie tridimensionnelle et usages du français*, Éditions du Seuil, Paris, 1988.

[7] Fouché, P., *Traité de prononciation française*, Éditions Klincksieck, Paris, 1969.

[8] Fougeron, C., Goldman, J.-P., Frauenfelder, U.H., "Liaison and schwa deletion in French: an effect of lexical frequency and competition", *Eurospeech*, Aalborg (pp. 639-642), 2001.

[9] Gauvain, J.-L. *et al.*, "The LIMSI 1998 HUB-4e transcription system", *DARPA Broadcast News Workshop*, Herndon (pp. 99-104), 1999.

[10] Hintze, M.-A., Pooley, T., Judge, A. (eds), *French accents: phonological and sociolinguistic perspectives*, AFLS/CiLT, London, 2000.

[11] Lamel, L.F., Gauvain, J.-L., Eskénazi, M., "BREF, a Large Vocabulary Spoken Corpus for French," *Eurospeech*, Genova (pp. 506-508), 1991.