Le “ h aspiré ” à l’état sauvage : décrire la disjonctivité dans les grands corpus de parole naturelle

Adèle Jatteau, Nicolas Audibert, Martine Adda-Decker, Lori Lamel, Eric Bilinski

Le « h aspiré » à l'état sauvage : décrire la disjonctivité dans les grands corpus de parole naturelle

Adèle Jatteau[1,*], *Nicolas* Audibert[2], *Martine* Adda-Decker[2], *Lori* Lamel[3] et *Eric* Bilinski[3]

[1]Laboratoire Savoirs, Textes, Langage (UMR 1863) - Université de Lille
[2]Laboratoire de Phonétique et Phonologie (UMR 7018), Université Sorbonne Nouvelle
[3]Laboratoire Interdisciplinaire des Sciences du Numérique (UMR 9015), Université Paris-Saclay

[*]a.jatteau@gmail.com

Résumé. Les mots à « h aspiré » ou « disjonctifs » en français forment un phénomène multifactoriel difficile à décrire : ils sont rares dans le discours, et sont associés à une charge prescriptive qui influence les locuteurs testés en laboratoire. Cette étude propose d'étudier la disjonctivité dans de grands corpus de parole naturelle, à l'aide des outils de traitement automatique de la parole. La question centrale est un point qui divise les travaux de modélisation théorique de la disjonctivité : s'agit-il d'une propriété lexicale catégorique, ou bien d'une propension variable selon les mots ? Les résultats suggèrent que deux sources de disjonctivité doivent être distinguées : la disjonctivité lexicale, qui apparaît comme largement catégorique, et une disjonctivité d'origine pragmatique, sémantique ou syntaxique, manifestée sous la forme d'une rupture prosodique avant le mot disjonctif. Cette analyse, bien que pondérée par les limitations de l'étude, oriente les futures recherches vers une analyse prosodique des contextes de disjonctivité.

Abstract. Words with "h aspiré" or "disjunctive words" in French constitute a multifactorial phenomenon which is difficult to describe: they are rare in speech, and they are associated with a prescriptive charge that influences the speakers tested in the laboratory. This study proposes to investigate disjunctivity in large corpora of natural speech, using automatic speech processing tools. The central question is one that divides work on the theoretical modeling of disjunctivity: is disjunctivity a categorical lexical property, or a word-variable gradient propensity? The results suggest that two sources of disjunctivity need to be distinguished: lexical disjunctivity, which appears to be largely categorical, and a disjunctivity of pragmatic, semantic or syntactic origin. This analysis, although tempered by the study's limitations, points the way for future research towards a prosodic analysis of disjunctive contexts.

1 Introduction

Le phénomène du « h aspiré » en français désigne les mots comme *héros* ou *hamac* qui bloquent des processus de sandhi externe comme la liaison (*les* | *héros* vs. *le*[z] *étaux*) ou l'effacement de schwa (*l*e *héros* vs. *l'étau*), alors qu'ils commencent phonétiquement par une voyelle. Le terme « h aspiré » des grammaires classiques du français (Fouché 1959) étant particulièrement malvenu – il n'y a pas de fricative glottale /h/ en français contemporain, et le comportement touche aussi des mots sans <h> graphique initial, comme *onze* -, nous lui préfèrerons le terme de « disjonction » (Cornulier 1981).

Alors que le français est connu pour préférer l'enchaînement syllabique CV au détriment des frontières de mot, les mots disjonctifs vont à rebours de cette tendance : ils bloquent la liaison (Tab. 1a), exigent la réalisation du schwa dans les clitiques Cə (Tab. 1b), ne permettent pas l'élision du [a] de *la*, ou encore celle du [y] de *tu* dans le registre familier (Tab. 1c), et sélectionnent les formes pré-consonantiques des adjectifs comme *beau/bel* ou *vieux/vieil* (Tab. 1d). De manière variable, ils peuvent également bloquer l'enchaînement (Tab. 1e), et Cornulier (1981) mentionne l'impossibilité de semi-consonantiser une voyelle précédant un mot disjonctif (Tab. 1f).

Tableau 1. Comportements des mots disjonctifs en sandhi externe qui les distinguent des mots #V-.

		#V-	**Disjonctifs**	**#C-**
a.	Liaison	*le*[z] *étaux*	*les* | *héros*	*les métaux*
b.	Schwa dans Cə	*l'étau*	*le héros*	*le métal* (*l'métal* possible)
c.	Elision de [a, y]	*l'amie* *t'aimes*	*la hache* *tu hais*	*la tache* *tu sèmes*
d.	Supplétion	*bel ami*	*beau hameau*	*beau tamis*
e.	Enchaînement	*cin.q a.mis*	*cinq . ha.meaux* *?cin.q ha.meaux*	*cinq . ta.mis*
f.	Semi-consonantisation	*qu[j] avale*	*qui hasarde*	*qui varie*

Ces différents comportements rapprochent les mots disjonctifs des mots à initiale consonantique (Grevisse & Goosse 2008), et ont conduit à des analyses représentationnelles de la disjonctivité comme une consonne sous-jacente sans réalisation phonétique (ex. Schane 1978). Toutefois, plusieurs différences avec les mots à initiale consonantique ont été signalées dans la littérature. Contrairement à ces derniers, les mots disjonctifs demandent la réalisation d'un schwa ou bien d'un coup de glotte après les mots comme *le* ou *une* (Tab. 2a ; voir Gabriel & Meisenburg 2009 pour le détail phonétique de ces réalisations). Cornulier (1981) signale également qu'ils se combinent difficilement avec des préfixes à finale consonantique.

Tableau 2. Comportements des mots disjonctifs en sandhi interne et externe qui les distinguent des mots #V- et #C-.

		#V-	Disjonctifs	#C-
a.	Schwa	*un(e) amie*	*un*[ə] *housse* ou *un*[ʔ] *housse*	*un(e) tache* (*un*[ə] *tache* possible)
b.	Composition	*ex-ami*	*?ex-héros*	*ex-belle-sœur*

Ces différents comportements en sandhi conduisent de nombreux auteurs à caractériser la disjonctivité plutôt comme une contrainte d'alignement entre le début du mot et le début d'une syllabe (Dell 1973, Côté 2008, Zuraw & Hayes 2017 ; voir Zuraw & Hayes 2017 pour une revue récente des différentes approches proposées). Notons que les mots à semi-voyelle initiale présentent à leur tour des comportements spécifiques, que nous n'aborderons pas dans cette étude.

Bien que membre de la « Sainte Trinité de la phonologie française » (Côté 2008), la disjonctivité a fait couler moins d'encre que la liaison et le schwa dans la littérature. Une première difficulté est qu'elle cumule sa propre complexité à celle de ces deux autres grands sujets de la phonologie française, comme le montre le Tab. 1. Mais la difficulté principale réside dans l'importante variation de la disjonction rapportée dans la littérature : toutes les études font en effet état d'une grande variabilité des usages, sous l'effet d'une non moins grande variété de facteurs. Au centre du cercle se trouvent les mots « idiosyncratiquement disjonctifs », comme *hasard* et *héros*, qui peuvent être listés comme disjonctifs dans le dictionnaire. Sont recensés également dans les grammaires les chiffres comme *un, huit* ou *onze* et leurs dérivés (*le onzième jour*), qui sont régulièrement disjonctifs (Grevisse & Goosse 2008). On mentionne en outre le cas des noms de lettres (*le A*), des acronymes (*la SNCF*), et Plénat (1995) signale le cas des mots de verlan (*un* | *ouf*). Plusieurs facteurs d'ordre sémantique ou pragmatique peuvent en outre conduire à l'emploi disjonctif de mots non recensés comme canoniquement disjonctifs. Le Grevisse mentionne ainsi les mots en emploi autonymique (*le diminutif de aube*), et signale une tendance à traiter les noms propres comme disjonctifs (*le livre de Unamuno* ; Grevisse & Goosse 2008). Dans la même veine, les mots techniques, nouveaux ou archaïques relèveraient du même « besoin de démarquer l'expression ou le mot », en transposant dans la phonologie « une espèce d'hétérogénéité de la chose prise pour le nom avec la phrase qui se l'incorpore » (Cornulier 1981).

En plus de ces différents facteurs touchant les mots disjonctifs, la littérature relève une variation reposant sur le mot précédent (« mot 1 »). Cornulier (1981) considère par exemple que *hasard* n'est pas de la même classe que *héros* parce qu'il tolère l'enchaînement, comme dans *par hasard* ; d'après Fouché (1956), *oui*

est disjonctif, mais peut ne pas l'être après *que*. Certaines constructions syntaxiques, comme *vers les une heure*, exigeraient la disjonction. Le type de sandhi externe est aussi mentionné : la résistance à la resyllabation est plus nette en contexte de liaison ou de schwa qu'en contexte d'enchaînement (Côté 2008, Göhring 2017). Les mots courts marquent plus facilement la disjonction : pour Cornulier (1981), *chapitr(e) onze mille* est ainsi plus acceptable que *chapitr(e) onze* ; le nombre de consonnes à la fin du mot 1 jouerait également un rôle. Dans une tâche de lecture de mots et de non-mots, Tessier et al. (2023) constatent en effet cette influence de la longueur du mot, ainsi qu'une plus grande propension à la disjonctivité lorsque le mot 2 commence par [u-] par opposition à [a-]. Les bases seraient plus facilement disjonctives que les dérivés, avec *le*[z] *handicapés* à côté de *le handicap*. A cette variation sur les propriétés phonologiques, morphologiques, syntaxiques et sémantiques du mot 1 et du mot 2, s'ajoute une variation sociolinguistique importante. « La langue populaire », dit le Grevisse, « ne respecte guère la disjonction devant le *h* aspiré » (Grevisse & Goosse 2008 : 55) ; le site de l'Académie Française consacre une page au statut disjonctif du mot *haricot*, sans parvenir à mettre un terme aux débats animés des blogs sur internet. Les expériences de Green & Hintze (2004), Gabriel & Meisenburg (2009), Tessier et al. (2023) ou encore Scheer (à paraître) montrent des jugements et des productions très variés selon les locuteurs et les mots. Göhring (2019) parle de « véritable doute linguistique » ; Cornulier (1981) propose la (triste) conclusion que « beaucoup de gens censés bien parler semblent chaque fois tirer à pile ou face ».

La propension d'un item lexical à se comporter comme disjonctif est-elle vraiment aléatoire ? Cette question divise les traitements formels de la disjonction en deux grandes catégories. D'un côté, de nombreuses études reposent (implicitement) sur un comportement phonologique catégorique des items lexicaux, donc sur une uniformité de comportement des mots disjonctifs (par exemple Pagliano 2003, Boersma 2007, Côté 2008, Gabriel & Meisenburg 2009). Dans cette approche, la disjonctivité est inscrite dans la représentation lexicale, ou dans un diacritique associé à cette représentation. De l'autre, plusieurs auteurs proposent une grammaire ou des contraintes phonologiques spécifiques à chaque mot concerné, réduisant la disjonction à une constellation de faits indépendants (Tranel & Del Gobbo 2002). Plus récemment, Zuraw & Hayes (2017) analysent la disjonction comme une propension lexicale graduelle (*lexical propensity*) : certains mots seraient plus enclins que d'autres à se comporter comme disjonctifs le long d'un continuum entre les deux extrêmes. Cornulier (1981) exprimait déjà une opinion comparable, en classant les mots disjonctifs en catégoriques (« contrainte de séparation syllabique », classe de *héros*) vs. variables (« contrainte de séparabilité syllabique », classe de *hasard*) ; le deuxième type incluerait « la grande majorité des mots dits à « h aspirée », voire la totalité pour de nombreux locuteurs ».

La difficulté, pour éclairer cette question, est la rareté des études de corpus sur la disjonctivité. La majorité des travaux sont basés sur les intuitions linguistiques des auteurs, les annotations des grammaires et des anecdotes collectées dans leur expérience de locuteur natif. Cette approche ne permet pas de quantifier la disjonction, et de juger si on a bien affaire à un continuum. Une autre veine de recherche construit des données par des expériences d'élicitation en laboratoire. Or, la disjonctivité fait l'objet de prescriptions des grammairiens ; Scheer (à paraître) en particulier souligne que la situation d'enregistrement en laboratoire suscite des réactions normatives et des hypercorrections. D'un autre côté, l'étude sur corpus de la disjonctivité se heurte à un problème de taille : le phénomène est très rare dans la parole. Green & Hintze (2004) ne relèvent en tout que 42 occurrences de mots disjonctifs dans le « corpus de Lille », et calculent que l'intervalle moyen entre deux mots disjonctifs est de 11 minutes. Göhring (2017), qui travaille sur le corpus Phonologie du Français Contemporain (PFC, Durand et al. 2002), ne recense que 313 occurrences, relevant de 43 lemmes. Parmi eux, 3 lemmes seulement apparaissent plus de 20 fois (*haut/hauteur, hasard* et *haie*). A la lumière de la théorie des exemplaires, elle suggère que la stabilité de ces lemmes fréquents vis-à-vis de la disjonction montre un « effet conservateur » : les items lexicaux fréquents sont aussi plus « forts » et faciles d'accès dans le lexique. Son corpus ne lui permet toutefois pas de démontrer la contrepartie : les items les plus rares devraient être moins stables.

L'alternative est de passer aux corpus écrits. Zuraw & Hayes (2017), en travaillant avec le corpus Google n-grams et les publications postérieures à 1900, assemblent une liste de 358 mots apparaissant après des mots 1 dont la forme orthographique change en fonction de l'initiale du mot suivant (ex. *le/l'handicap*), pour un total de 98 millions de bigrammes. L'échelle inégalée de cette étude permet de quantifier des

phénomènes décrits de manière impressionniste dans la littérature. Zuraw & Hayes proposent ainsi que les différents mots 1 ne favorisent pas la (dis)jonctivité au même degré, et les regroupent en trois classes : *beau/bel, vieux/vieil, mon/ma* sont ceux qui favorisent le moins la disjonction, suivis par *au/à l', de/d', la/l',* tandis que *du/de l', le/l'* la favorisent le plus. Ils analysent également le caractère disjonctif du mot 2 comme une propriété lexicale graduelle (*lexical propensity*) : les mots du français s'échelonneraient selon un continuum du 100% jonctif au 100% disjonctif, qu'ils répartissent en 5 classes pour les besoins de la modélisation statistique. Toutefois, ce passage à l'échelle comporte ses inconvénients. La restriction de l'étude aux couples mot 1 + mot 2 ignore ainsi complètement le contexte syntaxique. Nous n'avons pas d'information non plus sur le contexte pragmatique, comme l'emploi autonymique des mots ou leur éventuel « besoin de démarcation ». Une partie des contextes déterminants pour la disjonctivité, dont notamment la liaison, ne peuvent être inclus parce qu'ils n'ont pas de formes différentes à l'écrit. De plus, il reste à savoir si ces résultats sur l'écrit s'étendent également à l'oral, dans lequel la prosodie joue un rôle important.

L'objectif de la présente étude est de compléter et augmenter ces études sur corpus en étudiant la disjonction dans un grand corpus de parole naturelle (au sens où il n'a pas été enregistré pour les besoins de l'étude linguistique) et/ou spontanée (cf. la description des corpus dans la section 2). Nous espérons ainsi permettre une étude quantitative du phénomène dans un registre oral peu ou pas surveillé. Notre démarche se distingue des travaux de Göhring (2017) et Zuraw & Hayes (2017) sur un point important. Ces deux études sont fondées sur une liste d'items lexicaux pré-établie, constituée de mots recensés dans la littérature comme disjonctifs ou variables, et largement limitée aux mots à <h> graphique initial. Pour étudier la disjonctivité « à l'état sauvage », notre démarche consiste au contraire à lancer un large filet, pour récupérer toutes les occurrences disjonctives de nos corpus sans liste pré-conçue. Cette méthode nous permet d'inclure des types de mots généralement laissés de côté dans les études sur la disjonctivité, comme les sigles.

Pour permettre ce changement d'échelle, nous nous appuyons sur une combinaison de tâches automatiques et de travail manuel : le corpus est segmenté par alignement forcé, assorti d'une vérification manuelle d'une partie des items recueillis. Cette méthodologie s'inscrit dans le courant du *big data* qui a touché les Humanités ces dernières années, en exploitant les outils et méthodes des technologies de la parole pour l'étude des phénomènes linguistiques.

2 Méthodologie

Trois corpus ont été rassemblés pour cette étude : ESTER (Galliano et al. 2005), ETAPE (Gravier et al. 2012) et NCCFr (Torreira et al. 2010). Le corpus ESTER contient 90h de journaux d'information radio enregistrés entre 1998 et 2003. Le corpus ETAPE contient 13,5 heures d'émissions radio et 29 heures d'émissions TV (débats et émissions de divertissement), dont 77% de parole (donc 38h), enregistrés en 2011-2012. Le corpus NCCFr, enfin, contient 36h de discours informel entre étudiants qui se connaissent, enregistré en 2007-2008. L'ensemble contient donc environ 160 heures de parole naturelle de registres variés.

Ce corpus a ensuite été segmenté automatiquement avec le système de reconnaissance vocale automatique du laboratoire (Gauvain et al. 2002), en utilisant la méthode d'alignement automatique avec variantes (Adda-Decker & Lamel 2000). Pour un mot comme *un*, par exemple, le système peut choisir de segmenter comme [ɛ̃], [ɛ̃n] avec liaison, ou encore [n] seul ; un mot comme *cette* peut être aligné comme [sɛt], [sɛtə] avec schwa, ou encore [st] ou [stə].

Nous avons ensuite récupéré des séquences de deux mots dans le corpus, dont le mot 1 appartient à la liste en (1). Les mots 1 ont été choisis de manière à sélectionner des séquences de forte cohésion syntaxique (déterminant + nom ou adjectif, pronom + verbe) dans lesquels la liaison est fortement attendue (ex. *un*, *les, on*), où le schwa peut être prononcé ou non (ex. *le, une*), ou bien où une forme différente est attendue devant consonne et voyelle (ex. *du, au, beau*).

(1) Mots 1 (29) : *un, une, des, le, la, l', les, mon, ton, son, mes, tes, ses, ce, cet, cette, ces, on, nous, vous, ils, elles, de, d', du, au, aux, beau, bel*

La base de données ainsi constituée inclut environ 220 500 bigrammes. Nous avons ensuite annoté cette base en parties du discours et en lemmes à l'aide du package *udpipe* (Straka & Strakova 2017) du logiciel R (R Core Team 2021). Pour cette étude, nous avons exclu les mots 2 commençant par une consonne et une semi-voyelle phonétique, réduisant le corpus à un peu plus de 78 000 bigrammes. Cette base de données a ensuite été nettoyée à la main, de manière à enlever les contextes non pertinents pour l'étude (par exemple, les cas de pronoms toniques *nous, on…* dans lesquels aucune liaison n'est attendue), et à corriger la lemmatisation et la partie du discours si nécessaire. Cela aboutit à une base de 75 644 bigrammes. L'absence de liaison entre mot 1 et mot 2, la présence de schwa (ex. *l*[ə]*, cett*[ə]), ou d'une forme pré-consonantique (ex. *du, au*) ont été codées comme disjonctives. Enfin, environ 1800 bigrammes ont été vérifiés auditivement et visuellement sur Praat (Boersma & Weenink 2023) par le premier auteur. En particulier, toutes les occurrences annotées comme disjonctives dans la base de données finale ont été vérifiées. Dans les cas de disjonction, la présence d'un coup de glotte a été annotée à l'aide du logiciel Praat. Ces coups de glotte seront ponctuellement mentionnés dans les résultats, mais ne font pas l'objet ici d'une étude systématique.

3 Résultats

Sur les 75 644 bigrammes de la base de données, 1142 présentent une disjonction, soit 1,51%. Le mot 2 inclut 3567 lemmes différents, dont 210 présentent au moins une occurrence disjonctive, soit 5,86%. L'écoute de ces occurrences a révélé 8 cas de [h] initial, dans les mots *Habous, Hamas* et *Hassania* (tous les cas de *Hassania*). Ces occurrences ont été exclues de l'étude, et toutes les données présentées ci-dessus partent de la base de 75 636 bigrammes dont 1134 disjonctifs. Parmi les 209 lemmes présentant au moins une occurrence disjonctive, 157 le sont dans 100% des occurrences. Toutefois, comme nous allons le voir ci-après, il s'agit en majorité d'hapax, ou de lemmes très peu fréquents. Dans la présente étude, nous nous limiterons donc à une description essentiellement qualitative des données (voir section 3.8 pour une approche statistique inférentielle et ses limites). Pour chaque lemme cité, nous précisons entre parenthèses le nombre d'occurrences disjonctives et le nombre d'occurrences totales : « *handicap* (7/10) » indique donc que le lemme *handicap* apparaît 10 fois dans la base, et qu'il est 7 fois disjonctif. La barre dans *des* | *heurts* indique l'absence de liaison.

3.1 Noms communs, adjectifs et verbes

35 noms communs et verbes apparaissent uniquement en contexte disjonctif. La grande majorité commence par un <h> graphique, et figure dans les listes canoniques des grammaires du français (ex. Grevisse & Goosse 2008 : 55).

(2)

hachoir (1/1)	*handball* (1/1)	*haut* (96/96)	*hisser* (2/2)	*hotte* (1/1)
haine (10/10)	*hangar* (1/1)	*haut-commissaire* (1/1)	*hit-parade* (1/1)	*houlette* (1/1)
hall (1/1)	*hantise* (2/2)	*haut-parleur* (1/1)	*hockey* (1/1)	*houligan* (1/1)
halle (2/2)	*harpe* (1/1)	*hauteur* (26/26)	*hold-up* (1/1)	*hourra* (1/1)
halte (2/2)	*hasard* (10/10)	*héros* (12/12)	*honte* (6/6)	*hurlement* (1/1)
hameau (3/3)	*has-been* (1/1)	*heurt* (6/6)	*hors-la-loi* (1/1)	
hand (1/1)	*haschich* (1/1)	*hip-hop* (3/3)	*hotline* (2/2)	

S'y ajoute un mot fréquent sans <h> graphique, la *une* des journaux (81/81), qui se distingue de *une* pronom (comme dans *l'une et l'autre*). On trouve également le mot d'emprunt *oujdi* (1/1).
6 autres lemmes présentent de la variation. Le nom du *handicap* (7/10) apparaît trois fois jonctif (*l'handicap*), chez le même locuteur. Les autres occurrences proviennent de 3 locuteurs différents. Le dérivé *handicapé* (4/5) fait la liaison une fois. Le nom de la *hausse* (93/103) se démarque de *haut* et *hauteur* par 10 occurrences jonctives, toutes situées après *une*. La capacité des déterminants à deux voyelles (*une, cette*) à être prononcés sans schwa devant les mots par ailleurs disjonctifs – ce que Cornulier (1981) appelle un « droit d'e », par opposition à l'élision dans *j(e), c(e)*, etc. – explique sans doute aussi l'occurrence de *un(e) holding* (*holding* 6/7) sans schwa, ainsi que de *un(e) SNCF* (*SNCF* 58/59). Le mot *hamburger* (2/3)

apparaît jonctif dans *vous parliez d'hamburgers* (vs. *découverte du hamburger* et *le hamburger américain*), et le verbe *hurler* (1/2) dans *obligé d'hurler* (vs. *ils* [ʔ]*hurlent,* avec emphase).
Ces lemmes s'opposent aux très nombreux noms communs et verbes qui sont toujours jonctifs (*heure* 0/969, *hiver* 0/60, *habiter* 0/6, etc.). Notons que les verbes disjonctifs sont très rares dans notre corpus (*hisser* et *hurler*) ; leur rareté pourrait contribuer à expliquer les flottements dans les jugements de grammaticalité observés par Scheer (à paraître) dans les séquences pronom sujet + verbe.

3.2 Lettres et sigles

Dans notre base de données, peu de noms de lettres apparaissent. *A* (1/1) est disjonctif dans la formule *de A à D* ; l'expression *les points sur les I* (3/4) fait la liaison une fois. Le terme *Xième* (1/1) est disjonctif dans *la Xième crise*, mais *énième* (1/3) est disjonctif dans une occurrence sur 3, sous l'emphase (*cett*[ə/ʔ] *énième procédure* avec courte pause entre les deux mots).

Les sigles et abréviations se divisent en deux catégories : ils sont jonctifs lorsqu'ils commencent par une voyelle (3a), généralement disjonctifs lorsqu'ils commencent par une consonne (3b).

(3) a. *ADN* (0/12), *A86* (0/6), *EDF* (0/10), *ILPGA* (0/10), *ONG* (0/30), *UEFA* (0/19)…

b. *FN* (43/43), *FMI* (32/32), *RMI* (5/5) (et *RMIste* 2/2), *RPR* (76/76), *SRPJ* (4/4), *SDF* (10/10)…

On trouve 3 exceptions pour les sigles à initiale vocalique : *cett*[ə/ʔ] *EPO chinoise* est prononcé en détachant les mots, avec une courte pause entre chacun (*EPO* 1/9) ; dans *le UPK* (1/1) ,chaque lettre est martelée pour introduire un sigle nouveau dans le discours ; enfin, *une SNCF* (SNCF 58/59) a été commenté dans la section précédente.

Les sigles à initiale consonantique ont un comportement plus complexe. Parmi les sigles commençant par <H>, *HIR* (1/1) et *HCR* (12/12) sont disjonctifs, mais *HLM* (0/2) ne l'est pas. Le comportement du sigle semble s'aligner sur le mot qui sous-tend le <H> : *haut* pour HCR, *habitation* pour HLM. Le sigle *HEC* (1/2) est jonctif lorsqu'il désigne l'école (*l'HEC*), mais disjonctif, au masculin, lorsqu'il désigne les diplômés en emploi massique : *ils importent genre du HEC*. Une asymétrie intéressante apparaît par ailleurs pour les sigles à <R> initial : lorsqu'il s'agit d'un nom de radio, le sigle est systématiquement jonctif (*RFI* 0/8, *RTL* 0/2, *RMC* 0/1). Un contraste intéressant apparaît pour le sigle *RTM* (1/64), qui est jonctif chaque fois qu'il désigne *Radio Télévision Maroc*, mais disjonctif dans l'occurrence où il désigne la *Régie des Transports Métropolitains*. On note que sur ces deux noms propres, seul le deuxième prend l'article. C'est peut-être la même logique qui conduit à l'emploi jonctif du nom de la chaîne de télévision *M6* (0/1). Enfin, la marque *M&M'S* (0/1) apparaît en emploi jonctif dans *paquet d'M&M'S*.

3.3 Numéraux

Le chiffre *onze* (47/47) et le dérivé *onzième* (10/10) sont toujours disjonctifs. Dans le cas de *onze*, son comportement est donc stable qu'il serve de numéral (*ces onze mois de scandale*) ou de nom (*le onze national*). *Un* (8/2016) apparaît disjonctif dans les expressions numériques complexes suivantes.

(4) *de un mètre quatre-vingt*
plus de un dollar zéro neuf
chute de un virgule zéro quatre pourcent
les un virgule huit milliard
plus de la moitié des un milliard huit cents millions de francs
de l'ordre de un cas environ pour trente mille patients traités
la règle du un sur deux
la règle du un pour deux
l' hypothèse des du un tiers (dysfluence)

Une étude spécifique sur *un* serait nécessaire pour comprendre ce qui motive la disjonction dans ces emplois, par opposition à *près d'un accident sur deux, plus d'un milliard de francs* ou *près d'un kilomètre*.

Enfin, nous avons mentionné *Xième* (0/1) et *énième* (2/3) dans la section 3.2, qui sont dérivés de lettres mais utilisés comme adjectif numéral.

3.4 Noms propres

Dans leur grande majorité, les noms propres à <h> initial sont soit catégoriquement disjonctifs (5a), soit catégoriquement jonctifs (5b).

(5) a. *Hamas* (61/61), *Hainan* (1/1), *Harlem* (3/3), *Hezbollah* (32/32), *Highlanders* (2/2), *Hollande* (3/3), *Honduras* (1/1), *Hong-Kong* (2/2), *Hongrie* (3/3), *Hutu* (7/7), etc.

b. *Haïti* (0/4), *Harry Potter* (0/2), *Hammurabi* (0/2), *Helsinki* (0/3), *Himalaya* (0/5), etc.

Le fait de transformer un nom commun en nom propre peut déclencher la disjonctivité, même sans <h> graphique initial : *aux* | *Actes Sud Junior* (1/1, vs. *acte* 0/137), *de Action Logement* (1/1, vs. *action* 0/279). On note également le nom de marque dans *du Omo* (1/1). Les noms de personne sont généralement jonctifs (y compris avec <h> initial), avec deux exceptions dans notre base de données : *cette Anne Coesens que je trouve (...) exemplaire*, où la disjonction semble liée à un effet d'emphase (*Anne* 1/3), et *un André Dussollier qui est magistral* (*André* 1/12).

Le caractère étranger du nom propre semble également jouer un rôle dans les cas de *Amal* (2/2), *All Black* (2/2), *Emmy* (1/1), *Oumma* (4/4) *Onassis* (1/1), *Once* (1/1), *Independent* (1/1). Mais d'autres noms, comme *Independenza* (0/3), *Independence Day* (0/2), ou encore *Act Up* (0/4) et *Athletic* (0/3), sont jonctifs.

4 noms propres présentent de la variation. *Hollande* (1/3) et *Hollandais* (3/3) sont disjonctifs lorsqu'il s'agit des Pays-Bas, mais *Hollande* est jonctif dans les deux cas où il désigne François Hollande. *Hollywood* (3/5) seul est jonctif, mais il est disjonctif dans les titres *Hollywood Actor Award* et *Hollywood Spotlight Award*. De la même manière, *Astérix* (1/2) est jonctif dans *trente-trois tomes d'Astérix*, mais disjonctif dans *en mixage de Astérix et Obélix mission Cléopâtre*. Le Grevisse signale cette possibilité de disjonction devant les titres (Grevisse & Goosse 2008 : 65) ; il pourrait s'agir ici d'un facteur sémantique (référent plus spécifique et moins connu) ou prosodique (taille du constituant). L'effet de la complexité du constituant suivant sur la liaison est depuis longtemps signalée dans la littérature (par exemple Morin & Kaye 1982).

Enfin, *Haïtien* (1/4) fait la liaison une seule fois, tandis que *Haïti* (0/4) est toujours jonctif.

On peut souligner ici qu'on n'observe pas dans la base de données de tendance graduelle à traiter les dérivés comme jonctifs par opposition aux bases. *Héros* (12/12) est toujours disjonctif, et *héroïne* (0/20) toujours jonctif, que le mot désigne le féminin de *héros* ou la drogue. *Onze* (47/47) et *onzième* (10/10) sont systématiquement disjonctifs, de même que *Hongrie* (3/3) et *Hongrois* (3/3). *Handicap* (7/10) est plus variable que *handicapé* (2/3), de même que *Hollande* (1/3) par rapport à *Hollandais* (3/3).

3.5 Disjonctions marginales

Enfin, les cas de disjonction qui restent concernent pour la majorité des mots très fréquents, qui présentent marginalement un emploi disjonctif. 26 lemmes apparaissent ainsi disjonctifs dans moins de 5% de leurs occurrences (*un* et *RTM* ne sont pas inclus ici, ayant été discutés dans les sections précédentes). La majeure partie de ces mots, comme *étude* (1/141), *année* (2/255), *élection* (1/417), *information* (1/356), *opportunité* (1/46), etc. sont prononcés de manière disjonctive sous l'effet de facteurs pragmatiques comme l'emphase ou l'hésitation, lorsque le locuteur semble chercher le mot juste. Notons que sur 31 occurrences, 16 apparaissent après *cette*. La cohésion syntaxique n'implique pas nécessairement une cohésion prosodique ; lorsque cette dernière est rompue, la suite mot 1 – mot 2 porte les marques phonologiques de la disjonction. Il faut souligner que certains des items commentés dans les sections précédentes, qui semblent lexicalement disjonctifs, présentent également un contour intonatif et un rythme « non-neutre » ; nous avons ponctuellement signalé certains cas d'emphase ou d'insistance.

Certaines disjonctions marginales ne rentrent pas dans ce cadre. Les auxiliaires *être* et *avoir* apparaissent 4 fois en contexte disjonctif. *Ils* | *ont interdit de fumer* correspond peut-être à un emploi dialectal attesté. *Elles*

| *étaient entretenues* est dit très rapidement (ces deux derniers exemples viennent de NCCFr). Le cas de *ils et elles* | *étaient profondément républicains* peut être attribué à la taille du sujet, qui constitue un groupe prosodique autonome et non plus un clitique. *Elles*| *aient* | *été réduites* est prononcé de manière hachée, avec un coup de glotte marqué devant *aient* et *été*.

Enfin, quelques occurrences semblent correspondre à des lapsus purs et simples, comme *la issue* (*issue* 1/179), *un* | *échec* (*échec* 1/76), *elles* | *allaient être appliquées* ou *des* | *humeurs* (*humeur* 1/7).

3.6 Lemmes fréquents

Comme le montrent les chiffres des sections précédentes, la plupart des lemmes avec au moins une occurrence disjonctive sont des hapax ou des mots rares dans la base de données. Si l'on se concentre sur les lemmes apparaissant au moins 20 fois dans la base de données, la figure (1) montre que la distribution est clairement bimodale : 12 des 39 lemmes restants (parmi les lemmes qui présentent au moins une occurrence disjonctive) présentent un taux de disjonction supérieur à 90%, tandis que les 27 autres sont en-dessous de 5%. Les 12 lemmes en question sont *FMI, FN, Hamas, haut, hauteur, Hezbollah, onze, RPR, une* (nom commun ; ces mots sont 100% disjonctifs)*, SNCF* (98,3%) et *hausse* (90,3%) – une liste qui reflète le lexique du discours journalistique.

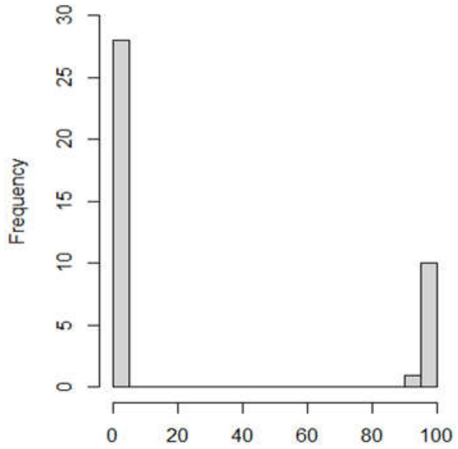

Figure 1. Taux de disjonction pour les lemmes ayant au moins une occurrence disjonctive et apparaissant au moins 20 fois dans la base.

3.7 Bilan intermédiaire

Les résultats présentés dans les sections précédentes suggèrent qu'il existe au moins deux sources de disjonctivité distinctes en français : la disjonctivité lexicale, sous la forme de lemmes disjonctifs dans la grande majorité de leurs occurrences, et la disjonctivité venant de facteurs pragmatiques, sémantiques et syntaxiques, sous la forme de lemmes n'apparaissant en emploi disjonctif que de manière marginale. Notre base de données dans son état actuel ne distingue pas ces deux types. Nous émettons l'hypothèse que le deuxième se distingue par la présence d'une rupture prosodique entre le mot 1 et le mot 2. Pour tester cette hypothèse, une étude complète de la disjonctivité en français devrait donc inclure une étude prosodique, de manière à départager ces deux ensembles et leur intersection.

Un autre point important à souligner est l'importance du contexte large. Une étude se limitant à l'étude du mot 1 et du mot 2 ne permet pas de comprendre, par exemple, que *Hollywood* et *Astérix* sont disjonctifs seulement lorsqu'ils appartiennent à un syntagme plus grand. Ces deux points doivent être gardés en tête pour la lecture de la section suivante.

3.8 Effet du mot 1 et du type de sandhi

A des fins de comparaison avec la littérature, nous nous sommes penchés sur l'effet du type de sandhi (liaison, schwa, allomorphie) et du mot 1 sur le caractère disjonctif ou non d'un bigramme.

Göhring (2017) montre que le taux de disjonction d'un lemme varie en fonction du type de sandhi dans lequel il est impliqué. Elle distingue quatre types de sandhi : l'élision, comme dans *le/la/l'*, la liaison, la prononciation de schwa dans les clitiques dissyllabiques comme *une*[1]*, et* l'enchaînement. Dans son corpus, les clitiques monosyllabiques et la liaison montrent de hauts taux de disjonction (respectivement 100% et 89,34%), alors que l'enchaînement et les clitiques « plurisyllabiques » de l'autre (comme *une*), montrent des taux plus bas (62,5% pour le schwa, 8,3% pour l'enchaînement). Pour tester cette hypothèse dans notre base de données, nous avons établi un modèle linéaire généralisé (régression binomiale simple). La variable dépendante est le type de sandhi, codé de manière à permettre la comparaison avec les trois premières catégories de Göhring (2017) : élision, liaison, schwa des clitiques non monosyllabiques (l'enchaînement n'est pas inclus parce qu'il n'est pas codé dans notre base données). Notons que son corpus est limité à des mots sélectionnés pour être disjonctifs, alors que le nôtre inclut la totalité de la base. Les résultats montrent que la catégorie des clitiques dissyllabiques (*une, cette*) est associée au plus fort taux de disjonction (2%), significativement plus élevé que celui de la liaison (1,1% ; $z = 5.503$; $p < .001$). La liaison est à son tour associée à un taux de disjonction plus fort que les clitiques monosyllabiques (0,2% ; $z = 6.931$; $p < .001$). Cette échelle dissyllabes > liaison > monosyllabes est l'inverse de celle trouvée par Göhring (2017). Elle surprend au regard des résultats présentés en §3.1 (les 10 occurrences jonctives de *hauteur* apparaissent après *une*), mais est cohérente avec l'association fréquente de *cette* à avec les disjonctions « marginales » mentionnée en §3.5. Il est donc possible que ce résultat soit lié à ce que nous avons analysé en §3.7 comme la disjonctivité non lexicale, sous l'effet de facteurs prosodiques et pragmatiques non contrôlés dans la base de données.

L'effet du mot 1 sur la disjonction est souvent commenté dans la littérature, en suggèrant que la disjonctivité est une propriété de la construction plutôt que du mot 2 seul. Zuraw & Hayes (2017) sont les seuls à notre connaissance à montrer ce point par une étude quantitative. Dans leurs résultats, *du/de l'* et *le/l'* par exemple sont plus souvent associés à des contextes disjonctifs que *beau/bel, vieux/vieil* et *ma/mon ; le/l'* est aussi plus favorable à la disjonction que *de/d'*. La structure des données ne permettant pas une analyse conjointe de l'effet de l'ensemble des facteurs dans un même modèle, nous avons testé ce point en établissant un second modèle généralisé de régression binomiale de réalisation de la disjonction. La variable indépendante est le mot 1, dont la liste de départ figure en (1) ci-dessus. Certains de ces « mots 1 » sont des variantes du même lexème : *l'* par exemple peut renvoyer à *le* ou à *la* ; l'article est inclus dans *du* et *au.* Nous avons donc recodé certains mots 1 en classes de mots 1 : la classe « *le* » inclut *le-la-l'-du* et *au*, la classe « *de* » inclut *de-d'*, et la classe « *ce* » inclut *ce-cet*. Les 23 catégories ainsi constituées ont ensuite été réduites à celles présentant au moins 100 occurrences, éliminant *bel, ce, me, tes* et *ton*. Les résultats suggèrent que les classes *le*, *un, des, ces* et *cette* figurent parmi les mots 1 associés au plus grand taux de disjonctivité (contrairement à ce qu'on pourrait attendre pour *cette*), suivis par *les* et *une*, puis les classes *de, les, aux, ce, son, ses, on* et *ils* (*nous, vous* et *mon* n'apparaissent dans aucun contexte disjonctif). La classe « *de* » en particulier montre un taux de jonction significativement plus élevé que la classe « *le* » (99,9% contre 97,8% ; $z = 13,08$, $p < .001$), en accord avec les résultats de Zuraw & Hayes (2017) : il semble que l'article défini se prête plus facilement à la disjonction que l'article indéfini. La situation de *une* en-deçà d'autres mots 1 confirme par ailleurs les résultats commentés dans la section 3.1 (cas de *une hausse* en particulier).

Toutefois, ces résultats doivent être pris avec précaution. Les 4 occurrences de *elles* disjonctif, par exemple, incluent le cas de *ils et elles | étaient*, où la disjonction n'est probablement pas dûe aux propriétés du mot *elles*, et des cas atypiques devant les verbes *avoir, être* et *aller,* commentés dans la section 3.5. La position de *cette* dans le peloton de tête, alors que nous avons vu qu'il apparaît particulièrement fréquemment dans les disjonctions d'origine pragmatique (section 3.5), suggère qu'une telle analyse quantitative doit prendre

[1] La prononciation de schwa dans les clitiques d'une seule syllabe (type *le, de, ce*) est généralement distinguée de celle des clitiques plus longs comme relevant de l'élision.

en compte plus finement les différents facteurs impliqués dans la disjonction. Au vu des résultats des sections précédentes, il est vraisemblable que l'effet du mot 1 (et peut-être du type de sandhi) diffère *a minima* dans ces deux grandes catégories que nous avons identifiées, disjonction lexicale et non-lexicale.

4 Discussion

Dans leur rapport « Mining Years and Years of Speech » (2019), Coleman et al. estiment que pour obtenir un échantillon « raisonnable » de paires de deux mots fréquents, plus de 1000 heures de corpus sont nécessaires ; pour des paires de mots arbitraires, il faut plus de 100 ans. Ces ordres de grandeur révèlent une difficulté majeure pour étudier la disjonction en corpus : le phénomène est très rare dans le discours. Bien que les occurrences rassemblées ici dépassent celles de Green & Hintze (2004 ; 42 occurrences) et de Göhring (2017 : 313 occurrences – ces deux études étant limitées aux mots à <h> initial), les 160 heures de parole naturelle réunies pour cette étude et leurs 1134 occurrences disjonctives restent insuffisantes pour tester et quantifier de manière exhaustive la grande variété de facteurs qui interviennent dans la disjonction.

Avec ses limites, l'approche que nous avons adoptée dans cette étude suggère néanmoins une réponse à notre question initiale : dans l'ensemble, la disjonctivité se comporte comme une propriété catégorique des items lexicaux. En tout, 75% des lemmes qui présentent au moins une occurrence disjonctive le sont dans toutes leurs occurrences ; lorsqu'on observe uniquement les lemmes attestés plus de 20 fois, une distribution bimodale émerge clairement. Cela suggère que certains lemmes sont lexicalement marqués comme lexicaux, et d'autres non. L'apparition de « paires minimales » constitue un argument dans ce sens : le comportement de *une* nom (*la une des journaux*) est ainsi catégoriquement différent de celui de *une* pronom ou article. De manière plus anecdotique, nous avons également relevé un comportement de *Hollande* différent selon le référent (le pays ou l'homme politique), de même que pour *RTM* (*Radio Télévision Maroc* ou *Régie des Transports Métropolitains*).

Il reste bien sûr de la variation. Le facteur principal que nous avons identifié concerne le mot 1 *une*, prononcé parfois sans schwa devant *hausse, holding* et *SNCF.* Nous avons également pu repérer, dans le cas de *handicapé*, une variation inter-locuteurs. Le comportement jonctif singulier des sigles de noms de radio provient peut-être d'un effet de fréquence ou de familiarité avec le sigle (une grande partie des occurrences de *RTM* apparaît par exemple dans la formule figée *à l'écoute d'RTM*) ; mais ce comportement est régulier, et ne contredit pas l'hypothèse d'une catégorie de mots disjonctifs opposée à celle d'une catégorie de mots non-disjonctifs. Les autres cas de variation observés – par exemple la liaison devant *Haïtiens* ou l'absence de schwa dans *d'hamburgers* – forment un résidu que l'insuffisance des données ne permet pas d'expliquer. Ce résidu est toutefois réduit par rapport au nombre de lemmes dont le comportement est catégorique.

Le deuxième résultat suggéré par cette étude est qu'il existe deux sources distinctes de disjonctivité en français. A l'ensemble des lemmes lexicalement marqués comme disjonctifs s'oppose un ensemble de lemmes lexicalement jonctifs, qui peuvent apparaître occasionnellement en emploi disjonctif. Cette analyse s'oppose à celle de Zuraw & Hayes (2017) qui voient la disjonctivité comme une propriété graduelle le long d'un continuum. Les facteurs qui déclenchent la disjonctivité sont d'ordre pragmatique, sémantique ou syntaxique : les principaux sont l'emphase, l'hésitation, la taille du constituant auquel appartient le mot 1 ou le mot 2. Nous faisons l'hypothèse que ces facteurs introduisent une rupture prosodique entre le mot 1 et le mot 2. Cette rupture serait marquée par les manifestations phonologiques de la disjonction (rupture du sandhi externe), mais aussi dans le contour intonatif et le rythme. Les mots lexicalement disjonctifs, par contraste, peuvent apparaître dans des contextes de forte cohésion prosodique (ils peuvent bien sûr aussi faire ponctuellement l'objet d'une démarcation pragmatique). Notons que cette hypothèse s'oppose *a priori* à la proposition de Scheer (à paraître), pour qui les mots disjonctifs ont la capacité de déclencher un nouveau domaine de production dans la planification de la parole (Wagner 2012). L'idée est que lorsque le mot 2 n'est pas déjà planifié au moment où le mot 1 est computé, le mot 2 se retrouve à l'initiale de domaine ; une rupture prosodique entre mot 1 et mot 2 est alors plausible. Notre prédiction serait que les mots lexicalement disjonctifs ne sont pas nécessairement associés à une rupture prosodique, alors que les emplois marginalement disjonctifs de mots par ailleurs jonctifs le sont.

La prochaine étape est donc d'annoter le corpus selon le profil prosodique des bigrammes, de manière à distinguer les deux sources de disjonctivité. Cette étape est nécessaire, à notre avis, avant de pouvoir procéder à une étude plus avancée des différents facteurs, comme le mot 1 ou le type de sandhi. Elle permettrait également de tester dans la parole spontanée la prédiction de Pagliano (2003), selon qui la réalisation des séquences du type *une housse* avec schwa et coup de glotte est limitée aux contextes d'emphase (cf. Gabriel & Meisenburg 2009 pour un test de cette prédiction en laboratoire). Il serait également intéressant, dans un second temps, d'annoter le corpus pour les différents facteurs identifiés comme favorables à la disjonction, afin d'en comparer l'impact dans un seul modèle.

Ces résultats doivent toutefois être nuancés par les nombreuses limites de notre étude. Comme mentionné plus haut, le corpus n'atteint pas des dimensions suffisantes pour quantifier précisément le degré de disjonctivité d'un grand nombre de lemmes. Par ailleurs, la majorité des locuteurs de notre corpus sont des locuteurs professionnels, dont l'entraînement à la prise de parole en public a pu contribuer à stabiliser les emplois jonctifs ou disjonctifs des différents mots. Notre méthodologie nous a permis de nous pencher sur la réalisation de la liaison, du schwa et des mots 1 dont la forme change en fonction de l'initiale du mot 2. Nous n'avons donc pas de données sur l'enchaînement, dont la détection automatique semble difficile, alors que l'étude de Göhring (2017) montre que son comportement vis-à-vis de la disjonction est beaucoup plus variable que pour les autres types de sandhi. Etant donnée la tension entre la grande quantité de données nécessaires et l'analyse fine des différents facteurs portant sur la disjonctivité, chaque méthodologie éclaire seulement une partie du problème. La présente étude vient donc compléter les travaux précédents en offrant, avec sa méthodologie nouvelle, une lumière partielle mais complémentaire au problème de la disjonction en français.

Références bibliographiques

Adda-Decker, M., & Lamel, L. (2000). The use of lexica in automatic speech recognition. In F. Van Eynde & D. Gibbon (éds.), *Lexicon Development for Speech and Language Processing* (pp. 235–266). Springer.

Boersma, P. (2007). Some listener-oriented accounts of h-aspiré in French. *Lingua*, 117, 1989–2054.

Boersma, P. & Weenink, D. (2023). Praat: doing phonetics by computer [Computer program]. Version 6.4, <http://www.praat.org/>.

Coleman, J., Libermann, M., Kochanski, G., Yuan, J., Grau, S., Cieri, C., Baghai-Ravary, L., & Burnard, L. (2011). Mining years and years of speech. Final report of the Digging into Data project "Mining a Year of Speech." University of Oxford Phonetics Laboratory.

Cornulier, B. de. (1981). H-aspirée et la syllabation. Expressions disjonctives. In D. Goyvaerts (éd.), *Phonology in the 1980's* (pp. 183–230). Story-Scientia.

Côté, M.-H. (2008). Empty elements in schwa, liaison and h-aspiré: The French Holy Trinity revisited. In J. Hartmann (éd.), *Sounds of silence: Empty elements in syntax and phonology* (pp. 61–103). Brill.

Durand, J., Laks, B. & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In C. Pusch & W. Raible (éds.), *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics – Corpora and Spoken Language* (pp. 93-106). Gunter Narr Verlag.

Fouché, P. (1959). *Traité de prononciation française* (2e éd.). Paris : Klincksieck.

Gabriel, C., & Meisenburg, T. (2009). Silent onsets? An optimality-theoretic approach to French h aspiré words. In F. Kügler, C. Féry & R. van de Vijver (éds.), *Variation and gradience in phonetics and phonology* (pp. 163–184). Mouton de Gruyter.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., & Gravier, G. (2005). The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News. *Proceedings of ISCA Interspeech*, 1149–1152.

Gauvain, J.-L., Lamel, L., & Adda, G. (2002). The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2), 89–108.

Göhring, T. (2017). L'état actuel du h disjonctif (h aspiré): Une approche fondée sur la fréquence d'emploi. *Romanische Forschungen*, 129(2), 147–168.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., & Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *Proceedings of LREC - 8th International Conference on Language Resources and Evaluation*, 3995-3999.

Green, J. N., & Hintze, M.-A. (2004). Le h aspiré en français contemporain: stabilité, variation ou déclin ? In A. Coveney, M.-A. Hintze, & C. Sanders (éds.), *Variation et francophonie* (pp. 241–280). L'Harmattan.

Grevisse, M., & Goosse, A. (2008). *Le bon usage. Grammaire française* (14e éd.). De Boeck & Duculot.

Morin, Y. C., & Kaye, J. (1982). The syntactic bases for French liaison. *Journal of Linguistics*, 18(2), 291–330.

Pagliano, C. (2003). *L'épenthèse consonantique en français : ce que la syntaxe, la sémantique et la morphologie peuvent faire à la phonologie : parles-en de ta numérotation ?impossible*. Thèse de l'Université de Nice.

PFC = Phonologie du Français Contemporain, <www.projet-pfc.net>

Plénat, M. (1995). Une approche prosodique de la morphologie du verlan. *Lingua*, 95, 97–129.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. <http://www.R-project.org/>, <http://www.rstudio.com>

Schane, S. (1978). L'emploi des frontières de mots en français. In B. de Cornulier & F. Dell (éds.), *Etudes de phonologie française* (pp. 133–147). Editions du CNRS.

Scheer, T. (à paraître). Glottal stop insertion and production planning domains in French. *The Linguistic Review.*

Straka, M., & Strakova, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 88–99.

Tessier, A.-M., Jesney, K., Vesik, K., Lo, R., & Bouchard, M.-E. (2023). The Productive Status of Laurentian French Liaison: Variation across Words and Grammar. *Proceedings of the 2022 Annual Meeting on Phonology*.

Torreira, F., Adda-Decker, M., & Ernestus, M. (2010). The Nijmegen corpus of casual French. *Speech Communication*, 10(3), 201–212.

Tranel, B., Del Gobbo, F. (2002). Local Conjunction in Italian and French Phonology. In C. R. Wiltshire & J. Camps (éds.), *Romance Phonology and Variation* (pp. 191–218). Benjamins.

von Heusinger, K., & Wespel, J. (2019). Indefinite proper names and quantification over manifestations. In E. Puig-Waldmüller (éd.), *Proceedings of Sinn Und Bedeutung* 11, 332–345.

Wagner, M. (2012). Locality in Phonology and Production Planning. In A. McKillen & J. Loughran (éds.), *Proceedings of the Montreal-Ottawa-Toronto (MOT) Phonology Workshop 2011. Phonology in the 21st Century: In Honour of Glyne Piggott. McGill Working Papers in Linguistics* (Vol. 22).

Zuraw, K., & Hayes, B. (2017). Intersecting constraint families: An argument for Harmonic Grammar. *Language*, 93(3), 497–548.