

Prediction of liaison in French by Measures of Information Theory

THOMAS CULETTO
University of Oxford
thomas@culetto.net

ABSTRACT. Using measures of Information Theory, I first generalise Zelig Harris's (1955) morphological segmentation approach to include Parts of Speech, and then apply the generalized approach to French liaison. If the distribution of the successors of a PoS is not too dispersed, conditional entropy is low. It can be assumed that low entropy shows a strong cohesion between the (syntactic) units and that liaison is more likely to occur in that context. Using a tagged corpus to calculate conditional entropy and comparing the results with data presented in the experimental literature, I conclude that conditional entropy after a PoS (with 8 categories) is strongly correlated ($r = .8$, $p < 1\%$) with the probability of liaison after the same PoS.

1 Introduction

Liaison in French is generally defined as the production of a normally silent word-final consonant before a word starting with a vowel: compare *les garçons* /legaʁsõ/, where the final /z/ of *les* is not heard, with *les enfants* /lezãfã/ where the consonant appears. Liaison occurs with a limited number of consonants (/z/, /t/, /n/, /R/, /p/) and is normally present with *enchaînement* as in /le.zã.fã/. It is considered an indicator of juncture between words that belong to a larger unit.

The traditional classification (for instance Delattre 1951; Encrevé 1988) relies on the Parts of Speech (henceforth abbreviated PoS) and distinguishes three different cases to explain the occurrence of liaison. In the compulsory case, the absence of liaison is faulty as in *les enfants*; in the forbidden category, realisation of liaison is improper as in *un soldat | anglais*; the optional case is the most interesting as the speaker has the choice to use liaison or not, generally depending on the social status or the register of the speech: in *pas encore* a higher register would condition the appearance of liaison, whilst a more laid-back style allows the absence of a liaison consonant. Syntactic cohesion is the principle that is generally invoked to justify liaison.

This approach takes the form of a long list of rules and does not define very precisely the notion of syntactic cohesion. Moreover, the separation between compulsory, optional and forbidden might prove artificial. Experimental results (Boula de Mareüil et al. 2003) indicate that liaison rates can form a continuum: compulsory liaison is realised in more than 70% of the cases and varies from 95% (determiner) to 70% (monosyllabic adverb other than *pas*); forbidden liaison

ranges from 1% to 10%. Therefore “optional” liaison might have a rate of occurrence between 10% to 70%. Some lexical items such as *mais* allow a nearly complete freedom (44%, i.e. $\frac{1}{2}$ of the data). As mentioned before, register and social status influences significantly rates of occurrence, making them even more uncertain.

Different domains have been proposed to account for liaison. Selkirk (1974) first advocated a purely syntactic approach. Using an X-Bar framework, she shows that c-command is a necessary but not sufficient condition to trigger liaison. However according to Fougeron et al. 2001b, 50% of the satisfying context liaison are not realised. Some models have focused more on the prosody; we can mention the Phonological Phrase (PhP, Selkirk 1986) and Accentual Group (GA, Scarborough and Jun 2003). According to Fougeron and Delais 2004 and Post 2000, the GA seems to yield a better prediction as many liaisons are overgeneralized between PhPs. On the other hand, many liaisons are not realised between GAs (Fougeron and Delais 2004). Bybee (2001) represents a purely lexical approach. In her view, liaison is preserved in frequent units. Therefore, words which occur together frequently are more likely to have liaison. However, this frequency effect could not be attested by experimental studies (Fougeron et al. 2001a, 2001b).

As it has proved difficult to model liaison occurrence by a deterministic model, it is reasonable to think that uncertainty is intrinsic to the phenomenon of liaison and that a robust quantitative model is needed to account for its inherent variation. Bybee’s model suggests a quantitative approach that is purely lexical. Although lexical particularities play an important role, studies by Boula de Mareüil et al. (2003) and Fougeron et al. (2001b) have shown that liaison is strongly linked to syntactic factors.

2 Model proposed

The model I propose is inspired by the observations of Zellig Harris (1955). Devising a method to segment a phonologically transcribed statement, he noticed that the number of succeeding items increases at morphological boundaries. In the simplest version of his method, morphological segmentation can thus be achieved by simply thresholding the number of successors.

The notion of the number of successors can be formalised, if we take into account the factor of frequency, using the conditional entropy (see Xanthos 2001 for a more detailed account). The conditional entropy quantifies the dispersion of the distribution of a categorical variable, in this case the successors of a PoS. It is considered a robust measure of uncertainty of this kind of variable. To take a concrete example, if the determiner is followed in more than 88% of cases by a noun, and that the 7 other word classes are left in the 12% remainder, it proves easy to predict the next PoS (a noun) and conditional entropy proves rather low: 0.73 bits (with 8 PoS). The latter figure once compared with the 2.77 bits of the

verb, which has a more evenly spread distribution of successors, indicates that it proves less difficult to predict the successors of the determiner than those of the verb. To make this measurement more explicit, it is considered that a bit corresponds to a binary question, to which one can answer by yes or no. The whole model is a probabilistic finite-state machine, where every PoS constitutes a state; therefore no long-distance dependency can be considered.

To compute the conditional entropy after a r -gram w , we need to have the appearance probabilities of all successors a drawn from an alphabet A :

$$h(w) = - \sum_{a \in A} p(a | w) \log p(a | w) \quad (1)$$

The conditional entropy is equivalent to the divergence of Kullback-Leibler (notated K) between the conditional distribution c of the successors of one r -gram and a distribution of reference, the uniform distribution u , in which each r -gram has the same probability of appearance. The Kullback-Leibler divergence is defined as:

$$K(f \parallel g) = \sum_{j=1}^m f_j \log \frac{f_j}{g_j} \quad (2)$$

More briefly, $K(c \parallel u) = 0$ iff $H(c)$ is maximum, i.e. dissimilarity between the distribution of the successors and the uniform distribution is null if and only if the conditional entropy of the distribution of the successors is maximal, which is to say that it is \log_m . If we consider the empirical distribution of the symbols for the whole of the corpus as the distribution of reference and calculate the divergence of Kullback-Leibler on this empirical distribution, we obtain the average mutual information (although we will not tackle this measure in this paper).

The less uncertainty about successors we can find, the stronger the cohesion between the units. If we consider that liaison occurs more frequently between strongly linked units, then the main assumption of this paper is the following: the lower the entropy, the more likely the occurrence of liaison.

3 Methodology and corpus

My corpus was composed of literary (Balzac and Stendhal), political and scientific texts of my own choice drawn mainly from the Gallica website of the Bibliothèque Nationale Française. To have the necessary amount of data for the Markov model, I gathered 1.5 million words. The text was tagged with the CORDIAL software retailed by Synapse Développement. I recoded the tagset produced by CORDIAL and, on this data, I computed the count, frequency and entropy measures using a program I wrote in JAVA.

The data concerning the percentage of liaison after each PoS were drawn directly from two existing studies. For a general percentage after the usual classification

in 8 PoS, I relied on Fougeron et al. (2001b) who used the SYLSWISS corpus consisting of 5h speech produced by 10 Swiss locutors. For a breakdown of specific configurations or individual items, I took advantage of data provided by Boula de Mareüil et al. (2003) that relied on the BREF corpus composed of 66,500 sentences read by 120 speakers (around 26k words).

I compared my results to these two different samples of French. This choice, mainly dictated by reasons of availability, had the advantage of avoiding the risk of results biased by the peculiarities of a corpus.

4 Experimental results

The model can be configured in different ways as it allows different granularities, i.e. a different level of precision in the definition of the PoS. For instance the subordinating and coordinating conjunctions can be split into two classes or lumped together under a single label, as “conjunctions”. Order, i.e. the number of symbols taken into account to predict the next symbol, is the second parameter. A higher order generally gives a more accurate description, but it also increases the number of symbol combinations exponentially, requiring more data and proving harder to interpret.

I focused on order 1 as it proves straightforward to interpret at first hand. I chose the 8 traditional PoS ($m=8$) to enable a comparison with the rate of liaison after each PoS provided in Fougeron 2001b. Table 1 shows the successor count of each PoS. Σ /PoS gives the overall count of a specific PoS and %/PoS gives the overall proportion of each PoS. This distribution is used to calculate the overall entropy (H) of 2.86 bits, i.e. the uncertainty linked to the choice at random of a PoS.

N	Adj	Adv	Conj	Det	Noun	Prep	Pron	Verb	Σ /PoS	%/PoS
Adj	3177	4902	13468	12040	22853	14516	11015	6520	88491	6.65%
Adv	12178	10234	6709	13728	3502	12418	14252	22933	95954	7.22%
Conj	4359	6277	3076	20060	3863	11035	21399	5679	75748	5.7%
Det	16393	2619	379	3299	196109	344	2393	688	222224	16.71%
Noun	37324	19073	30576	38177	16603	70361	45505	36415	294034	22.11%
Prep	5161	3731	857	75981	40377	1924	17326	18608	163965	12.33%
Pron	1179	13671	3084	8468	2327	6813	35361	96971	167874	12.62%
Verb	8720	35447	17598	50471	8401	46554	20623	33690	221504	16.66%

Table 1: count of successors (columns) of a PoS (rows)

Each row of Table 2 gives the distribution of the successors of a definite PoS. The uncertainty relative to this distribution is then summarised by the conditional entropy (notated h) in bits.

%	Adj	Adv	Conj	Det	Noun	Prep	Pron	Verb	h (bits)
Adj	3.6	5.5	15.2	13.6	25.8	16.4	12.4	7.4	2.79
Adv	12.7	10.7	7.0	14.3	3.6	12.9	14.9	23.9	2.85
Conj	5.8	8.3	4.1	26.5	5.1	14.6	28.3	7.5	2.65
Det	7.4	1.2	0.2	1.5	88.2	0.2	1.1	0.3	0.73
Noun	12.7	6.5	10.4	13.0	5.6	23.9	15.5	12.4	2.87
Prep	3.1	2.3	0.5	46.3	24.6	1.2	10.6	11.3	2.11
Pron	0.7	8.1	1.8	5.0	1.4	4.1	21.1	57.8	1.87
Verb	3.9	16.0	7.9	22.8	3.8	21.0	9.3	15.2	2.77

Table 2: distribution of the successors of a PoS

Let us consider the histogram of the conditional entropy associated to each PoS and ordered by growing entropy:

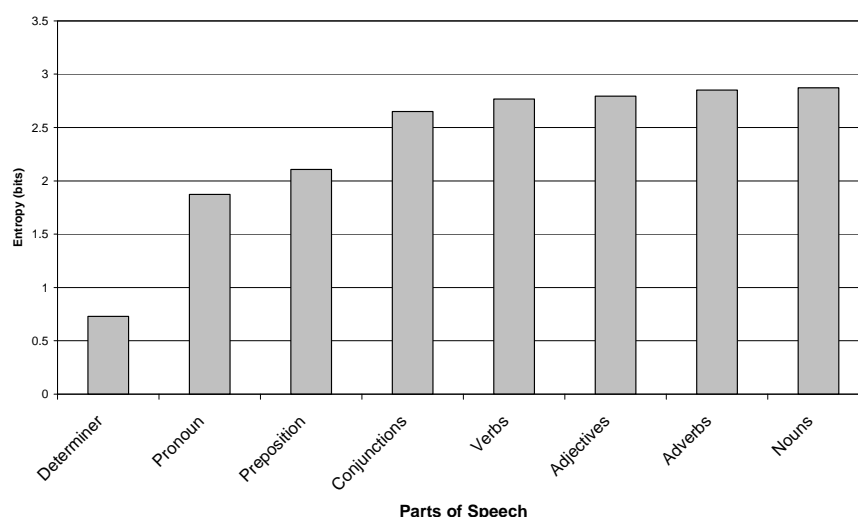


Figure 1: Histogram of conditional entropy of 8 PoS ranked in growing order

The closed classes find themselves on the left, whilst the open classes, which display a very similar rate of entropy, are grouped on the right. Our hypothesis is that the lower the entropy is, the more often liaison occurs. Does our first overview of the data confirm this assumption? If we consider a threshold between 2 and 2.5 bits, the categories under that threshold should have a higher probability of triggering liaison. The determiner confirms this theory as it is compulsory in 95% of occurrence (all percentages given in this section are from Boula de Mareuil et al. 2003 except when otherwise stated); the pronoun also triggers compulsory liaison when clitic as in *vous avez* (83%), with an overall rate of 85% according to Fougeron et al. 2001b; as for the preposition, liaison is compulsory if monosyllabic (90.5%) *en Angleterre*, otherwise optional, *devant eux*. The monosyllabic case is by far the most frequent and Fougeron et al. 2001b quote an overall rate of 99% of liaison after the prepositions in their corpus. This suggests that there is a wide variation within the items of a syntactic category and that a model with a granularity of 8 cannot give a perfect account of the particular behaviour of certain lexical items. However, those PoS under the threshold that

we defined do not seem to display cases of forbidden liaison, which would seriously endanger our hypothesis.

Those PoS over the threshold show more diversity in their behaviour. The conjunctions can be forbidden *et* (1%), optional *mais* (44%) or even compulsory *quand* (94%). Verbs prove slightly more problematic: a main verb followed by a non-clitic pronoun does not allow liaison (5%) as in *tu perds un temps fou*; when followed by a pronoun (99%) *prends-en* or auxiliary (81%), it is nevertheless compulsory, which contradicts our hypothesis. The noun behaves differently when singular (forbidden, 10.4%) or plural (optional, 28%). Adjectives very often trigger liaison when prenominal (72.5%) and rarely when postnominal (6%). Finally, the adverbs, a large residual category, display a certain diversity: when monosyllabic, they offer a wide gamut of possibilities (*loin* 0%, *pas* 41%, *très* 94%); when polysyllabic they do not allow liaison after them.

When we consider the preceding paragraph, it seems difficult to give a clear-cut rule for large categories without taking into consideration the peculiarities of the lexical item. However, if we do not want to predict every single occurrence and allow a certain margin of error, is it possible to find a general statistical relation between entropy and the rate of liaison after each PoS? To do so, I compared the rate of conditional entropy as calculated on my corpus with data found in Fougeron et al. 2001b. In this paper overall percentages of occurrence of liaison read as in Table 3:

Adj	Adv	Conj	Det	Noun	Prep	Pron	Verb
23 %	35 %	59 %	95 %	3 %	99 %	85 %	31 %

Table 3: % of realisation of liaison in function of the grammatical category of the linking word.

Crossing this data with our measures of conditional entropy, we obtain Figure 2:

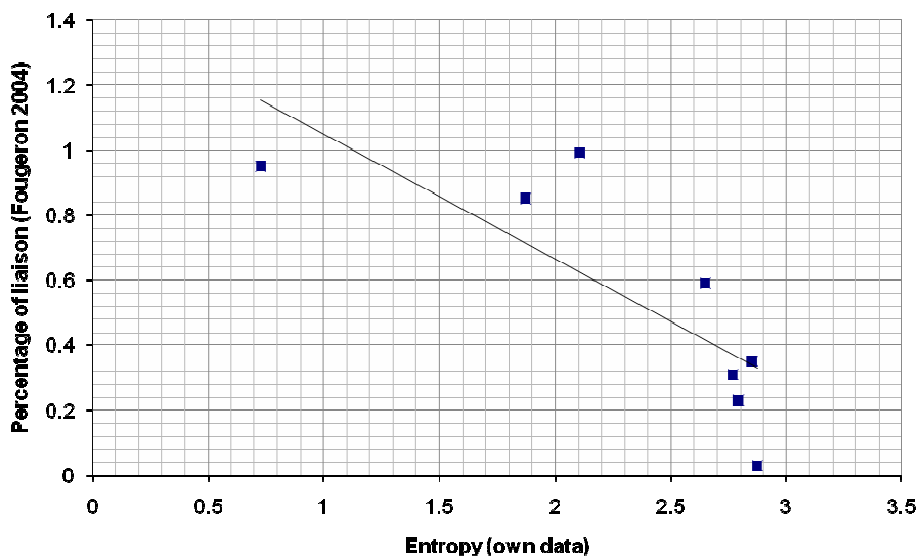


Figure 2: % of liaison according to conditional entropy

The shape of Figure 2 indicates a linear relation that we can formally confirm with the calculation of the coefficient of correlation and with a statistical test.

	% of liaison	cond. entropy		
Mean	0.537500	2.330053	T-Test	-3.222697
Variance	0.114294	0.487101	Critical 2-sided T-value (5%)	2.447000
Standard Error	0.338074	0.697926	2-sided p-value	0.018076
Covariance	-0.187848		Critical 1-sided T-value (5%)	1.943000
Correlation	-0.796134		1-sided p-value	0.009038
Determination	0.633829		Degrees of Freedom	6

Table 4: Standard statistical measures, Pearson correlation and T-tests linked to % of liaison and conditional entropy

Table 4 indicates an important correlation (Pearson) between rate of liaison and conditional entropy (r : -.8). The 1-sided p-value under 1% shows that this result is significant.

5 Discussion

The results of the correlation test confirmed the hypothesis that liaison is, on the whole, linked to conditional entropy, even if many individual lexical items do not conform to this prediction. As it is generally considered that liaison occurs between words that are strongly connected, my research would imply that this cohesion can be measured by conditional entropy. From the point of view of formal language theory, a simple probabilistic finite-state machine is sufficient to account for this phonological phenomenon, without the need of more elaborate syntax.

The choice of the granularity of the PoS is quite empirical, and it remains to be seen if the hypothesis will remain valid given finer granularity. For instance, separating the subordinating and coordinating conjunctions might improve the prediction as the coordinating conjunctions *et* (forbidden) and *mais* (optional) show a much higher entropy than *quand* (compulsory). Another path to explore includes considering bigrams, i.e. predicting the next token with two tokens instead of one. In the case of the adjective, this seems to yield benefits: the bigram Det + Adj is generally followed by a noun as in *les petits enfants* and displays a relatively low conditional entropy. This observation seems to correspond with the fact that the prenominal adjective requires compulsory liaison. Another study would therefore be required to see to which extent such refinements might improve the precision of the model and explain the variations we observe within a syntactic category.

References

- Boula de Mareüil, Ph., M. Adda-Decker, and V. Gendner (2003). Liaisons in French: a corpus-based study using morpho-syntactic information. *15th ICPHS*, 1329-1332, Barcelona.
- Bybee, J., (2001). Frequency effects on French liaison. In J. Bybee and P. Hopper (Eds.). *Frequency and the emergence of linguistic structure*, 337-359. Amsterdam: John Benjamins.
- Delattre, P. (1951). *Principes de phonétique française à l'usage des étudiants anglo-américains*. Middlebury College.
- Encrevé, P. (1988). *La liaison avec et sans enchaînement: Phonologie tridimensionnelle et usages du français*. Paris: Éditions du Seuil.
- Fougeron, C., J.-P. Goldman, and U. Frauenfelder (2001a). Liaison and schwa deletion in French: an effect of lexical frequency and competition, Paper presented at Eurospeech conference, in Aalborg, Denmark.
- Fougeron, C., J.-P. Goldman, A. Dart, L. Guelat and C. Jeager (2001b). Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. Paper presented at 8ème TALN, Tours.
- Fougeron, C. and E. Delais (2004). Liaisons et enchaînements : « Fais_en à Fez_en parlant ». Paper presented at Actes des Journées d'Etudes sur la Parole 2004, Fès, Morocco.
- Harris, Z. (1955). From phoneme to morpheme. *Language* 31: 190-222.
- Scarborough, R. and S-A Jun (2003). Accentual Phrase and the domain of liaison in French. Poster presented at 15th ICPHS, Barcelona.
- Selkirk, E. (1974). French liaison and the X-bar convention. *Linguistic Inquiry* 5, 573-590.
- Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook* 3: 371-405.
- Xanthos, A. (2001). Du k-gramme au mot: variation sur un thème distributionnaliste. MA dissertation, University of Lausanne.