Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français

Cécile Fougeron (1), Jean-Philippe Goldman (2),
Alicia Dart (1), Laurence Guélat (1), Clémentine Jeager (1)

(1) Laboratoire de Psycholinguistique Expérimentale - Université de Genève
40 boulevard du Pont d'Arve, 1211 Genève 4 – Suisse
Cecile.Fougeron@pse.unige.ch

(2) Laboratoire d'Analyse et de Technologie du Langage - Université de Genève
2 rue de Candolle, 1211 Genève 4 – Suisse
Jean-Philippe.Goldman@lettres.unige.ch

Résumé – Abstract

Les nombreuses recherches portant sur le phénomène de la liaison en français ont pu mettre en évidence l'influence de divers paramètres linguistiques et para-linguistiques sur la réalisation des liaisons. Notre contribution vise à déterminer la contribution relative de certains de ces facteurs en tirant parti d'une méthodologie robuste ainsi que d'outils de traitement automatique du langage. A partir d'un corpus de 5h de parole produit par 10 locuteurs, nous étudions les effets du style de parole (lecture oralisée/parole spontanée), du débit de parole (lecture normale/rapide), ainsi que la contribution de facteurs syntaxiques et lexicaux (longueur et fréquence lexicale) sur la réalisation de la liaison. Les résultats montrent que si plusieurs facteurs étudiés prédisent certaines liaisons, ces facteurs sont souvent interdépendants et ne permettent pas de modéliser avec exactitude la réalisation des liaisons.

Various studies on liaison phenomena in French have shown the influence of several linguistics as well as para-linguistics factors on liaison realization. In this study we aim at determining the relative contribution of certain of these factors by using a robust methodology and tools used in automatic language processing. In a 5 hours speech corpus, produced by 10 speakers, we study the effect of speech style (oral reading/spontaneous speech), speech rate, as well as the contribution of syntactic and lexical (word length and frequency) factors on liaison realization. Results show that even if several factors can contribute to predict some liaisons, these factors are often interdependant and do not allow a sufficient prediction of liaison realization.

Mots clés : liaison, lecture, spontané, débit de parole, longueur, fréquence lexicale.

1 Introduction

La liaison est un phénomène phonologique caractéristique du français par lequel la consonne finale de certains mots (dite consonne latente de liaison) apparaît en surface quand le mot suivant commence par une voyelle, mais pas par une consonne. Par exemple, « les » sera prononcé [lez] devant « enfants », mais [le] devant « garçons ». Souvent reconnus comme l'un

des problèmes les plus épineux du phonétisme français, la liaison est intéressante sur plusieurs points.

D'une part, elle soulève les questions linguistiques de la représentation lexicale des formes avec consonne latente, de la modélisation de la nature alternante de cette consonne et de son affiliation syllabique, de la définition des facteurs conditionnant son apparition et de son domaine d'application. Ainsi, la modélisation de la réalisation des liaisons en français à mis à l'épreuve la plupart des théories phonologiques du français (ex. Shane '68 ; Selkirk '74 ; Morin & Kaye '82 ; Clements & Keyser '83 ; Klausburger '84 ; Dell '85, Tranel '00). La liaison se trouve également au cœur du débat sur les relations entre les différentes composantes de la Grammaire (ex. Inkelas & Zec '95), sa réalisation ne pouvant s'expliquer qu'en faisant appel à des critères à la fois morpho-syntaxiques, lexicaux, prosodiques, stylistiques et para-linguistiques.

D'autre part, la liaison, en tant que variante de production, soulève des questions relatives au traitement du langage qu'il soit naturel ou automatique. En psycholinguistique, la réalisation de la liaison est étudiée du point de vue des processus de traitement mis en jeu dans la production et la perception de mots prenant des formes variables en surface. Un des débats actuels concerne l'incidence possible d'un coût sur la reconnaissance lexicale des formes liées, du fait (1) de l'épenthèse d'une consonne dont l'affiliation lexicale est déterminée par la nature du segment commençant le mot suivant, et (2) de la resyllabation de cette consonne au travers des frontières de mot. Ces phénomènes soulèvent les questions du traitement séquentiel des mots de la parole, de la nature des unités de segmentation (syllabiques et/ou supra-lexicales), et de l'intégration d'informations linguistiques variées. Il apparaît donc évident qu'une meilleure connaissance des facteurs conditionnant la réalisation des liaisons est indispensable à la compréhension des processus cognitifs de traitement. Du point de vue du traitement automatique, en synthèse de la parole par exemple, une bonne modélisation des facteurs conditionnant la liaison est cruciale pour l'intelligibilité. En effet, omettre une liaison obligatoire ou réaliser une liaison interdite est immédiatement perçu comme une erreur de prononciation et peut nuire à la compréhension, comme l'illustre l'exemple suivant :

« Le petit ¦ arbitre la rencontre de demain. » vs. « Le petit arbitre la rencontre chez elle ».

Le phénomène de la liaison en français faisant l'objet d'études de domaines linguistiques variés, les travaux existants ont pu mettre en évidence la complexité de la modélisation des facteurs influençant sa réalisation. En effet, si certains cas de liaisons peuvent être expliqués par des règles établies, leurs domaines de réalisations (syntaxiques ou prosodiques) ainsi que les facteurs rendant compte de la variabilité dans la réalisation des liaisons restent encore débattus. Que ce soit pour le traitement automatique ou naturel, notre objectif est donc de savoir quels types d'informations doivent être pris en compte par l'homme ou la machine dans le traitement de la liaison..

De nombreux facteurs ont été décrits dans la littérature comme influençant la réalisation de la liaison. Ils sont de nature variée et relèvent de composantes à la fois linguistiques et para-linguistiques : organisation morpho-syntaxique et prosodique, caractéristiques lexicales, stylistiques, situationnelles, géographiques et socio-culturelles. Il ne s'agira pas ici de faire une revue de tous ces facteurs mais d'en choisir certains selon un critère qui répond à un de nos objectifs à long terme : pouvoir déterminer automatiquement, d'après ces facteurs, la probabilité d'occurrence d'une liaison. Pour cela, nous avons choisi un certain nombre de variables syntaxiques et lexicales facilement implémentables dans un système de synthèse vocale tel que FIPSVox (Gaudinat et al. '97) développé au LATL à Genève. En plus de ces facteurs, nous étudierons la contribution d'informations telles que le style ou la fréquence lexicale qui sont connus pour être intégrés dans les processus de traitement naturel de la parole en production et en perception. Nous n'aborderons pas, dans cette étude, l'effet de la structuration prosodique,

bien que sa contribution nous paraisse essentielle au conditionnement de la liaison. Les facteurs que nous avons étudiés sont les suivants :

Facteurs syntaxiques. La liaison est soumise à des conditions syntaxiques (1) d'appartenance catégorielle et (2) de configuration syntaxique. (1) Pour qu'il y ait liaison, le lieur et le lié doivent former une paire de catégories syntaxiques valides. Pour certaines d'entre elles, la liaison est considérée comme obligatoire (déterminant ou adjectif suivi d'un adjectif ou nom, pronom clitique précédant ou suivant un verbe). Pour d'autres, elle est facultative comme entre un nom et un adjectif, et c'est alors la fréquence de la co-occurrence de la paire qui, entre autre, influencera la réalisation de la liaison (p.ex. *«relations amoureuses»*). (2) D'un point de vue configurationel, il a été établi que la c-commande, relation asymétrique entre deux constituants d'une arborescence syntaxique selon la théorie X', est une condition nécessaire, mais non suffisante à la liaison (Selkirk '74). La c-commande, essentiellement utilisée pour les relations anaphoriques et les règles d'accord, est définie chez Reinhart ('81), comme :

> *A c-commande B, si A ne domine pas B, B ne domine pas A, et le premier nœud branchant dominant A domine également B.*

Sur l'exemple de la figure 1, les flèches indiquent les contextes de mots contigus présentant une suite consonne-voyelle. Pour les liaisons autorisées (trait plein et discontinu), il y a relation de c-commande entre le lieur et le lié. Par exemple, le déterminant «*les*» c-commande l'adjectif «*adorables*», qui lui même c-commande le nom «*enfants*». Dans le cas de «*enfants ont*» (trait pointillé), la condition de c-commande n'est pas réalisée : le nœud NP, qui est le premier nœud branchant dominant le nœud N «*enfants*», ne domine pas l'auxiliaire «*ont*» en position T', et la liaison n'a pas lieu. Le contexte «*conduits en*» montre que la c-commande est une condition nécessaire mais pas suffisante pour la liaison : le lieur c-commande le lié, mais la liaison est intuitivement peu probable, voir interdite.

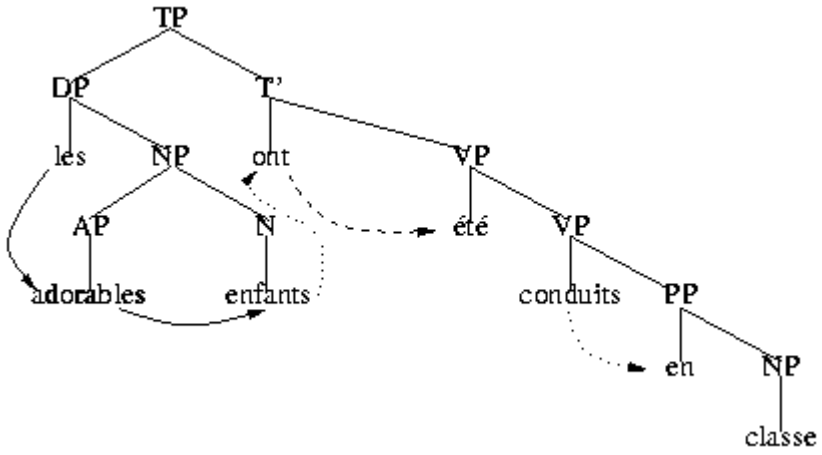

Figure 1 : Analyse syntaxique de « les adorables enfants | ont été conduits | en classe. », par FIPS (LATL).

Facteurs stylistiques. Le niveau stylistique du discours est considéré par certains comme l'un des facteurs les plus importants pour le conditionnement des liaisons facultatives, avec un accroissement des occurrences de liaison dans les registres qualifiés de « recherchés, élaborés, soutenus, soignés » (Delattre '66, Agren '73, Anderson '75). Dans la présente étude, nous testerons l'effet de ce facteur en comparant les occurrences de liaisons dans les productions d'un texte lu et dans des conversations spontanées. Outre les différences dans la situation de communication et le caractère improvisé ou non du message, ces deux types de productions diffèrent dans leur forme et leur substance. La lecture est celle d'un texte journalistique, alors que la conversation se fait entre interlocuteurs intimes dans un niveau de langue plus familier. L'influence du débit de parole sur la réalisation des liaisons a rarement été étudiée en soi. Cependant, les différents styles qui ont été observés mettent également en jeu des

caractéristiques de débit différentes (voir p.ex. Lucci '83). Nous chercherons donc ici à déterminer en quoi le débit de parole peut influencer la réalisation des liaisons. Facteurs lexicaux. Parmi les facteurs « lexicaux », l'influence de la fréquence lexicale sur la production de la parole connaît un intérêt croissant, en particulier en ce qui concerne son impact sur les variations en parole continue. . Plusieurs auteurs ont montré que les variantes de production (p.ex. l'assimilation) sont sensibles au poids informatif des mots : elles sont moins fréquentes sur des mots rares, portant un focus (Cooper & Paccia-Cooper, '80) ou une information nouvelle, et sur des mots avec une forte densité de compétiteurs lexicaux (Wright '98). En tant que variante de surface, la liaison peut également être affectée par des contraintes « orientées vers la sortie » d'intelligibilité et de minimisation des ambiguïtés, et donc être sensible à la fréquence lexicale des mots. Un tel effet de fréquence lexicale a été relevé dans plusieurs corpora, où les liaisons semblent plus souvent réalisées avec des mots lieurs fréquents qu'avec des mots lieurs rares (Agren, '73 ; Bybee to appear). Par exemple, dans leur corpus de 155h de parole, Adda et collègues ('99) montrent que si l'on considère les mots lieurs les plus fréquents (256 mots), 75% des liaisons sont réalisées, alors que si l'on étend l'échantillon aux 2560 mots les plus fréquents, le taux de réalisation diminue à 64%, contre 55% si l'on considère tout le corpus.

Un autre facteur de type lexical est celui de la longueur des mots impliqués dans la liaison. Il semblerait que la liaison soit plus fréquemment réalisée avec des mots lieurs courts (Agren '73, DeJong '91). Par exemple, Encrevé ('88) note qu'avec des lieurs monosyllabiques la liaison est réalisée à 77%, contre 29% pour les lieurs polysyllabiques. La longueur du mot lié, cette fois, influence aussi l'occurrence des liaisons d'après Morin et Kaye ('82) qui observent des liaisons plus fréquentes avec des compléments courts (voir aussi Delattre '66). Par contre, à notre connaissance, aucune étude ne spécifie si cette contrainte de longueur est fonction de la longueur des items liés (lieurs et/ou liés) ou si elle dépend de la longueur totale de la séquence liée. En effet, l'effet de longueur du mot lieurs ou du mot lié peut n'être qu'une conséquence des caractéristiques de fréquence ou de catégorie grammaticale de ces mots (les mots courts incluent les déterminants, prépositions, auxiliaires et sont souvent plus fréquents). En revanche, le facteur de longueur peut avoir un effet propre si l'on considère que la liaison a pour fonction le groupement de deux mots au sein d'un composant prosodique comme le groupe accentuel. Puisque la longueur de ce groupe est limitée (3 syllabes en moyenne et, en général, pas plus de 6-7 syllabes, Jun et Fougeron '00), une liaison créant un groupe plus long pourrait être évitée. En résumé, il apparaît que les facteurs invoqués dans la littérature sont souvent interdépendants. Nous tâcherons de les étudier dans une approche comparée afin de déterminer leur influence relative. Notre démarche consistera à observer les conditions de réalisation des liaisons sur un corpus en manipulant quelques facteurs tels que le débit et le style, puis en observant sur ce corpus les caractéristiques des liaisons réalisées selon les facteurs décrits précédemment.

2 Méthode

2.1 Description des corpora

Les occurrences de liaisons ont été observées dans trois situations produites par les mêmes locuteurs : la lecture d'un texte à débit normal (LN), la lecture du même texte à débit rapide (LR), une conversation spontanée (CS).

Corpus de lecture. Le texte utilisé pour les deux conditions de lecture a été conçu à partir d'extraits d'un corpus d'environ 30000 mots provenant d'articles du journal « le Monde » de janvier 1987. A l'origine, ce corpus avait été créé dans le cadre d'un projet sur l'évaluation de systèmes de synthèse (Action de Recherche Concertée B3 «synthèse de la parole», AUPELF-UREF 1996-99) pour lequel il avait été phonétisé automatiquement, et vérifié manuellement.

Outre l'exactitude de la transcription phonétique, ce corpus présente l'avantage de contenir les variantes de prononciation pour chaque mot en incluant : les consonnes latentes de liaison, les élisions possibles, ainsi que les variantes de prononciation des nombres comme 90 (nonante/quatre-vingt-dix). Afin de maximiser le nombre de liaisons potentielles dans notre texte pour la lecture, les phrases du corpus comportant au moins deux contextes de liaison possibles ont été sélectionnées et regroupées pour former un texte cohérent avec, au besoin, l'ajout de phrases introductives. Au total, le texte comportait 1860 mots, donnant un échantillon de parole d'environ 10 minutes à débit normal (voir tableau I). .

Corpus spontané. Les productions de conversation spontanée sont issues de dialogues entre les locuteurs et les expérimentateurs. Ces conversations ont un style relaché, les interlocuteurs étant amis. Ce style moins formel apparaît dans le choix d'un lexique plus familier et dans les nombreuses répétitions, réductions, hésitations, faux-départs, etc…. Les thèmes abordés dans ces conversations sont variés. Une fois l'élimination du temps de parole de l'expérimentateur, les échantillons de productions spontanées varient entre 10 et 17 minutes et 1569 à 3700 mots (voir tableau I, pour les données par locuteurs). .

Locuteurs et procédure. Dix locuteurs francophones suisses, 6 femmes (loc. 1 à 6) et 4 hommes (loc. 7 à 10), âgés entre 20 et 30 ans, issus d'un milieu socio-culturellement favorisé, ont été enregistrés. Durant les sessions d'enregistrement, le locuteur prenait connaissance du texte en le lisant silencieusement, puis le lisait une fois à débit normal, puis à débit rapide. Les deux débits étaient définis par le locuteur lui-même, avec pour seule consigne le fait que le débit rapide devait être supérieur au débit normal (voir tableau I). La session de lecture était suivie d'une conversation spontanée entre le locuteur et l'expérimentateur. Tous les locuteurs étaient naïfs vis à vis de l'objet de l'étude.

2.2 Analyses et facteurs étudiés

Contextes de liaison possible et liaisons réalisées. Afin de pouvoir comparer la réalisation des liaisons dans les différents échantillons de parole, le nombre de liaisons réalisées a été comparé au nombre de liaisons possibles dans l'échantillon de parole. Pour ce faire, nous avons défini comme « contexte de liaisons possibles (CL) » les séquences présentant un mot1 se terminant par un consonne latente de liaison (/t, n, z, r, p/) et un mot2 commençant par une voyelle ou un « h » muet. Cette définition repose donc uniquement sur des critères orthographiques et lexicaux (lexique de mot finissant par une consonne de liaison, et lexique de mot commençant par un « h » muet ou aspiré), critères qui sont facilement implémentables dans un système automatique. De cette façon, la définition des CL ne prend pas en compte des critères de jugements linguistiques, syntaxiques ou prosodiques sur les liaisons obligatoires, interdites ou facultatives, afin de pouvoir examiner l'influence de certains de ces facteurs sur la réalisation des liaisons. Pour la parole spontanée, cette procédure a été appliquée sur des transcriptions épurées des interjections et hésitations. Parallèlement, les occurrences de liaisons réalisées en lecture et en spontané ont été relevées indépendamment par 2 juges. .

Facteurs étudiés. Outre l'examen de l'incidence des facteurs manipulés « style » et « débit » sur la réalisation de la liaison, nous avons examiné, sur le corpus de lecture uniquement (les deux débits confondus), les facteurs syntaxiques de « catégorie grammaticale » et de « distance syntaxique », et les facteurs lexicaux de « longueur », « fréquence lexicale », « type de consonne de liaison ». Ces facteurs seront décrits dans les sections de résultats correspondantes.

	loc. 1	loc. 2	loc. 3	loc. 4	loc. 5	loc. 6	loc. 7	loc. 8	loc. 9	loc. 10	moy.
Temps de parole (min.)											
LN	12:11	09:07	11:35	09:58	10:52	17:19	10:51	11:11	11:11	12:06	11:38
LR	10:14	10:24	11:20	11:59	15:22	10:39	11:06	09:37	13:09	10:09	11:23
CS	07:39	08:19	08:04	09:17	12:20	09:38	09:43	08:23	11:20	07:41	9:14
Débit de parole (nombre de mots/min.)											
LN	182	179	164	155	121	175	168	193	141	183	166
LR	243	224	231	200	151	193	191	222	164	242	206
CS	191	234	208	234	144	214	153	195	168	195	194
Contexte de Liaison : nombre d'occurrences / occurrences par 100 mots											
LN et LR :	243 / 13										
CS	213 / 9	172 / 8	178 / 7	133 / 6	172/ 11	398/ 11	169/ 10	280/ 13	96 / 5	149 / 6	196/8.6
Liaisons réalisées : nombre d'occurrences / occurrences par 100 mots											
LN	95 / 5	89 / 5	95 / 5	87 / 5	94 / 5	114 / 6	131 / 7	88 / 5	93 / 5	127 / 7	101/5.5
LR	93 / 5	93 / 5	103 / 6	90 / 5	95 / 5	116 / 6	136 / 7	136 / 7	90 / 5	123 / 7	107/5.8
CS	68 / 3	48 / 2	61 / 3	50 / 2	56 / 4	157 / 4	67 / 4	67 / 3	53 / 3	59 / 2	69/3
Liaisons réalisées par rapport au nombre de Contexte de Liaison (%)											
LN	39	37	39	36	39	47	54	36	38	52	42
LR	38	38	42	37	39	48	56	56	37	51	44
CS	32	28	34	38	33	39	40	24	55	40	36

Tableau I : Caractéristiques des productions des locuteurs (1 à 10) dans les deux styles « lecture (L) » et « spontané (CS) » et aux deux débits de lecture normal (LN) et rapide (LR).

3 Résultats

3.1 Occurrence des liaisons en fonction du débit et du style.

Le tableau I présente les caractéristiques de débit, les occurrences de liaison observées et les contextes de liaisons possibles dans les 3 situations de production. Entre la lecture rapide et la lecture normale, on observe une augmentation moyenne du débit de 24% (11 à 40% selon les locuteurs), mais le pourcentage de liaisons réalisées sur les 243 contextes de liaisons possibles ne varie que très peu avec, en moyenne, 5.4 et 5.7 liaisons tous les 100 mots. Tous locuteurs confondus, le débit n'a donc pas d'effet significatif sur l'occurrence des liaisons ($\chi^2(1)=3.2, p=0.07$). Seul le locuteur 8 montre une augmentation significative d'environ 20% de réalisation des liaisons en débit rapide ($\chi^2(1)= 1{,}19, p<.0001$). En revanche, en conversation spontanée, on observe une tendance à diminuer la réalisation des liaisons par rapport à la lecture avec, en moyenne, 3 liaisons tous les 100 mots (CS vs. LN ($\chi^2(1)= 20.5, p<.0001$), vs. LR ($\chi^2(1)= 38.5, p<.0001$) tous locuteurs confondus). Cependant, cet effet dépend des locuteurs. Lorsque l'on compare la CS avec les LN, une diminution significative de la réalisation des liaisons (de 12 à 14%) apparaît pour les locuteurs 7, 8 et 10 (à $p<.01$). Comparés à la LR, ces 3 locuteurs et le locuteur 6 présentent une diminution significative de leur taux de liaison de 8 à 32%. En revanche, le locuteur 9 augmente de manière significative la réalisation des liaisons en CS (17-18%) par rapport aux deux types de lecture. Pour les autres locuteurs, l'occurrence des liaisons est similaire en lecture et en spontané. Il est à noter que les 4 locuteurs masculins (7 à 10) ont les taux de réalisation des liaisons les plus élevés et présentent les plus fortes variations en fonction du style de parole . Cependant, cette variabilité dans la réalisation des liaisons ne semble pas corrélée à une variation particulière de débit en spontané par rapport à la lecture. En effet, le nombre de mots produits par minute en CS vs. LN ou LR varie de façon différente pour ces 4 locuteurs.

3.2 Occurrences des liaisons en fonction des facteurs syntaxiques et lexicaux

3.2.1 Facteurs syntaxiques

Dans un contexte de liaison, la définition de la c-commande est simplifiée car A (le lieur) et B (le lié) sont des têtes lexicales, et l'un ne peut dominer l'autre. De plus, il n'y pas d'autre contrainte entre B et l'ancêtre commun dominant à A et B (noté C) que la dominance du dernier sur le premier, qui est vérifiée par définition. Donc la c-commande équivaut simplement à dire que la distance entre A et C doit être égale à 1. Pour modéliser cette configuration syntaxique et la faire intervenir comme facteur dans notre étude, nous avons simplement calculé, pour chaque contexte de liaison, la distance entre le mot lieur et le nœud dominant le lieur et le lié. Cette distance syntaxique a été calculée grâce à l'analyseur syntaxique FIPS du LATL (http://www.latl.ch). Cet outil linguistique robuste est basé sur le modèle générativiste chomskyen. Il prend en entrée un texte brut et tente de le découper en phrases, associant à chacune d'elles une ou plusieurs structures syntaxiques correspondant aux structures de surface enrichies. Quelques ajustements de cet analyseur ont permis d'extraire automatiquement, pour chaque contexte de liaison, les informations syntaxiques utiles pour notre étude : distance syntaxique, catégorie grammaticale du lieur et du lié.

Le tableau II présente le pourcentage moyen de réalisation des liaisons selon la catégorie grammaticale. Par exemple, pour les adjectifs, sur les 29 contextes de liaisons possibles avec un déterminant en tant que lieur, le taux moyen de réalisation des liaisons est de 95% avec un écart type de 19%.

	Adj.	**Adv.**	**Conj.**	**Det.**	**Nom**	**Prép.**	**Pron.**	**Verbe**
lieur	23 % (17;39)	35 % (36;37)	59 % (8; 36)	95 % (29; 19)	3 % (45; 6)	99 % (15; 2)	85 % (13; 37)	31 % (77; 38)
lié	49 % (25; 45)	26 % (13; 35)	6 % (33; 23)	35 % (32; 42)	85 % (33; 31)	12 % (47; 24)	61 % (12; 48)	66 % (45; 37)
dist	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
	50 % (144;44)	43 % (56; 42)	18 % (12; 38)	8 % (12; 29)	0 % (7; 0)	2 % (3; 3)	0 % (2; 0)	

Tableau II : Taux de réalisation des liaisons en fonction de la catégorie grammaticale des mots lieurs et liés et de la distance syntaxique entre les mots. Nombre d'item et écart type entre ( ).

En ce qui concerne les lieurs, il est possible d'opposer les catégories grammaticales fermées (ex. déterminants, pronoms, prépositions), dont le taux de réalisation dépasse 80%, aux classes ouvertes (ex. noms, adjectifs, adverbes et verbes) qui ont un taux de réalisation moindre. Il faut noter que la classe fermée contient un nombre plus restreint d'items différents (par ex. 9 types de déterminants « des, un, aux, mes, leurs, les, trois, son, mon », 5 types de pronoms « en, on, les, eux, ils »). De la même façon, les catégories verbe et adverbe de la classe ouverte contiennent aussi plusieurs répétitions de certains items (p.ex. 9 « est », 9 « pas ») ce qui contribue à augmenter le taux de liaisons pour ces catégories. L'appartenance catégorielle du mot lieur apparaît également comme un bon prédicteur pour les noms : sur les 42 items nominaux pluriels de notre corpus, le taux de réalisation de liaisons est en moyenne de 3% avec une faible variation entre les items (σ=6%). Dans la seconde ligne du tableau, les résultats concernant les mots liés ne permettent pas de distinguer clairement les catégories grammaticales. En effet, pour une plus grande variété de catégories, les liaisons apparaissent à plus de 50%. On remarque pourtant que les trois catégories les moins liées sont des classes fermées. En conclusion, la catégorie du lieur et du lié ne semble pas être un indice fiable quand il est considéré de façon globale comme ici. La forte variabilité (voir les écarts types) au sein de chaque catégorie montre que le taux de liaisons dépend plutôt de la nature des items lexicaux considérés.

En ce qui concerne la mesure de cohésion syntaxique dans la séquence, on observe une corrélation négative significative (r=-.38, p<.01) entre le taux de réalisation des liaisons et la distance syntaxique entre le lieur et le lié. Cette relation apparaît sur la dernière ligne du tableau II, où l'on remarque que les liaisons sont plus fréquemment réalisées avec une distance de 1 (c-commande) ou 2. Pourtant, cet indice ne permet pas de prédire la réalisation des liaisons à lui seul, puisque 50% des liaisons avec une distance 1 ne sont pas réalisées.

3.2.2 Facteur longueur

L'effet de la longueur des mots sur l'occurrence de la liaison est évaluée en observant la fréquence de réalisation des liaisons en fonction du nombre de syllabes dans le mot lieur, le mot lié, ainsi que la séquence liée. La proportion de liaisons réalisées en fonction de ce critère, ainsi que le coefficient de corrélation (pearson) entre les deux variables longueur/réalisation, sont donnés dans le tableau III. Nos observations confirment les données de la littérature en ce qui concerne la forte proportion de liaisons réalisées avec des mots lieurs courts. Pourtant, il est à noter que cet effet ne concerne que les lieurs mono-syllabiques, et que cette tendance affecte particulièrement les conjonctions, déterminants, prépositions, pronoms et verbes qui ont un taux de réalisation supérieur à 60% (figure 2). En revanche, les mots de deux syllabes et les monosyllabiques appartenant aux autres catégories syntaxiques, ne sont pas des lieurs fréquents (ex. 5% pour « des laits écrémés », 30% pour « mais aussi »).

	1 syll	2 syll	3 syll	4 syll	5 syll	6 syll	7 syll	r
lieur	71 (127,39)	15 (65,26)	6 (41, 16)	3 (9, 8)				-0.6
lié	29 (129,41)	51 (54, 42)	63 (44, 42)	59 (14, 46)				0.3
séq.		62 (53, 43)	35 (71, 43)	45 (76, 44)	33 (28, 42)	15 (11, 29)	9 (4, 18)	-0.2

Tableau III : Taux moyen de liaisons et coefficient de corrélation en fonction du nombre de syllabes des lieurs, liés et de la séquence. nombre d'items et écart-type entre ().

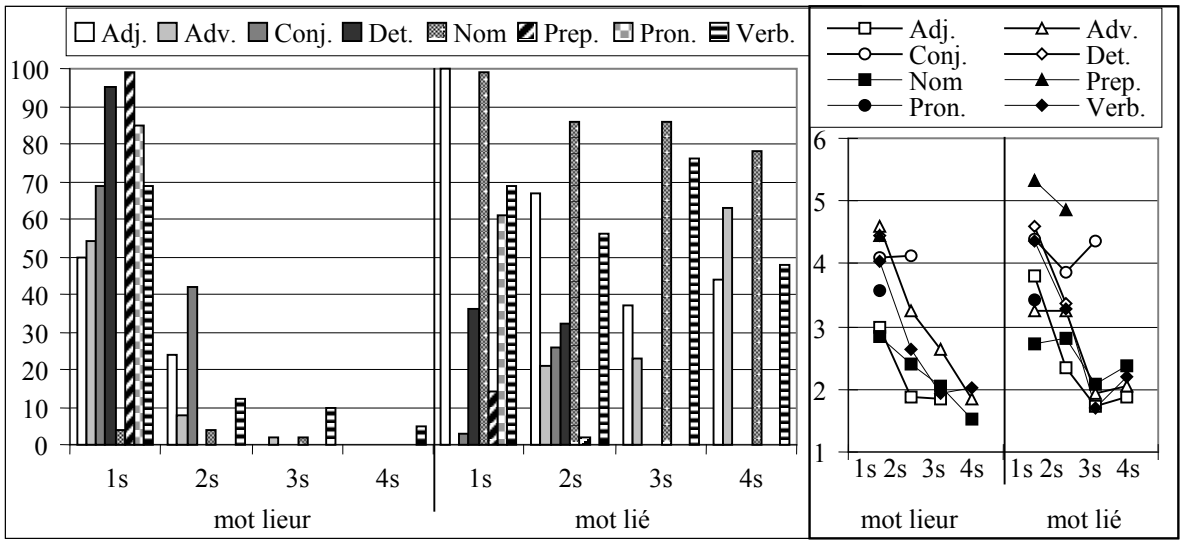

Figure 2: Taux moyen de liaisons en % (à gauche) et fréquence lexicale en log(occurrence) (à droite) des mots lieurs et liés en fonction de leur longueur (s = syllabe) et de leur catégorie grammaticale.

En ce qui concerne la longueur du mot lié, on observe une tendance inverse : les mots monosyllabiques sont moins souvent liés que les 2 à 4 syllabes. Toutefois, il apparaît sur la figure 2 que cet effet est aussi fonction de la catégorie syntaxique : les adjectifs, noms, pronom et verbes monosyllabiques de notre corpus ont un taux supérieur à 50%. De plus, l'effet de

longueur du mot lié apparaît négligeable lorsque l'on compare les proportions de réalisation des liaisons pour une même catégorie syntaxique aux différentes longueurs. Par exemple, les noms sont liés à 80% en moyenne, indépendamment de leur longueur (p.ex « premières années» 100%, « leurs appareils » 75%). Seuls les adjectifs montrent une diminution de leur pourcentage de liaison avec une augmentation de leur longueur. Cependant, cet effet semble être fonction de la nature du lieur précédant ces adjectifs dans notre corpus, avec principalement des monosyllabiques grammaticaux pour les adjectifs courts (p.ex. « très heureux » 100%) et des noms pour les adjectifs longs (p.ex. « défits américains » 5%). En ce qui concerne la longueur de la séquence liée (lieur+lié), on observe que dans des séquences courtes de 2 syllabes, les liaisons sont réalisées plus fréquemment (62% en moyenne) que dans les séquences plus longues. Au delà de 5 syllabes, les liaisons tendent à diminuer fortement, mais cette tendance reste à vérifier compte tenu du faible nombre d'item de ce type dans notre corpus.
En résumé, en ce qui concerne le facteur longueur, la réalisation des liaisons est principalement affectée par le nombre de syllabes du mot lieur avec lequel elle est corrélée négativement.

3.2.3 Facteur fréquence

Afin de mesurer un effet de «fréquence lexicale» sur la réalisation des liaisons, les fréquences d'occurrence des mots lieurs et liés ont été calculées à partir d'une base de données d'articles de presse d'un total de 12MO de motsCes mesures sont exprimées selon le logarithme du nombre d'occurrences pour linéariser les données. Pour la séquence mot1+mot2, deux mesures ont été calculées à partir du même corpus : (1) le nombre de co-occurrence des formes orthographiques m1 et m2 sans marque de ponctuation les séparant (les homographes sont confondus), et (2) la probabilité transitionnelle (occ. séquence/occ. mot1) qui donne la probabilité d'occurrence de la séquence sachant la probabilité d'occurrence du lieur.
Les données observées présentent une corrélation positive significative ($p<.01$) entre le taux de réalisations des liaisons et la fréquence du mot lieur ($r=.6$), du mot lié ($r=.4$) et de la co-occurrence mot1-mot2 ($r=.4$). En revanche, le taux de liaisons n'est pas fonction de la probabilité transitionnelle de la séquence ($r=.1$). La fréquence de la séquence seule est donc importante, indépendamment de la fréquence du lieur. Les liaisons apparaissent donc plus fréquemment lorsque les lieurs et liés sont des mot fréquents et quand ils forment une séquence de mots fréquente dans la langue. Toutefois, cet effet de fréquence est à nuancer par le fait que les caractéristiques de fréquence sont fortement corrélées à celles de longueur ($r=.6$ pour le lieur, et $r=.8$ pour le lié). Comme on peut le voir sur la partie droite de la figure 2, et comme le prédit la loi de Zipf ('49), les mots courts, qu'ils soient lieurs ou liés, sont aussi les plus fréquents. Ainsi en ce qui concerne les mots lieurs, l'influence de la fréquence est similaire à celle observée pour la longueur : les lieurs les plus fréquents (les monosyllabiques) ont un taux de réalisation plus important. En revanche, si la fréquence des mots liés diminue avec leur allongement, il apparaît que les mots moins fréquents et longs (3-4 syllabes) ne sont pas moins liés que les autres (avec par ex. plus de 70% de liaisons pour les noms comme « gros actionnaire », « mes adjectifs »).

4 Discussion et conclusion

A partir des échantillons de parole recueillis auprès de dix locuteurs, une approche statistique globale telle qu'elle est exercée ici permet de mettre en évidence la contribution de certains facteurs sur la réalisation des liaisons. Cependant, ces facteurs ne permettent pas de prédire avec une précision suffisante la probabilité d'occurrence d'une liaison dans un contexte donné. En effet, la fréquence de réalisation de la liaison semble fortement dépendre des items lexicaux impliqués. Nos résultats ont toutefois permis de confirmer l'influence de plusieurs facteurs sur la réalisation des liaisons. Parmi eux, la catégorie syntaxique des mots lieurs et liés apparaît comme essentielle puisqu'elle relativise l'effet des autres facteurs étudiés. Une description de la réalisation des liaisons basée sur l'appartenance catégorielle des mots impliqués, telle qu'elle est

donnée dans la plupart des grammaires, semble donc justifiée. Pourtant, notre étude montre qu'une modélisation des liaisons basée sur ce seul critère n'est pas adéquate au vu de la variabilité dans la réalisation des liaisons au sein même d'une même catégorie grammaticale (voir tableau II). Les facteurs de cohésion syntaxique, de longueur du mot lieur, et de fréquence des lieurs, des liés et de la séquence, doivent être pris en considération dans le modèle. En effet, avec un modèle de régressions multiples incluant ces 5 facteurs, 53% de la variance des taux de réalisation des liaisons peut être expliquée. Cependant, la contribution relative de ces facteurs est à prendre avec réserve compte tenu du manque d'orthogonalité de certains facteurs. Enfin, pour ce qui est du style de parole, nos données montrent que le débit de parole n'a d'effet sur la réalisation des liaisons que pour un locuteur sur 10. La liaison ne semble donc pas contrainte comme d'autres variantes phonologiques (p.ex. l'élision) par les « fast speech rules ». Quant aux variations entre lecture et conversation spontanée, l'examen des productions de dix locuteurs nous permet de mettre en évidence une variabilité inter-individuelle pour ce type d'effet. Un examen plus détaillé des productions spontanées à partir des critères présentés pour le corpus de lecture et incluant l'analyse de facteurs prosodiques nous permettrons, à l'avenir, de valider les résultats présentés ici.

Remerciements

Nous remercions S. Dubé pour son aide. Le 1er auteur est financé par le projet FNRS 1114-059532, le 2ème auteur est en partie financé par le Programme Plurifacultaire «prosodie» de l'Univ. de Genève et le projet CTI n°4607.1.

Références

Adda-Decker M &al (1999): Pronunciation Variants in French: schwa & liaison*, ICPhS 99*

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

Anderson, S. (1975). Liaison in french. *The French review*, 48, 848-855.

Bybee, J. (to appear). Frequency effects on French Liaison. In *Frequency and the Emergence of Language Structure*. J. Bybee and P. Hopper (eds.) Amsterdam: John Benjamins.

Clements, G.N. & Keyser, S.J. (1983). CV phonology: A generative theory of the syllable. Linguistic Inquiry Monograph 9. Cambridge, MA: MIT Press.

Cooper, W.E. & Paccia Cooper J. (1980) Syntax and speech. Cambridge Mass.: Harvard U. Press.

De Jong, D. (1991). La sociophonologie de la liaison orléanaise. Symposium on Romance Languages.

Delattre, P. (1966). *Studies in French and comparative linguistics*. Mouton: The Hague.

Dell, F. (1985). Les règles et les sons: introduction à la phonologie générative. Paris: Hermann.

Encrevé, P.(1988). *La liaison avec et sans enchaînement*. Paris: Seuil

Gaudinat A., Wehrli E. (1997) Analyse syntaxique et synthèse de la parole, *TAL* , vol. 38,n°1:121-134

Inkelas, S. & Zec, D. (1995). Syntax-phonology interface. In J. Goldsmith (ed.) *The Handbook of Phonological Theory*. Cambridge, MA: Blackwell, pp. 535-549.

Jun S.-A. & Fougeron C. (2000), A Phonological model of French intonation. In A. Botinis (ed.) Intonation: Analysis, Modeling and Technology. Dordrecht : Kluwer. pp.209-242

Klausburger, J. (1984) French Liaison and Linguistic Theory. Stuttgart: Franz Steiner.

Lucci, V. (1983). *Etude phonétique du français contemporain à travers la variation situationnelle*. Grenoble: Publ. U. de Grenoble.

Morin YC & Kaye J (1982) The syntactic bases for French liaison. *Jounal of Linguistics* 18 :291-330

Reinhart, T. (1981) Definite NP anaphora and c-command, Linguistic Inquiry 12, 605-635.

Selkirk E. (1974): French liaison and the X-bar convention, *Linguistic Inquiry* 5, pp. 573-590.

Shane, S.A. (1968). French phonology and morphology. Cambridge: MIT Press.

Tranel B. (2000) Aspect de la phonologie du français et la théorie de l'optimalité. *Langue française*

Wright, R. (1998) Factors of lexical competition in vowel articulation, to appear in Labphon6.

Zipf, G.K. (1949). Human Behaviour and the Principle of the Least Effort. Cambridge : Addison-Wesley.