CrossMark

Journal of French Language Studies **27** (2017), 27–40, © Cambridge University Press
doi:10.1017/S0959269516000363

Analyse fréquentielle de la liaison variable dans un corpus de français parlé

GIULIA BARRECA[a] et GEORGE CHRISTODOULIDES[b]

[a] *Université Paris Ouest – MoDyCo*
[b] *Université Catholique de Milan*

(Received January 2016; revised July 2016)

RÉSUMÉ

Notre contribution se propose de mener une analyse fréquentielle de la liaison variable à partir des données de parole semi-spontanée extraites du corpus « Phonologie du Français Contemporain » (Durand, Laks et Lyche, 2002; 2009). Les résultats de notre étude confirment la présence d'une distribution de Mandelbrot-Zipf des réalisations, selon laquelle un nombre limité de séquences permet de rendre compte de plus de la moitié des occurrences. Ces données semblent davantage remettre en discussion un traitement unitaire de la liaison basé sur des règles et apporter des arguments en faveur d'une approche exemplariste d'étude de la liaison.

INTRODUCTION

Les descriptions de la liaison ont été soumises à plusieurs révisions en fonction des méthodologies et des normes de prononciation retenues (Fouché, 1959; Encrevé, 1988 ; Durand et al., 2011). Malgré cette hétérogénéité des traitements classificatoires, tous les auteurs s'accordent à faire la différence entre les liaisons qui se réalisent toujours (appelées obligatoires ou catégoriques selon les auteurs) et celles dont la production est variable de façon plus au moins importante (liaisons facultatives ou variables selon les auteurs) en fonction des différents facteurs phonologiques, prosodiques, sociolinguistiques et lexicales (Encrevé, 1988 ; Durand et Lyche, 2016). La pluralité de facteurs susceptibles d'influencer la réalisation de la liaison rend sa prédiction difficile (Boula de Mareüil, Adda-Decker et Gendner, 2003; Fougeron et al., 2001). Toute la complexité de l'analyse des liaisons variables réside donc dans la prévision de la possibilité plus ou moins forte que ces liaisons soient produites. Plusieurs statuts des consonnes de liaison (Côté, 2005) ont été théorisés selon les différentes approches afin de rendre compte de la réalisation variable de la liaison. Les modèles autosegmentaux proposent un traitement unitaire de la liaison et considèrent la variation de la liaison comme le résultat d'un « double flottement » (squelettal et syllabique) de la consonne de liaison (Encrevé, 1988). En revanche, dans le cadre des approches lexicales (Bybee, 2005 ; Laks, 2005 ; Côté, 2013), deux hypothèses différentes ont été avancées. La première relève des approches exemplaristes d'étude de la liaison (Bybee, 2005; Chevrot, Chabanal et Dugua, 2007) et attribue un rôle central aux variables lexicales comme facteurs

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

Tableau 1. *Calcul de l'indice de la fréquence d'une séquence de liaison d'après Chevrot, Chabanal et Dugua, 2007 (tiré de Côté, 2013).*

Fréquence de la séquence de liaison =	Nombre d'occurrences du M1
	Nombre de M2 différents

explicatifs de la productivité de la liaison. La deuxième est issue de l'étude de Côté (2013) et voit dans les probabilités transitionnelles qui caractérisent les séquences de liaison une possible explication de la fréquence de réalisation de la liaison.

En particulier, Chevrot et al. (2007) proposent de considérer la productivité de la liaison comme le résultat du rapport existant entre le nombre d'occurrences du M1 dans un contexte donné et le nombre de différents M2 susceptibles d'apparaître après M1 à l'intérieur d'une séquence de liaison (cf. Tableau 1).

D'après Côté (2013), les différentes hypothèses de fréquence formulées par les approches exemplaristes n'expliquent pas l'hétérogénéité observée au sein de certains types de liaison (ex. mots invariables monosyllabiques). C'est pour cette raison que Côté (2013) propose une analyse basée sur la probabilité transitionnelle des séquences de liaison à partir d'un corpus de variétés du français laurentien. Elle conduit une étude sur la base d'un ensemble de mots invariables monosyllabiques constitué par des adverbes (« mieux », « moins », « pas », « plus », « très » et « trop ») des prépositions (« chez » et « dans ») et des connecteurs (« mais », « puis », « dont » et « quand »).

À la suite de cette analyse, Côté (2013) suggère d'analyser la productivité d'une séquence de liaison en termes de probabilité transitionnelle des séquences de liaison. L'idée de base est que la fréquence de réalisation d'une liaison entre un M1 et un M2 (ex. « quand il ») est conditionnée par la probabilité transitionnelle qu'un M2 (ex. « il ») ou sa catégorie morphosyntaxique (ex. pronom personnel sujet) suive le M1 (ex. « quand »). Autrement dit, parmi les connecteurs, « quand » et « dont » déclencheraient un nombre plus important de liaisons puisqu'ils sont suivis dans la plupart des cas par les seuls pronoms personnels « elle (s) », « il(s) » et « on ». En revanche, bien que les connecteurs « mais » et « puis » apparaissent plus souvent que « quand » devant ces pronoms, ils possèdent des taux de réalisation de la liaison très faibles puisqu'ils sont susceptibles d'être suivis par un plus grand nombre de M2.

Côté (2013) étend également cette analyse basée sur des probabilités transitionnelles aux séquences de liaison ayant comme M1 des adverbes monosyllabiques. La productivité de la réalisation de la liaison après les adverbes « très » et « plus »[1] pourrait donc être expliquée comme la conséquence de la forte

[1] L'adverbe négatif « plus » n'est pas analysé par cette étude puisqu'il correspond dans la variété de français laurentien à une prononciation distincte [py] qui exclut la réalisation d'une liaison.
« *Plus* 'more': This word poses difficulties since the written form plus corresponds to three different prononunciations: [ply], [plʏs] and [py]. The distribution of the three variants is complex, but may be summarised as follows. The reduced form [py] is found only in negative contexts (e.g. *je parle plus* [py] 'I don't speak anymore'), while the positive contexts

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

probabilité transitionnelle pour ces adverbes d'être suivis par des adjectifs et des adverbes. En revanche, dans le cas des autres adverbes « trop », « moins », « pas » et « mieux », après lesquels la liaison est peu productive, la probabilité transitionnelle d'apparaître devant des adjectifs et des adverbes varie de 6% à 53% puisqu'ils sont également susceptibles de se trouver devant d'autres catégories (ex. prépositions, des verbes, etc.). Pour résumer, d'après l'étude de Côté (2013), les comportements de la liaison à l'intérieur d'une même classe morphosyntaxique (dans ce cas celles des adverbes et des connecteurs monosyllabiques) seraient influencés par la diversité des probabilités transitionnelles qui caractérisent les différentes unités et non pas des variables fréquentielles[2] .

Des analyses quantitatives (Boula de Mareüil, Adda-Decker et Gendner, 2003; Durand et al., 2011; Laks, Calderone et Celata, 2014) ont été également menées sur de grands corpus de français oral afin d'étudier la distribution de la fréquence des réalisations de la liaison (catégorique et variable) selon les contextes morphosyntaxiques de liaison. Ceux-ci sont généralement définis sur la base de la catégorie morphosyntaxique d'appartenance du premier et du deuxième mot de liaison (dorénavant respectivement M1 et M2) qui permettent d'identifier le « type de liaison » (ex. la liaison entre un déterminant (DET) et le nom suivant (NOM)). Ces généralisations regroupent l'ensemble des occurrences de liaison (« tokens ») dont les M1 et M2 relèvent des mêmes catégories morphosyntaxiques (ex. « les amis », « les arbres » etc.).

Les résultats de ces travaux ont montré que la répartition statistique des fréquences d'apparition des réalisations de la liaison catégorique et variable dans les corpus observés suit une distribution statistique de la fréquence décrite initialement par Zipf (1949) et ensuite définie par Mandelbrot (1957) (et ainsi connue sous le nom de distribution de Mandelbrot-Zipf). Celle-ci représente un modèle de la loi de puissance qui définit une relation mathématique entre deux quantités (l'une correspondant à la fréquence d'un évènement et l'autre indiquant la taille d'un évènement) où la décroissance des fréquences d'un évènement est moins importante lorsque la taille de l'évènement augmente.

L'observation de la présence d'une distribution de Zipf (1949) parmi les données de réalisation de la liaison indique qu'un groupe restreint de contextes morphosyntaxiques, à savoir de types de liaison (cf. ci-dessus), correspond à la plupart des réalisations totales de la liaison (catégorique et variable).

are shared between [plʏs] and [ply] (e.g. *je parle plus* [plʏs] 'I speak more'; *plus* [ply] *vite* 'faster'). The form [ply] is also found in a number of locutions (e.g. *plus ou moins* 'more or less', *non plus* 'neither') and it may replace [py] and, in some cases, [plʏs] in very careful registers. Crucially, only [ply] may be followed by a liaison [z] and is relevant here. All instances of *plus* pronounced [py] or [plʏs] were therefore excluded from consideration ». (Côté, 2013 :160).

[2] La seule exception à cette généralisation serait représentée par le taux élevé de réalisations attesté après la forme verbale « est » pour lequel l'influence de la fréquence a été plusieurs fois évoquée dans la littérature.

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

La forte diversité des comportements qui caractérise les réalisations de la liaison variable mise en évidence par ces études, suggère que pour traiter de la liaison, il est nécessaire d'analyser très précisément chaque contexte et chaque occurrence. Ces résultats renforcent l'idée déjà avancée par Morin et Kaye (1982) selon laquelle le contexte syntaxique de réalisation de la liaison doit être défini catégorie par catégorie mais aussi mot par mot, et que la dimension lexicale doit donc être intégrée dans l'étude du contexte syntaxique de réalisation de la liaison.

OBJECTIFS

À la suite des travaux précédents (Boula de Mareüil, Adda-Decker et Gendner, 2003; Laks, Calderone et Celata, 2014), ce travail a pour objectif d'étudier la distribution des seules réalisations de la liaison variable à partir des données de parole semi-spontanée extraites du corpus PFC.

De plus, à travers une analyse des occurrences de liaison variable extraites du corpus PFC, nous nous proposons d'analyser les écarts de réalisation entre les différentes séquences issues (tokens) des mêmes contextes morphosyntaxiques de liaison (type de liaison) afin d'essayer de rendre compte de ces idiosyncrasies. L'observation d'une distribution de Zipf des réalisations des liaisons variables ainsi que la présence de différences importantes de réalisation au sein des mêmes types de liaison pourraient contribuer à remettre en cause un traitement unitaire et homogène de la liaison proposé par les modèles génératifs (Encrevé, 1988).

CORPUS ET MÉTHODOLOGIE

La ressource orale sur laquelle nous nous appuyons est le corpus « Phonologie du Français Contemporain » (Durand, Laks et Lyche, 2009), l'un des plus grands corpus de français parlé issu non seulement de la France hexagonale mais également de différents pays francophones (Belgique, Suisse, Québec, Afrique francophone, Louisiane etc.). Notre échantillon correspond en particulier à 4296 occurrences de liaisons variables réalisées qui ont été extraites des transcriptions des entretiens guidés et des conversations libres de 394 locuteurs natifs issus de 38 points d'enquête (correspondant à un total de 1.368.044 tokens). Dans le cadre de cette étude nous avons décidé d'écarter les données des lectures oralisées pour lesquelles la fréquence de réalisation pourrait avoir été influencée par la nature même de la tâche.

L'ensemble de ces données (ainsi que la totalité du corpus PFC) a été soumis à un important travail de ré-annotation en parties de discours (POS) à l'aide de l'étiqueteur morphosyntaxique *DisMo* (Christodoulides, Avanzi et Goldman, 2014; Christodoulides, Barreca et Avanzi, 2015). La ré-annotation automatique (suivi d'un post-traitement de correction manuelle) nous a permis d'améliorer la précision d'étiquetage des contextes morphosyntaxiques des liaisons (types de liaison). À la suite de ce traitement, nous avons mené une analyse fréquentielle de la répartition des réalisations de la liaison variable par types de liaison à l'aide du package *Zipf-R* développé par Baroni et Evert (2014) (exploitable dans le cadre du

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

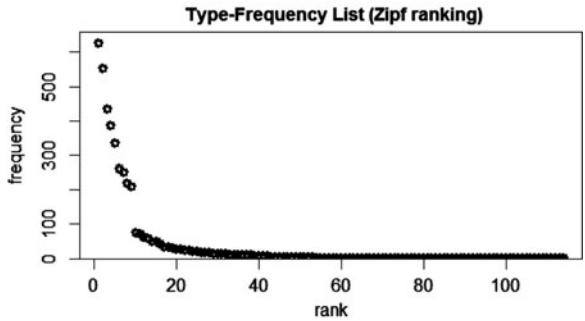


Figure 1. *Distribution de la fréquence des réalisations de la liaison variable dans la parole semi-spontanée du corpus PFC (abscisse : types de liaison ; ordonnée : tokens).*

logiciel statistique R) afin de vérifier la présence d'une distribution de Zipf des réalisations des liaisons variables.

RÉSULTATS

Les résultats qui suivent sont issus d'une part de l'analyse de la distribution de la fréquence des liaisons variables réalisées par type de liaison menée par biais du package *Zipf R* et d'autre part de l'étude des idiosyncrasies qui caractérisent la réalisation de la liaison. Ces données nous permettront à la fois de vérifier le degré d'homogénéité de la répartition de la réalisation de la liaison variable entre les différents types de liaison et au sein de ces mêmes contextes morphosyntaxiques.

Distribution de la liaison variable

Les données issues de l'étude de la distribution des fréquences des liaisons variables réalisées sont illustrées dans la figure ci-dessous (cf. Figure 1). Sont représentés en abscisse les types (*rank* correspondant à 114 types différents de liaison) et en ordonnée les tokens (à savoir les occurrences de liaison réalisées représentant 4296 tokens). Les 9 premiers types, c'est-à-dire les 9 contextes de liaison variable les plus fréquents sur un total de 114 types, rendent compte de 76% des réalisations totales (à savoir 3280 / 4296 tokens).

Ces résultats montrent que la distribution des réalisations de la seule liaison variable en parole semi-spontanée suit également une distribution de Zipf (1949), déjà observée par les études précédentes sur l'ensemble des types de liaison (catégoriques et variables).

Autrement dit, un nombre restreint de types (9 types de liaison) représente la presque totalité des réalisations de la liaison variable. Ceux-ci correspondent à des séquences constituées aussi bien de mots issus des classes fermées (prépositions,

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

conjonctions et adverbes) que des classes ouvertes (verbes et adjectifs)[3] . Le tableau ci-dessous (cf. Tableau 2) résume ces types et le nombre de réalisations de la liaison correspondantes (tokens).

L'observation d'une telle répartition des fréquences indique qu'il existe de nombreux types possédant très peu de réalisations (n. tokens < 100) de liaison ; en revanche, le nombre de types qui possèdent un taux élevé de réalisation de liaison est limité.

Afin de vérifier la validité de ces résultats, nous avons mené à l'aide du package *Zipf-R* (Baroni et Evert, 2014) une analyse d'ajustement (*goodness of fit*) d'un modèle statistique paramétrique *Large Number of Rare Events model* (dorénavant modèle LNRE) (Baayen, 2001) pour la distribution des fréquences d'une population (en l'occurrence, la fréquence de réalisation de la liaison) aux données observées. Cette analyse statistique paramétrique basée sur un test multi-varié du *khi* deux (Baayen, 2001) nous permet d'estimer la distance existant entre la distribution observée et celle attendue sur la base du modèle LNRE. L'ajustement du modèle statistique aux données observées est d'autant plus valide que le *khi* carré (x^2) est petit et que la valeur *p* d'ajustement des données de liaison au modèle LNRE est importante. Autrement dit, la valeur *p* doit être plus grande que 0,05 ($p > 0,05$) pour que l'hypothèse soit vérifiée. L'estimation de l'ajustement (*goodness of fit*) du modèle testé montre que le coefficient du *khi* carré est très petit: 1,7. En revanche, la valeur *p* est beaucoup plus grande que 0,05 ($p = 0,6$). De ces résultats, nous pouvons conclure que la réalisation de la liaison variable observée suit une loi de puissance de type Mandelbrot-Zipf.

En effet, si nous examinons de plus près les résultats de ces types de liaison variable, nous observons une forte idiosyncrasie inter-catégorielle. Ces différences sont évidentes si nous comparons les résultats des 9 premiers types de liaison issus de l'analyse fréquentielle (en gras dans le tableau ci-dessous) (cf. Tableau 3) avec ceux des types similaires appartenant au même contexte morphosyntaxique de liaison tel qu'il est décrit par les classifications traditionnelles de la liaison (ex. « CON-PRO »). Le tableau ci-dessous résume (cf. Tableau 3) pour chaque type le nombre de liaisons réalisées (LR), non-réalisées (LNR) ainsi que le taux de réalisation de la liaison (%LR).

Les données issues de cette analyse confirment les tendances déjà observées par les études précédentes menées sur des corpus oraux (Mallet, 2008 ; Côté, 2013 ; Durand et al., 2011). La liaison est plus fréquente au présent (« VER:pres ») (en particulier après les formes du verbe « être ») qu'après les autres temps verbaux. En revanche la liaison après les adverbes de négation (« ADV:neg ») est caractérisée par un taux de réalisation faible.

Cependant ces résultats ne nous disent rien quant aux fréquences de réalisation au sein de chaque type de liaison. En effet, comme nous l'avons

[3] Le jeu d'étiquette utilisé pour l'annotation des types de liaison est celui proposé par l'annotateur morphosyntaxique *DisMo* (cf. Christodoulides, Avanzi et Goldman, 2014 ; Christodoulides, Barreca et Avanzi, 2015).

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

Tableau 2. *Types de liaison variables réalisées les plus fréquents (n. tokens > 100).*

	Type	Token		Type	Token
CON:sub-PRO:per:sjt	Conjonction de subordination – Pronom personnel sujet	627	PRP-NOM:pro	Préposition-Nom	263
PRP-DET:ind	Préposition-Déterminant indéfini	553	ADV:deg-ADJ	Adverbe d'intensité-Adjectif	251
VER:pres-DET:ind	Verbe présent-Déterminant indéfini	435	VER:pres-PRP	Verbe présent-Préposition	220
PRP-NOM:com	Préposition-Nom commun	386	ADJ-NOM:com	Adjectif épithète-Nom commun	209
VER:pres:aux-VER:ppas	Verbe auxiliaire au présent-Verbe au participe passé	336			

Tableau 3. *Comparaison entre les types de liaison variable les plus fréquents (n. tokens > 100) (en gras) et les types issus des mêmes classes morphosyntaxiques.*

	Type	LR	LNR	%LR
CON-PRO	**CON:sub-PRO:per:sjt**	**627**	**205**	**75%**
	CON:coo-PRO:per:sjt	1	857	0%
PRP-DET	**PRP-DET:ind**	**553**	**159**	**78%**
VER-DET	**VER:pres-DET:ind**	**435**	**1476**	**23%**
	VER:cond-DET:ind	0	37	0%
	VER:fut-DET:ind	0	13	0%
	VER:impe-DET:ind	0	7	0%
	VER:impf-DET:ind	34	976	3%
	VER:inf-DET:ind	2	283	1%
	VER:ppas-DET:ind	12	285	4%
	VER:ppre-DET:ind	1	7	13%
	VER:simp-DET:ind	0	4	0%
	VER:subp-DET:ind	1	26	4%
PRP-NOM	**PRP-NOM:com**	**386**	**10**	**97%**
	PRP-NOM:pro	**263**	**23**	**92%**
VER:aux-VER	**VER:pres:aux-VER:ppas**	**336**	**735**	**31%**
	VER:cond- VER:ppas	0	45	0%
	VER:fut- VER:ppas	0	2	0%
	VER:impe- VER:ppas	1	2	0%
	VER:impf- VER:ppas	17	331	5%
	VER:subp- VER:ppas	0	11	0%
ADV-ADJ	**ADV:deg-ADJ**	**251**	**46**	**85%**
	ADV:neg-ADJ	1	106	0%
VER-PRP	**VER:pres-PRP**	**1152**	**220**	**16%**
	VER:cond-PRP	0	15	0%
	VER:fut-PRP	0	7	0%
	VER:impe-PRP	2	23	8%
	VER:impf-PRP	21	922	2%
	VER:inf-PRP	1	552	0%
	VER:ppas-PRP	13	341	4%
	VER:ppre-PRP	1	9	10%
	VER:simp-PRP	0	2	0%
	VER:subp-PRP	0	6	0%
ADJ-NOM	**ADJ-NOM**	**46**	**209**	**82%**

évoqué précédemment une description basée uniquement sur les contextes morphosyntaxiques de liaison ne permet pas de prédire avec une précision suffisante la réalisation de la liaison (Boula de Mareüil, Adda-Decker et Gendner, 2003). Il nous est donc apparu nécessaire de mener une étude plus approfondie des occurrences de liaison afin de mettre en évidence les éventuelles idiosyncrasies intra-catégorielles pour les types de liaison décrits dans le tableau 3.

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

Idiosyncrasies de la liaison variable

Les données issues de l'analyse conduite à l'intérieur des types variables les plus fréquents montrent qu'un nombre restreint d'items lexicaux permet de rendre compte de plus de la moitié des productions à l'intérieur du même type de liaison. Ces unités ne relèvent pas seulement des classes fermées mais également des catégories morphosyntaxiques telles que les adjectifs (cf. type « ADJ-NOM ») et les verbes (cf. types « VER-DET » et « VER -PRP »).

Le tableau ci-dessous (cf. Tableau 4) résume les séquences observées pour les types les plus fréquents de liaison variable ainsi que le pourcentage de réalisation (« % sur LR tot ») associé à ces séquences productives (ex. « quand-t-PRO ») par rapport au nombre total de réalisations au sein du même type (ex. « CON-PRO »).

Les seules séquences décrites dans le tableau ci-dessus (cf. Tableau 4) correspondent à 64% des réalisations totales de la liaison variable (à savoir 2755 occurrences de liaisons réalisées sur 4296 liaisons totales réalisées).

DISCUSSION

L'analyse fréquentielle menée sur la parole semi-spontanée dans le corpus PFC nous a permis de montrer que la liaison variable est aussi limitée à un petit nombre de types qui représentent 76% des réalisations totales. De plus au sein de ces types de liaison variable, plus de la moitié des réalisations correspondent à un nombre restreint de séquences très productives. Il reste cependant à identifier les causes de cette forte idiosyncrasie. Comme nous l'avons déjà souligné, les résultats de l'étude menée par Côté (2013) suggèrent que la seule variable de la fréquence ne permet pas d'expliquer les écarts de réalisation observés au sein de certains types de liaison (ex. mots invariables monosyllabiques). Selon cette approche, la productivité d'une séquence de liaison serait conditionnée par la probabilité transitionnelle qu'un M2 (ex. « il ») ou sa catégorie morphosyntaxique (ex. pronom personnel sujet) suive le M1 (ex. « quand »).

A la suite de notre étude nous proposons d'analyser les tendances dégagées par l'étude des probabilités transitionnelles (c.f. introduction) comme le résultat de la diversité qui caractérise les rapports de dépendances syntaxiques des mots de liaison. En effet, si nous reprenons l'analyse des adverbes monosyllabiques décrite dans la première partie de notre travail, l'adverbe « pas » est le plus souvent syntaxiquement gouverné par le verbe précédent alors que « très » dépend de l'adjectif suivant avec lequel il réalise la liaison de façon presque systématique. De même, la probabilité transitionnelle des adverbes « moins » et « trop » peut varier en fonction des valeurs de dépendance différentes de ces adverbes, qui peuvent être gouvernés par le verbe comme par l'adjectif. Il en va de même pour l'adverbe « plus » qui, dans des variétés autres que celle du français laurentien, est également susceptible de déclencher une liaison lorsqu'il possède une valeur négative (ex. « ils ont mal, ils arrivent plus [z] à marcher » Corpus PFC, Grenoble, 38aca1gg). En effet, si on adopte

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

Tableau 4. *Séquence internes aux types de liaison variable les plus fréquents (n. tokens > 100).*

Type	Constructions	% sur LR tot.	Type	Constructions	% sur LR tot.
CON-PRO (tot. CON = 5)	**Quand** - t - PRO	100%	**ADV-ADJ (tot. ADV monosyll = 5)**	**Très** - z -ADJ **Plus** - z - ADJ	96%
PRP-DET (tot. PRP monosyll = 7)	**Dans** - z - DET	92%	**VER-PRP (tot. VER = environ 400)**	**Est** - t -PRP **Sont** - t - PRP	93%
VER-DET (tot. VER = environ 300)	**Est** - t - DET **Était** - t -DET **Fait** - t -DET	95%	**ADJ-NOM (tot. ADJ = 53)**	**Bon** - n - NOM **Autres** - z -NOM **Petit** - t - NOM **Petit(e)s** - z - NOM **Grand** - t - NOM	77%
PRP-NOM (tot. PRP monosyll= 8)	**En**- n -NOM	96%			

une approche purement probabiliste, la présence de cette position comme site de liaison potentielle réduirait les probabilités transitionnelles de « plus » d'être suivi d'un adjectif ou d'un adverbe.

Quant aux connecteurs, il nous semble que la diversité des probabilités distributionnelles observée entre d'une part « quand » et « dont » et d'autre part « mais » et « puis » est la conséquence des rapports de subordination et coordination qui opposent ces connecteurs. La plus grande probabilité que le connecteur « quand » soit suivi par un pronom clitique (ce qui d'après Côté favoriserait la réalisation de la liaison) (cf. Côté, 2013 : pp. 162–163), pourrait donc être liée au fait qu'il introduit une proposition qui est syntaxiquement dépendante de l'énoncé principal. En revanche, la coordination, du fait que la proposition coordonnée est plus indépendante que la subordonnée, permettrait plus facilement l'ajout d'autres unités (ex. prépositions, adverbes, conjonctions) entre les connecteurs de coordination (« mais » et « puis ») et les sujets de la deuxième proposition.

En ce qui concerne la liaison après les adjectifs épithètes, si la probabilité transitionnelle représentait un facteur explicatif capable de distinguer le comportement de la liaison de l'adjectif antéposé de celui de la liaison après adjectif postposé au nom (cf. Côté, 2013), nous devrions alors pouvoir observer une différence de fréquence relative de réalisation de la liaison entre les adjectifs majoritairement antéposés (ex. « grand », « petit » etc.) et ceux ayant une distribution complémentaire (ex. « ancien », « prochain » etc.) (Benzitoun, 2014). Cependant, cette hypothèse ne semble pas être vérifiée. En effet, dans une étude récente menée sur un échantillon de 403 sites de liaison potentielle après adjectifs singuliers et pluriels,[4] Barreca et al. (2015) ont montré que les adjectifs singuliers majoritairement antéposés (« bon », « grand » et «petit » etc.) rendent compte d'environ 80% des liaisons non-réalisées totales (à savoir 10/13 occurrences). En revanche, d'autres adjectifs singuliers (parmi lesquels « ancien » et « prochain » qui possèdent une distribution complémentaire) présentent des taux faibles de non-réalisation de la liaison.

Pour conclure, ces observations nous suggèrent que si une analyse syntaxique superficielle des catégories morphosyntaxique d'appartenance des mots de liaison est insuffisante, une étude plus large des rapports de dépendances syntaxiques qui gouvernent les mots de liaison permettrait, au moins en partie, de rendre compte de la diversité de comportements qui caractérisent les types de liaison décrits par les classifications traditionnelles (Delattre, 1947 ; Encrevé, 1988). Cela ne revient pas à refuser une analyse basée sur les probabilités transitionnelles mais simplement à la traiter comme la conséquence de différences syntaxiques plus profondes.

CONCLUSIONS

À l'issue de ce travail, nous espérons avoir montré que l'analyse fréquentielle menée sur les corpus PFC ré-annotés a contribué d'une part à décrire des idiosyncrasies

[4] Extraites des données de parole semi-spontanée du corpus CFPP2000 (Branca-Rosoff et al., 2012) et du corpus PFC (Durand, Laks et Lyche, 2009).

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

internes aux contextes morphosyntaxiques de liaison, en dépassant les limites d'une classification basée sur les catégories morphosyntaxiques d'appartenance des mots de liaison. D'autre part, la distribution fréquentielle (distribution de Zipf) de la liaison variable, mise en évidence par notre analyse, remet davantage en discussion le traitement unitaire de la liaison proposé par les modèles phonologiques génératifs. Comment des analyses basées sur des règles (Schane, 1965 ; Dell, 1973) ou encore sur la notion de flottance squelettale et syllabique (Encrevé, 1988) pourraient rendre compte des écarts de réalisations observés au sein des séquences différentes de liaison issues des mêmes contextes morphosyntaxique ?

Afin de décrire l'hétérogénéité et la productivité de la liaison, il nous semble que, comme l'a suggéré Côté (2005), il serait opportun d'éviter une approche unitaire d'analyse de la liaison et surtout de la liaison variable. Comme nous l'avons montré, l'étude de la liaison variable reçoit un traitement différent selon le contexte. Si la liaison après les conjonctions et les prépositions monosyllabiques semble être dans l'ensemble assez homogène, d'autres, comme la liaison après des adverbes monosyllabiques, demandent une analyse syntaxique des dépendances qui dépassent les limites d'une classification basée sur les catégories morphosyntaxiques d'appartenance. En revanche d'autres types de liaison semblent être conditionnés principalement par des facteurs d'ordre lexical. C'est le cas par exemple de la liaison après des formes verbales où les seules formes « est » et « était » correspondent à la presque totalité des réalisations, indépendamment du contexte syntaxique dans lequel elles se trouvent (VER-DET, VER-PRP cf. Tableau 4). En ce qui concerne la liaison après les adjectifs épithètes, ni la distribution ni la probabilité transitionnelle n'influencent la productivité de la liaison puisque, comme nous l'avons vu, la liaison n'est pas plus souvent réalisée après des adjectifs majoritairement antéposés qu'après des adjectifs en distribution complémentaire. Si le facteur lexical ne semble pas pouvoir expliquer la variabilité de la liaison après les adjectifs épithètes singuliers (les adjectifs les plus fréquents « grand » et « petit » présentent le taux le plus important de non-réalisations), le nombre limité d'adjectifs susceptibles d'être antéposés à l'oral (Bybee, 2005 ; Durand et Lyche, 2008; Benzitoun, 2014 ; Barreca et al., 2015) confirme cependant une lexicalisation sous forme de constructions (Bybee, 2005) de ces contextes de liaison. En revanche, la liaison après les adjectifs pluriels, laquelle constitue une marque du pluriel, laisse la place à plusieurs traitements qui considèrent la consonne de liaison comme autonome par rapport au M1 et M2 (épenthétique, morphémique et exemplariste) (cf. Côté, 2005).

L'ensemble de ces considérations nous amène à nous interroger sur un traitement unitaire de la liaison variable basé sur des catégories morphosyntaxiques ne pourrait pas être complété par une étude plus fine des constructions (« dans - z - DET », « est - t - DET » , « était - t - PRP », « grand - t - NOM » etc.) qui émergent de notre analyse fréquentielle. En effet, indépendamment de leur type d'appartenance, ce petit nombre de constructions (cf. Tableau 4) représente 64% des réalisations totales de liaison variable. Les occurrences restantes de liaison, comme nous l'avons vu dans les pages précédentes, relèvent d'un nombre important de types différents. À la suite de Laks (2005), nous pouvons supposer que la production de ces liaisons variables

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

est davantage tributaire de l'influence des représentations orthographiques de mots de liaison. Ainsi la fréquence la plus importante de certaines occurrences de liaison attestées par la présente étude pourrait s'expliquer par une plus grande accessibilité de ces constructions favorisée par l'ensemble des facteurs lexicaux, syntaxiques et distributionnels que nous avons évoqués précédemment. En revanche, dans le cas des occurrences de liaison moins fréquentes, se situant au milieu et à la fin de la courbe distributionnelle mise en évidence dans cette analyse (cf. Figure 1), la production s'appuierait davantage sur des représentations orthographiques isolées des mots stockées dans le lexique mental du locuteur.

RÉFÉRENCES

Baayen, H. R. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Baroni, M. et Evert, S. (2014). The zipfR package for lexical statistics: A tutorial introduction. URL: http://zipfr.r-forge.r-project.org/materials/zipfr-tutorial.pdf, consulté le 14 février 2015.

Barreca, G., Christodoulides, G., Durand, J., Laks, B. et Lyche, C. (2015). La liaison de l'adjectif antéposé dans le français parlé. *Journées de l'Association of French Language Studies (AFLS 2015)*, Caen.

Benzitoun, C. (2014). La place de l'adjectif épithète en français: ce que nous apprennent les corpus oraux. *Actes du 4ème Congrès Mondial de Linguistique Française, CMLF2014*. Berlin, pp. 2333–2348.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F. et Pires, M. (2012). Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000), URL: http://cfpp2000.univ-paris3.fr/Corpus.html, consulté le 31 juillet 2015.

Boula de Mareüil, P., Adda-Decker, M. et Gendner, V. (2003). Liaisons in French: a corpus-based study using morpho-syntactic information. *Proceedings of 13th International Congress of Phonetic Sciences, ICPhS2003*. Barcelone, pp. 1329–1332.

Bybee, J. (2005). La liaison: effets de fréquence et constructions. *Langages*, 158: 24–37.

Chevrot, J.-P., Chabanal, D. et Dugua, C. (2007). Pour un modèle de l'acquisition des liaisons basé sur l'usage: trois etudes de cas. *Journal of French Language Studies*, 17.1: 103–128.

Christodoulides, G., Avanzi, M. et Goldman, J.-P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech. *Proceedings of 9th International Conference on Language Resources and Evaluation, LREC2014*. Reykjavik, pp. 3902–3907.

Christodoulides, G., Barreca, G. et Avanzi, M. (2015). DisMo: un annotateur multi-niveaux pour les corpus oraux. *Actes du 22ème colloque Conférence sur le Traitement Automatique des Langues Naturelles*, T*ALN2015*. Caen, France.

Côté, M.-H. (2005). Le statut lexical des consonnes de liaison. *Langages*, 158 : 66–78.

Côté, M.-H. (2013). Understanding cohesion in French liaison. *Language Sciences*, 39 : 156–166.

Delattre, P. (1947). La Liaison en Français, Tendances et Classification, *The French Review*, 21.2 : 148–157.

Dell, F. (1973). *Les règles et les sons*. Paris : Hermann.

Durand, J., Bernard, L. et Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In: C. Pusch et W. Raible (dir.), *Romanistische*

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press

Korpuslinguistik-Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language. Tübingen: Gunter Narr, pp. 93–106.
Durand, J. et Lyche, C. (2008). French liaison in the light of corpus data. *Journal of French Language Studies*, 18.1: 33–66.
Durand, J. et Lyche, C. (2016). Approaching variation in PFC: the liaison level. In: J. Durand, B. Laks et C. Lyche (dir.), *Varieties of Spoken French*. Oxford: Oxford University Press.
Durand, J., Laks, B., Calderone, B. et Tchobanov, A. (2011). Que savons-nous de la liaison aujourd'hui;? *Langue Française*, 169 : 103–136.
Durand, J., Laks, B. et Lyche, C. (eds) (2009). *Phonologie, variation et accents du français*. Paris: Hermès.
Encrevé, P. (1988). *La liaison avec et sans enchaînement: phonologie tridimensionnelle et usages du français*. Paris: Seuil.
Fouché, P. (1959). *Traité de prononciation française*. Paris: Klincksieck.
Fougeron, C., Goldman, J.-P., Dart A., Guelat, L. et Jeager, C. (2001). Influence de facteurs stylistiques, syntaxiques et lexicaux sur la réalisation de la liaison en français. *Actes de la 8ème Conférence Traitement Automatique des Langues Naturelles*, Tours, France.
Laks, B. (2005). La liaison et l'illusion. *Langages*, 158: 101–125.
Laks, B., Calderone, B. et Celata, C. (2014). French liaison and the lexical repository. In: C. Celata et S. Calamai (dir.), *Advances in Sociophonetics*. Amsterdam: John Benjamins, pp. 29–54.
Mallet, G.-M. (2008). La liaison en français: descriptions et analyses dans le corpus PFC. Thèse de doctorat, Université Paris Ouest Nanterre La Défense, Nanterre.
Mandelbrot, B. (1957). *Linguistique statistique marcoscopique. Théorie mathématique de la loi de Zipf (basée sur les méthodes de la théorie de l'information et de la thermodynamique)*. Paris: Institut Henri Poincaré, Seminaire de calcul des probabilites.
Morin, Y.-C. et Kaye, J. D. (1982). The syntactic bases for French liaison. *Journal of Linguistics*, 182: 291–330.
Schane, S. A. (1965). The phonological and morphological structure of French. Thèse de doctorat, MIT, Cambridge MA.
Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of 12th International Conference on New Methods in Language Processing, NeMLaP1994*. Manchester, Royaume Uni.
Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge: Addison-Wesley Press.

https://doi.org/10.1017/S0959269516000363 Published online by Cambridge University Press