

La liaison dans la parole spontanée familière: explorations semi-automatiques de grands corpus

Martine Adda-Decker^{1, 2}, Elisabeth Delais-Roussarie³, Cécile Fougeron¹
Cédric Gendrot¹, Lori Lamel².

(1) LPP, UMR 7018, 19, rue des Bernardins 75005 Paris

(2) LIMSI, UPR 3251, bât. 508, rue John von Neumann, 91403, Orsay

(3) LLF, UMR 7110 : 175 rue du Chevaleret, 75013, Paris

{madda, cfougeron, cgendrot}@univ-paris3.fr, {madda, lamel}@limsi.fr,
elisabeth.delais@linguist.jussieu.fr

RÉSUMÉ

Ce travail explore la réalisation des liaisons dans un parler spontané familier. En partant du constat que dans la parole familière les prononciations s'écartent souvent de leur forme canonique complète, et que l'on peut observer un taux de réduction temporelle plus élevé qu'en parole lue ou préparée, nous faisons l'hypothèse que le nombre de liaisons réalisées se trouve diminué dans ce type de parole. Nous mesurons les taux de réalisation des liaisons à l'aide d'alignements automatiques où nous proposons qu'une consonne de liaison peut apparaître devant un mot commençant par (semi)-voyelle. Le corpus utilisé est le corpus NCCFr (Nijmegen Corpus of Casual French) qui comprend 46 locuteurs (jeunes) et plus de 30 heures de parole. Les taux de réalisation sont mesurés pour les consonnes de liaison les plus fréquentes (/z/, /n/ et /t/) et dans des sites de liaison potentielle classés selon que la liaison y est obligatoire, facultative ou interdite. Nous proposons également une étude originale des taux de réalisation des liaisons en fonction d'une mesure de débit de parole et en fonction du locuteur.

ABSTRACT

French Liaison in casual speech : automatic and manual investigations

The realisation of the French Liaison is investigated in a large corpus of casual speech. Considering that casual speech gives rise to a large range of pronunciation variants and that overall temporal reduction increases (word and phone duration measurements) as compared to read and prepared speech, one may hypothesize that French liaison tends to be less productive in this kind of speaking style. We made use of automatic speech alignments to evaluate optional liaison realisations in potential liaison sites (word ending in a liaison consonant followed by a word-initial (semi)-vowel). Speech comes from the NCCFr corpus including 46 mostly young speakers with a total of more than 30 hours of speech. Realized liaisons were examined and measured for the most frequent liaison consonants (/z/, /n/ and /t/) as a function of a classification of the sites as mandatory, optional or forbidden with respect to liaison realization. An original contribution investigates liaison realization as a function of a speaker-dependent speech rate measure.

MOTS-CLÉS : Variantes de prononciation, liaison, parole familière, réduction temporelle.

KEYWORDS: Pronunciation variants, liaison, French, casual speech, temporal reduction.

1 Introduction

La liaison consiste en la réalisation d'une consonne finale normalement muette (suite à des changements linguistiques entre les 12^{ème} et 16^{ème} siècles) appartenant orthographiquement au mot (M_1) devant un mot (M_2) commençant par une voyelle (ou, dans quelques cas, une semi-voyelle). La liaison a été largement étudiée dans des travaux à visée descriptive et normative (orthoépique). Pendant longtemps, le matériel étudié provenait du phonéticien lui-même et les descriptions visaient à déterminer dans quels contextes, définis essentiellement sur la base des appartenances grammaticales de M_1 et M_2 , la liaison était obligatoire, facultative ou interdite (Delattre, 1966). Pour le cas des liaisons facultatives, il est apparu que de nombreux facteurs influencent la réalisation des consonnes de liaison : la catégorie grammaticale des mots, la nature de la consonne de liaison, la prosodie, le style de parole, la fréquence lexicale des mots, la fréquence de co-occurrence des bigrammes, etc. (pour une synthèse, voir (Côté, 2012)). En conséquence, la liaison peut être considérée comme un phénomène multi-factoriel et multi-niveaux largement influencé par des effets de fréquence (Mallet, 2008).

Depuis peu, le recours à de grands corpus oraux multilocuteurs (ex. Bref (Lamel *et al.*, 1991), PFC (Durand et Lyche, 2008)) contribue à la description et la quantification des phénomènes de liaison dans différents styles de parole (ex. (Adda-Decker *et al.*, 1999; Mallet, 2008)). Ces études sur des productions plus naturelles peuvent servir aussi bien pour le traitement automatique de la parole que pour la didactique des langues et la mise au point de modélisations phonologiques plus fines tenant compte du poids des différents facteurs.

Notre intérêt ici est l'examen de la réalisation des liaisons dans un registre de parole familier intime. Notre motivation est que dans ce type de parole où les réductions phonétiques sont fréquentes, un phénomène de variation phonologique tel que la liaison est influencé par des contraintes de production particulières, et une interaction entre phénomènes de variations phonétique et phonologique est à attendre. La présente étude est une première étape dans le traitement d'un large corpus de français familier. Nous proposons ici un premier bilan de la réalisation des liaisons dans ce corpus, en fonction de la nature de la consonne de liaison (/n, t, z/), du type du site de liaison (obligatoire, facultatif, interdit) et du débit de parole.

2 Corpus et Méthode

Corpus Le corpus utilisé dans cette étude est le corpus NCCFr (Nijmegen Corpus of Casual Speech (Torreira *et al.*, 2010)). Il a été conçu pour étudier la variation dans un registre familier intime. Le corpus a été enregistré fin 2007 au LPP (Laboratoire de Phonétique et de Phonologie). Il comprend des dialogues entre amis, étudiants de la région parisienne, les variables socioprofessionnelles et régionales étant ainsi relativement contrôlées. Les dialogues se déroulent parfois en présence d'une troisième personne, également du cercle d'amis, ayant comme rôle d'alimenter si nécessaire les échanges oraux entre les deux sujets. Ses contributions restent limitées et n'ont pas été exploitées. Le corpus a été transcrit manuellement au LIMSI, les transcripteurs ayant comme consigne de transcrire tous les événements audibles, dont les disfluences. De même, ils pouvaient mettre des ponctuations fortes là où cela leur semblait nécessaire. Ces ponctuations ont été retirées pour le traitement décrit ci-dessous. La table 1 donne une description du corpus en termes de mots (types et tokens) inclus dans le corpus, de mots M_1 à liaison potentielle indépendamment du mot suivant et de contextes M_1M_2 de liaison potentielle.

Contextes de liaison L'ensemble de mots « M_1 à \mathcal{L} potentielle » ont été extraits à partir de la liste de tous les mots du corpus finissant par une consonne orthographique "-s, -x, -z, -d, -t, -n, -r, -p". Cette méthode d'extraction produit une sur-génération des mots à liaisons potentielles, mais elle a l'avantage d'inclure tous les sites de liaison potentielle. Suite à un examen manuel (2 juges), cette liste a été réduite aux seuls mots contenant une consonne pouvant faire liaison (même si celle-ci est très rare). Les contextes M_1M_2 de \mathcal{L} potentielle ont ensuite été définis à partir d'une extraction des bigrammes M_1M_2 où M_2 commence par une voyelle (ou semi-voyelle) sans aucun jugement linguistique. Ont été filtrés les M_2 commençant par une semi-voyelle disjonctive, un h-aspiré, une interjection ou une hésitation, ainsi que les séquences dans lesquelles les mots M_1 et M_2 étaient clairement séparés sémantiquement ou syntaxiquement.

	tokens (k)	types (k)	
#total mots	271,9	8,7	sans silence, avec respiration et hésitation
#total mots M_1 à \mathcal{L} potentielle	105,0	4,1	sans tenir compte du contexte droit M_2
#tot. contextes M_1M_2 \mathcal{L} potentielle	26,9	2,3	M_2 commence par V ou semi-V

TABLE 1 – Description du corpus NCCFr en termes de mots lexicaux (occurrences dans le corpus=tokens, entrées lexicales=types) exprimés en milliers (k) ; nombre de mots M_1 à consonne de liaison potentielle, nombre de contextes de liaison potentielle.

Les séquences M_1M_2 de \mathcal{L} potentielle ont ensuite été classées manuellement, essentiellement d'après l'appartenance grammaticale des mots, en contextes de liaison obligatoire (p.ex *on_a faire*), facultative (p.ex *en_une heure*) ou interdite (p.ex *nous_ion*) (voir distribution Table 2). Comme cette classification s'est faite à partir d'informations limitées (transcription orthographique du trigramme M_1M_2 + mot précédent), nous avons exclu de notre étude préliminaire environ 14% des cas : les contextes où une écoute de la phrase aurait été nécessaire. Par exemple, nous avons éliminé certains M_1 "fait" où la consonne finale peut être muette (*il en fait*) ou non (*en fait*).

M_1	M_2	CL	#Occ.	Type	M_1	M_2	CL	#Occ.	Type
on	a	/n/	477	Obl	les	autres	/z/	81	Obl
ils	ont	/z/	402	Obl	les	enfants	/z/	38	Obl
on	est	/n/	375	Obl	les	hommes	/z/	34	Obl
en	a	/n/	375	Obl	les	élections	/z/	33	Obl
dans	un	/z/	148	Fac	les	a	/z/	25	Obl
pas	à	/z/	139	Fac	sais	ils	/z/	55	Int
fait	un	/t/	139	éliminé	quand	il	/t/	69	Fac
mais	il	/z/	138	Fac	gens	ils	/z/	68	Int
pas	un	/z/	112	Fac	fait	il	/t/	61	éliminé
mais	en	/z/	103	Fac	dit	oui	/t/	21	Int

TABLE 2 – Bigrammes M_1M_2 à contexte de liaison avec comptes d'occurrence et type de liaison Obl/Fac/Int ou marqués "éliminé". **Gauche** : 10 bigrammes les plus fréquents. **Droite** : bigrammes les-NOM les plus fréquents ; quelques bigrammes de fréquence moyenne représentatifs.

Alignement automatique & prononciations Afin de pouvoir déterminer la réalisation ou non des liaisons sur un si grand nombre de contextes de liaisons potentielles nous avons exploré notre corpus par alignement automatique en utilisant le système de reconnaissance du LIMSI (Gauvain *et al.*, 2005). La consonne de liaison (CL) est modélisée comme consonne finale optionnelle de la prononciation du mot M_1 dans le dictionnaire de prononciation (comme illustré dans la table 3). Nous rappelons rapidement quelques limites de ce traitement automatique : l’alignement automatique ne permet de choisir que parmi les prononciations incluses dans le dictionnaire de prononciation ; pour la grande majorité de mots, celles-ci restent souvent relativement « standard », i.e. incluant tous les phonèmes théoriquement prévus. Le style de parole familier intime inclut cependant une proportion importante de prononciations temporellement réduites et ayant une structure phonétique changée. Une partie des CL pourraient ainsi ne pas être détectées à cause de réductions temporelles extrêmes (p.ex. le mot “les” prononcé de manière très réduite [z] au lieu de [lez] pourrait entraîner un alignement avec la variante la plus courte [le] sans liaison alors que la réalisation acoustique est en fait réduite à la consonne de liaison). D’autre part si ce traitement automatique nous permet d’appréhender une quantité de données importante, nous sommes bien conscients qu’il sera nécessaire dans nos travaux ultérieurs d’aller vérifier à l’oreille cet alignement automatique.

Mots	Pron. canon.	$ \varphi_{\text{canon}} $	Variantes CL / réduites	Mots	Pron. canon.	$ \varphi_{\text{canon}} $	Variantes CL / réduites
ils	il	2	ilz / i iz	on	ɔ̃	1	ɔ̃n
animaux	animo	5	animoz	est	ɛ	1	ɛt
premier	prœmje	6	prœmjɛr	trop	tʁɔp	3	tʁɔp
maintenant	mɛ̃tənɑ̃	6	mɛ̃tɛ̃nɑ̃ mɛ̃nɑ̃	avec	avɛk	4	avɛk ɛk

TABLE 3 – Dictionnaire de prononciation avec liaisons potentielles et variantes réduites. $|\varphi_{\text{canon}}|$ indique la longueur de la prononciation canonique en nombre de phonèmes. La dernière ligne donne des variantes de réduction temporelle uniquement.

3 Résultats

Nous présentons les résultats pour les trois consonnes de liaison (CL) les plus productives : /z/, /t/ et /n/ (pour information, le /r/ générerait un peu plus de 1000 sites de liaison potentielle, et le /p/ ne se trouvait que dans 130 sites de liaison potentielle). Nous donnons pour chaque condition, le nombre de sites de liaison potentielle, le nombre de CL réalisés et ensuite le pourcentage de réalisation pour cette condition. Dans un premier temps, nous allons présenter les résultats globaux pour ces trois CL, avant de donner des détails en fonction de différents paramètres : nature de la consonne de liaison, considérations lexicales et syntaxiques, débit du locuteur.

3.1 Résultats globaux

Les résultats globaux sont présentés dans la table 4. Sur les 26946 sites de liaison potentielle de notre corpus, plus de la moitié (53%) sont des contextes de liaison facultative, 28% des contextes de liaison obligatoire et 19% des contextes de liaison interdite. Ce dernier pourcentage, élevé, reflète une des caractéristiques de l’oral familier qui se compose de successions de propositions courtes du genre *tu sais|| alors les gens|| ils _ont...*

Type	Sites	Réalisé	
	#	#	%
Obl ∪	7636	6358	83
Fac ∪?	14308	1560	11
Int	5002	275	6
All	26946	8193	30
All-Int	21944	7918	36

TABLE 4 – Résultats en fonction du type de liaison **Obl**(igatoire)/**Fac**(ultatif)/**Int**(erdit). Les lignes **All** et **All-Int** regroupent les 3 types, respectivement les 2 types Obl. et Fac. #**Sites** indique le nombre de sites potentiels de CL, Les deux dernières colonnes indiquent le nombre et le pourcentage de liaisons **réalisés**.

Environ un mot sur 10 (respectivement sur 12,5 après élimination de liaisons interdites) du corpus contient un site de liaison potentielle (26,9k sites, respectivement 21,9k sites, sans compter les liaisons interdites, sur 271,9k de mots traités). Comme attendu, les liaisons obligatoires ont des taux de réalisation élevés (83,3%) sans atteindre les 100%. Des vérifications manuelles en cours, vont permettre de quantifier la proportion de CL effectivement non réalisés par rapport à des problèmes d'alignement potentiels liés à la réduction temporelle, comme évoqués dans la section 2.

Les liaisons facultatives, les plus fréquentes, sont très peu réalisées, avec un taux moyen d'environ 11%. Ce taux très faible peut être rapproché de l'observation que notre type de parole familière inclut une proportion élevée de mots réduits temporellement, articulés très vite ou avec moins de segments. Nous allons revenir sur ce point dans la sous-section examinant les liaisons en fonction du débit. On peut faire l'hypothèse que la liaison permet un jeu syllabique résultant éventuellement dans un nombre de syllabes plus faible si la liaison n'est pas réalisée. En effet, dans bon nombre de cas où il y a un contact V#V, les deux syllabes consécutives (où la deuxième syllabe manque d'attaque consonantique) peuvent être articulées comme une seule syllabe avec les deux voyelles qui fusionnent en un seul segment vocalique, éventuellement entouré d'une semi-voyelle. Il faut aussi noter que la définition des sites de liaison facultative adoptée dans notre étude est très permissive (elle inclut des liaisons potentielles très rares). Elle est d'ailleurs radicalement différente de celle utilisée dans le corpus PFC. Toute comparaison des taux de réalisation entre études doit donc être interprétée avec précautions.

3.2 Tendances par CL

Vont maintenant être discutées les similarités et spécificités des différentes CL (voir Table 5). Elles seront présentées par ordre de fréquence décroissante des sites de liaison potentielle.

Consonne de liaison /z/ La liaison en /z/ présente le plus grand nombre de sites de liaison potentielle (14589, i.e. 54% de l'ensemble considéré). Les liaisons facultatives sont deux fois plus fréquentes que les liaisons obligatoires. Les mots M_1 les plus fréquents en contextes de liaison potentielle sont *pas, mais, sais, as, puis, vois, dans, plus, suis, es, après, vas, gens, deux* et cumulent à eux seuls plus de 5000 des 14589 sites. À part quelques adverbes et prépositions très fréquents, on peut remarquer que les première et deuxième personnes des formes verbales contribuent ensuite à créer un grand nombre des situations de liaison potentielle, en général dans

Type	/z/			/t/			/n/		
	Sites	Réalisé		Sites	Réalisé		Sites	Réalisé	
	#	#	%	#	#	%	#	#	%
Obl	3768	3197	85	194	149	77	3674	3012	82
Fac	7574	653	9	6482	779	12	252	128	51
Int	3247	108	3	854	60	7	901	107	12
All	14589	3958	27	7530	988	13	4827	3247	67
All-Int	11342	3850	34	6676	928	14	3926	3140	80

TABLE 5 – Résultats quantitatifs en fonction de la CL et le type de liaison **Obl**(igatoire)/**Fac**(ultatif)/**Int**(erdit). Les lignes **All** et **All-Int** regroupent les 3 types, respectivement les 2 types Obl. et Fac. **Sites** indique le nombre de sites potentiels de CL, Les deux dernières colonnes indiquent le nombre et le pourcentage de liaisons réalisés.

la catégorie des liaisons facultatives. Il est intéressant de noter que les déterminants (*les, des*) très fréquents en parole journalistique et en lecture de textes écrits, et pour lesquels la liaison est obligatoire avec les noms ou adjectifs suivants, n'apparaissent pas ici en premières positions. Par rapport aux résultats globaux présentés précédemment, on peut remarquer que le taux de liaison interdite n'est que de 3% et le taux de liaisons facultatives pour /z/ est le plus faible des CL avec seulement 9%.

Consonne de liaison /t/ La liaison en /t/ génère environ un tiers des sites de liaison potentielle (7530) correspondant en grande majorité à des liaisons facultatives. Les mots les plus fréquents générant des sites de liaison potentielle sont par ordre de fréquence décroissant *fait, est, tout, quand, avait, dit, sont, était, ont, vraiment* avec 3000 contextes de liaison potentielle. La distinction obligatoire/facultatif est souvent difficile : aucun de ces mots ne peut être mis dans la catégorie des liaisons obligatoires. En revanche, certains mots M_1 produisent obligatoirement une liaison, s'ils sont suivis par un ou plusieurs mots précis ou s'ils font partie d'une locution (p.ex. *tout* dans *tout à fait*) Pour le /t/, les taux de réalisation sont globalement les plus faibles (14%), ce qui s'explique d'abord par une proportion très faible de liaisons obligatoires. Ensuite, le taux de réalisation des liaisons obligatoires est le plus faible des 3 CL considérés. Il peut s'agir d'une réalité linguistique ou d'un biais méthodologique (le /t/ peut être omis lors de l'alignement automatique) : ceci doit être vérifié à la main.

Consonne de liaison /n/ Les sites de liaison potentielle sont de loin les moins fréquents pour la CL /n/ (moins de 20% de l'ensemble des sites). Cependant, la plupart des sites correspondent à des liaisons obligatoires (3674) réalisés à 82%. Ceci fait que le /n/ est presque aussi souvent réalisé dans notre parole spontanée que le /z/ (3140 /n/ vs 3850 /z/ réalisés en liaisons obligatoires et facultatives). Dans le corpus, parmi les mots générant le plus de sites de liaison potentielle en /n/, on trouve *on, en, un* responsables à eux seuls de 2000 contextes de liaison, très majoritairement de type obligatoire. D'autres mots fréquemment observés dans les sites de liaison potentielle sont *bien, bon, non, mon, rien, son, ton*. Une spécificité de la CL /n/ est de présenter un taux de liaison facultative de 52%, bien plus élevé que le /z/ et le /t/.

3.3 Influence du débit

Afin d'examiner la réalisation des CL en fonction du débit, il nous fallait introduire une mesure de débit de parole. Différentes mesures sont utilisées dans la littérature : le nombre moyen de mots par seconde, de syllabes par seconde, de segments phonétiques par seconde, incluant ou non les silences inter-mots.

Notre idée ici est de proposer une mesure de débit qui se calcule au niveau du mot (une mesure locale) et qui ne tient pas compte des silences ou autre temps écoulé hors production lexicale. De plus, nous voulons que la mesure permette de refléter les réductions temporelles des mots. Prenons l'exemple du mot *maintenant* pour lequel nous avons introduit plusieurs variantes plus courtes dans le dictionnaire. Une réalisation [mɛ̃nã] avec 4 segments et 2 syllabes est réduite par rapport à une réalisation canonique complète [mɛ̃tənã] avec 6 segments et 3 syllabes.

Notre mesure compte le nombre de segments de la forme canonique (théorique) et non pas le nombre de segments articulés de la forme de surface. Pour chaque mot du corpus, nous calculons un débit local grâce à l'équation 1. Un débit moyen peut ensuite être calculé par l'équation 2 soit pour l'ensemble du corpus, soit par locuteur, comme nous le proposons ici.

$$Debit(m_i) = \frac{|\varphi_{canonique}(m_i)|}{Duree(m_i)} \quad (1)$$

$$Debit = \frac{\sum_i |\varphi_{canonique}(m_i)|}{\sum_i Duree(m_i)} \quad (2)$$

D'après cette mesure le débit moyen d'un corpus journalistique (ESTER) est de 15 phonèmes par seconde (ce qui revient à une durée moyenne de 7 cs par segment phonémique). En revanche, cette même mesure appliquée à notre corpus de parole familière et plutôt intime (NCCFr) donne un débit moyen de 26 phonèmes par seconde, ce qui montre l'importance du phénomène de réduction temporelle en parole spontanée et qui pose un défi en modélisation acoustique et lexicale au traitement automatique de la parole.

Afin d'examiner des différences de réalisation de liaison entre locuteurs, nous avons appliqué notre mesure 2 sur l'ensemble des mots produits par locuteur afin de calculer un débit moyen par locuteur. Nous effectuons ensuite une analyse de régression prenant comme facteurs débit moyen et locuteur avec le taux de liaisons réalisées comme variable expliquée.

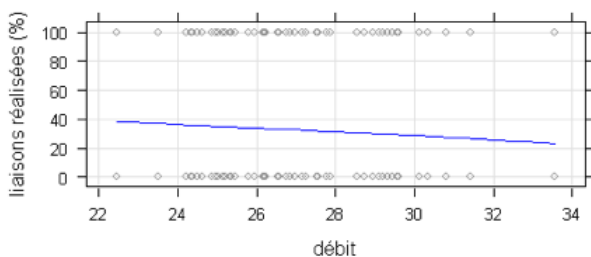


FIGURE 1 – Réalisation des consonnes de liaison en fonction du débit moyen du locuteur

4 Conclusions

Nous avons présenté des résultats de réalisation de la liaison dans un grand corpus de plus de 30 heures de parole spontanée dans un registre familial, intime incluant majoritairement des locuteurs jeunes de la région parisienne. Ce style de parole se caractérise par un taux de réalisation des liaisons en parole particulièrement faible (36% sans compter les liaisons interdites). Il se caractérise aussi par un débit très élevé (26 phonèmes par seconde par rapport à 15 phonèmes par seconde dans un corpus de parole journalistique ESTER) avec notre mesure (voir 2). Nous avons pu mesurer que le taux de liaison baisse si le débit moyen du locuteur augmente. Concernant la fréquence des sites potentiels de liaison, on a pu mesurer l'ordre suivant /z/ > /t/ > /n/, observé de manière général en français. En revanche, cette ordre change pour les réalisations de la manière suivante : /z/ >= /n/ » /t/.

Dans des travaux futurs, nous comptons examiner les découpages prosodiques et les intégrer dans l'analyse au même titre que le débit, le traitement du schwa et la nature des segments de liaison. Cela devrait permettre de voir s'il existe des relations entre ces différents éléments qui peuvent intervenir dans la réduction temporelle.

Remerciements

Le travail présenté a été en partie soutenu par le LabEx EFL (Empirical Foundations of Linguistics) et par le projet OSEO Quaero.

Références

- ADDA-DECKER, M., Boula de MAREÛIL, P. et LAMEL, L. (1999). Pronunciation variants in French : schwa & liaisons. In *International Congress of Phonetic Sciences*, pages 2239–2242, San Francisco, USA.
- CÔTÉ, M.-H. (2012). French liaison. In M. van OOSTENDORP, C. Ewen, E. H. et RICE, K., éditeurs : *Companion to phonology*, pages 2685–2710. Wiley-Blackwell, Malden, MA :.
- DELATTRE, P. (1966). *Studies in French and comparative phonetics*. Mouton & Co, La Haye.
- DURAND, J. et LYCHE, C. (2008). French Liaison in the light of corpus data. *Journal of French Language Studies*, 18:33–66.
- GAUVAIN, J.-L., ADDA, G., ADDA-DECKER, M., ALLAUZEN, A., GENDNER, V., LAMEL, L. et SCHWENK, H. (2005). Where are we in transcribing French broadcast news? In *Proc. Interspeech*, pages 1665–1668, Lisbon, Portugal.
- LAMEL, L., GAUVAIN, J. et ESKENAZI, M. (1991). BREF, a Large Vocabulary Spoken Corpus for French. In *Proc. EUROSPEECH - European Conference on Speech Communication and Technology*, Genova, Italy.
- MALLET, G. (2008). *La liaison en français : descriptions et analyses dans le corpus PFC*. PhD dissertation, Université Paris 10, Paris.
- TORREIRA, F., ADDA-DECKER, M. et ERNESTUS, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 52:201 – 212.