Multidimensionnalité de la liaison variable et difficultés de classification. Le cas des adverbes monosyllabiques

Giulia Barreca

Multidimensionnalité de la liaison variable et difficultés de classification. Le cas des adverbes monosyllabiques.

Giulia Barreca
Université Paris Ouest Nanterre La Défense, MoDyCo, CNRS, 200, avenue de la République, 92001 Nanterre
giulia.barreca@u-paris10.fr

RESUME

La combinaison des multiples facteurs qui influencent la réalisation de liaison variable constitue une source de complexité qui rend difficile son analyse et son traitement classificatoire. Le critère morphosyntaxique a le plus souvent guidé les classifications de la liaison variable. Dans cette première étude menée sur des séquences d'adverbes monosyllabiques suivis d'adjectifs, nous avons essayé de démontrer que la prise en compte du seul critère d'appartenance aux catégories morphosyntaxiques s'avère insuffisante pour décrire l'hétérogénéité des comportements associés aux éléments appartenant à une même classe. Ensuite, notre but à été celui de décrire de façon plus précise la variabilité de ce phénomène, analysant, à l'aide d'un apprentissage automatique, les nombreuses dimensions linguistiques (phonologique, syntaxique et lexicale) en jeu dans la réalisation de la liaison variable.

ABSTRACT

Multidimensionality of the variable liaison and difficulties in classification. The case of monosyllabic adverbs.

The combination of multiple factors that influence the realization of variable liaison is a source of complexity that makes its analysis and classificatory treatment difficult. The morphosyntactic criterion has often guided the classifications of variable liaison. In this first research of sequences of monosyllabic adverbs followed by adjectives, we have tried to prove that considering only the criterion of belonging to morphosyntactic categories is insufficient to describe the heterogeneity of behaviors associated with elements of the same class. The aim of this research was to describe more accurately the variability of this phenomenon, using machine learning, in order to analyze the largest number of linguistic dimensions (phonological, syntactic and lexical) involved in the realisation of variable.

MOTS-CLES : Liaison variable, hétérogénéité, analyse multidimensionnelle, classifications.
KEYWORDS: Variable liaison, heterogeneity, multidimensional analysis, classifications.

1 Introduction

La liaison a fait l'objet de nombreux travaux qui ont dû faire face à l'extraordinaire idiosyncrasie de ce phénomène et par conséquent à la difficulté de dégager et décrire ses nombreuses facettes. En effet, la description de la liaison variable oblige à prendre en compte de plusieurs dimensions d'analyse (phonologique, syntaxique et lexicale), ce qui la rend inévitablement complexe.

Dans les pages qui suivent, après avoir présenté un bref parcours des études menées sur liaison, nous allons mettre en évidence les limites des classifications de la liaison établies, le plus souvent, sur le seul facteur morphosyntaxique[1]. Pour ce faire, nous avons développé une première étude de la catégorie de liaison variable constituée par les adverbes monosyllabiques suivis d'un adjectif (désormais Adv+Adj) à partir des données extraites des corpus PFC (Phonologie du Français Contemporain) (Durand et al., 2002) et CFPP2000 (Corpus de Français Parlé Parisien des années 2000) (Branca-Rosoff et al., 2012). Nous essayerons de démontrer que la prise en compte du seul critère d'appartenance à la classe morphosyntaxique permet difficilement de rendre compte de la complexité et de l'hétérogénéité des comportements de la liaison associés aux adverbes monosyllabiques.

2 Les études précédentes: plusieurs facettes du même phénomène

Les premières études sur la liaison, contrairement aux études générativistes successives (Schane, 1967), ont le mérite d'avoir développé une problématisation de l'étude de la liaison et d'avoir mis en avant sa nature multidimensionnelle. En effet, Delattre (Delattre, 1966) considère que la syntaxe ne peut pas représenter le seul critère de classification des liaisons. Malgré la grille descriptive des trois catégories de liaison qu'il établit en 1947, il précise que ces catégorisations ne sont pas absolues, et peuvent varier sur la base de plusieurs facteurs, parmi lesquels le style, la prosodie et le degré phonématique de la consonne de liaison (désormais CL) (Delattre, 1966). En revanche, à partir des études générativistes de Schane (Schane, 1967) la liaison ne sera traitée que comme étant un phénomène strictement phonologique et unitaire sans jamais parvenir à expliquer son caractère variationnel. Les études successives se contenteront de décrire de façon partielle l'influence de l'une ou l'autre de ces dimensions, attribuant aux autres un rôle secondaire, finissant par se heurter à l'incapacité d'expliquer la variabilité de la liaison. Parmi celles-ci, par exemple, les études centrées sur la dimension phonologique qui considèrent la liaison comme le résultat d'un évitement du hiatus (Encrevé, 1988) ou de l'émergence des *consonnes latentes* (Tranel, 2000) ne s'appliquent qu'à des contextes de liaison catégorique. Le niveau stylistique est également considéré par de nombreuses études comme le facteur le plus déterminant. Bien que le choix du registre puisse faire varier la fréquence des liaisons, comme le souligne Mallet (Mallet, 2008), cela ne suffit pas à expliquer la variabilité des réalisations pour un même locuteur et la même catégorie morphosyntaxique. Une autre série de travaux se base sur l'analyse des facteurs lexicaux tels que la fréquence du premier et deuxième mot appartenant à la séquence liaisonnée (désormais respectivement M1 et M2) et de leur cooccurrence (Bybee, 2001).

[1] Nous faisons référence aux tripartitions classiques de la liaison en: liaison obligatoire/catégorique, facultative/variable et interdite/erratique. Parmi lesquelles celles qui ont été proposées par Delattre (Delattre, 1947) et Encrevé (Encrevé, 1988).

3 Les limites de la classification de la liaison variable sur base morphosyntaxique. Hétérogénéité des comportements des adverbes monosyllabiques

Comme nous l'avons vu dans les paragraphes précédents, la multidimensionnalité de l'analyse de la liaison, constitue un obstacle à son traitement classificatoire, car on reconnaît que les dimensions sont interdépendantes et on ne peut pas les traiter séparément (Mallet, 2008).

L'appartenance à la classe morphosyntaxique constitue le moyen pour opérer cette classification qui, encore qu'elle soit nécessaire sur le plan métalinguistique, résulte insuffisante pour décrire la réalité des usages linguistiques. Dans ce contexte, à travers cette première étude nous souhaitons décrire de façon plus précise l'hétérogénéité des fonctionnements de la liaison à l'intérieur des séquences Adv+Adj appartenant à la classe des adverbes monosyllabiques suivi d'un M2. Cette classe, comme d'autres, a été dans un premier temps traitée comme contexte de liaison obligatoire (Delattre, 1947). Des études successives (Mallet, 2008) ont intégré ce contexte à la classe des liaisons variables. Toutefois, malgré cette révision, la classification actuelle ne parvient pas à expliquer et donc à décrire la présence d'un grand écart des réalisations de la liaison existant entre les différents adverbes monosyllabiques. En effet, selon l'adverbe qu'on considère, soit la liaison est réalisée très souvent (96,55 % *très*, 64% *plus*) soit presque jamais (1,36 % *pas*)[2]. Une telle variabilité de comportements de la liaison peut être mise en relation, de façon plus générale, aux caractéristiques propres des adverbes. En effet, plusieurs études (Wilmet 2010) ont déjà souligné que les adverbes constituent une classe résiduelle regroupant des éléments invariables qui ne rentrent pas dans d'autres catégories. Par conséquent, il s'agit d'un ensemble très hétérogène qui est difficilement classifiable. Différents critères de classement ont été évoqués par les nombreux travaux classificatoires: critère fonctionnel, sémantique, degré d'intégration de l'adverbe dans la phrase (Melis, 1983) ainsi que la portée de l'adverbe et sa fonction (Blumethal, 1990). Dans les pages suivantes, nous essayerons de mettre en évidence la nécessité d'étendre la description de la variation de la liaison à d'autres champs d'analyse (phonologique, syntaxique et lexicale) que celui d'appartenance à la catégorie morphosyntaxique.

4 Méthodologie

Pour étudier un tel sujet, l'utilisation d'un corpus annoté s'impose. Dans cette première étude, nous allons analyser les séquences Adv+Adj des adverbes *pas*, *très*, *plus*, *moins*, *trop*[3] issues du corpus PFC (Durand et al., 2002) et du corpus CFPP2000 (Branca-Rosoff et al., 2012). En ce qui concerne le corpus PFC, les données prises en compte ont été extraites des transcriptions des entretiens guidés[4] et des entretiens libres[5] (tâches orales) de 134[6] locuteurs appartenant à 34

[2] Données extraites de Mallet (Mallet, 2008).

[3] Nous n'avons pas pris en compte, dans cette première étude, les séquences contenant les adverbes *tout*, *fort*, *mieux*, *bien* à cause de leur faible nombre d'occurrences dans notre échantillon.

[4] Il s'agit d'un entretien semi-dirigé par des questions posées par l'enquêteur d'un registre de langue plutôt soutenu.

[5] L'entretien libre consiste en une interaction plus riche (de registre plus familier) entre plus de deux personnes.

[6] Bien que nous ayons mené la recherche sur l'ensemble des locuteurs du corpus PFC (394 locuteurs pour un total de 36 enquêtes), nous n'avons obtenu qu'un échantillon très limité où seulement 134 locuteurs appartenant à 34 enquêtes différentes sont représentés. Cela est probablement lié au fait que l'annotation morphosyntaxique actuelle du corpus PFC, réalisée à l'aide de Treetagger, présente de nombreuses erreurs d'étiquetage. Actuellement, nous sommes en train de ré-annoter en morphosyntaxe (POS et lemmes) le corpus PFC à l'aide de l'étiqueteur DisMo (Christodoulides et al., 2014) en collaboration avec le laboratoire Valibel de l'Université Catholique de Louvain et l'Université de Genève.

enquêtes différentes. Les données issues du corpus CFPP2000 ont été extraites de l'ensemble des entretiens semi-directifs. L'échantillon des données que nous avons recueilli correspond à 505 occurrences totales de séquences Adv+Adj susceptibles d'être liaisonnées, ainsi réparties parmi les adverbes monosyllabiques suivants: *pas*, *très*, *plus*, *moins*, *trop*.

	pas	très	plus	moins	trop
PFC	81	67	44	8	8
CFPP2000	57	131	82	18	9
TOT	138	198	126	26	17

TABLE 1 – Occurrences totales des séquences Adv+Adj reparties par adverbe

En ce qui concerne les adjectifs nous avons enregistré un ensemble assez hétérogène correspondant à 70 lemmes différents pour le corpus PFC et 118 lemmes pour CFPP2000[7]. Ensuite, dans le but d'assurer une analyse multidimensionnelle de la liaison, pour chaque séquence nous avons recueilli des données relevant de trois principales dimensions d'analyse: phonologique, syntaxique et lexicale. De l'ensemble de ces dimensions résultent les variables dépendantes suivantes[8]:

- Variables phonologiques: la nature de la consonne de liaison (CL), le nombre de syllabes du M2, à savoir de l'adjectif (NSYLL M2);
- Variables syntaxiques: degré de cohésion syntagmatique (DCS) et rapports de dépendance entre l'adverbe et l'adjectif. Notamment si l'adverbe porte sur le verbe et est régi par le verbe (DV) ou s'il porte sur l'adjectif et donc il dépend de l'adjectif (DA);
- Variables lexicales[9]: la fréquence du M1, à savoir de l'adverbe (FREQ M1), la fréquence du M2, à savoir de l'adjectif (FREQ M2) et la fréquence de cooccurrence de la séquence Adv+Adj (COOC)[10].

Pour ce qui est de la variable indépendante, la réalisation de la liaison (REALISATION), afin d'assurer des résultats plus fiables sur un échantillon restreint, nous avons décidé de ne considérer que deux cas possibles: liaison réalisée (désormais LR), liaison non réalisée (désormais LNR).

L'ensemble des données ainsi organisées a été soumis au traitement de deux différents logiciels d'apprentissage automatique: *Sipina* (Zighed, 1992) et *Weka* (Holmes et al., 1994). Il s'agit de deux logiciels qui, dans le cadre de l'apprentissage supervisé, génèrent des arbres de décision. Ceux-ci permettent d'effectuer des classifications de données et d'aboutir au traitement d'un grand nombre de variables explicatives à partir de l'observation d'un échantillonnage. Autrement dit,

[7] Il s'agit plus précisément de 223 formes fléchies totales d'adjectifs différents possédant des indices de fréquence compris entre 0 et 248,45, extraits de la fréquence *freqfilm2* de la base de donnée lexicale *Lexique 3* (New, 2006).

[8] Nous n'avons pas analysé le facteur de variabilité inter/intra-individuelle étant donné le nombre faible d'occurrences par locuteur présent dans notre échantillon.

[9] Ces derniers facteurs de nature lexicale ont été extraits de la fréquence *freqfilm2* de la base de donnée lexicale *Lexique 3* (New, 2006) correspondant à la moyenne des fréquences par million des quatre sous-corpus de sous titres de films. Afin de vérifier la fiabilité des ces indices nous les avons comparé aux occurrences totales (tout contexte mélangé) des mêmes adverbes tirées du corpus CFPP2000. Les rapports de fréquence entre les adverbes correspondent.

[10] Cette valeur est représentée de manière absolue.

grâce à l'emploi de ces outils d'apprentissage automatique, nous espérons pouvoir représenter des tendances qui décrivent la variation de la liaison à partir de notre base de données d'apprentissage de séquences Adv+Adj. Il s'agira concrètement de prédire la valeur (LR ou LNR) de la variable indépendante (REALISATION) à partir de l'ensemble des variables prédictives (CL, NSYLL M2, DCS, FREQ M1, FREQ M2, COOC) afin d'établir une hiérarchie parmi chacune des différentes variables.

Cette méthode nous permettra, tout en gardant les propriétés d'économicité d'un outil descriptif, d'étudier conjointement les multiples facteurs en jeu dans la réalisation de la liaison variable des séquences Adv+Adj. Cela ne signifie pas pour autant que la catégorie morphosyntaxique d'appartenance ne joue pas un rôle important, mais simplement nous supposons qu'elle ne constitue pas un critère descriptif suffisant.

5 Résultats et discussion

Si nous analysons les résultats totaux de notre échantillon ne prenant en compte que le critère d'appartenance à la catégorie Adv+Adj nous observons 66% de LR et 34% de LNR. Cependant, en analysant les taux de réalisations de chaque adverbe, nous pouvons constater un fort déséquilibre interne à la catégorie. La liaison est beaucoup réalisée avec les adverbes *très* (95% LR) et *plus* (94% LR) alors que elle n'est jamais réalisée avec *pas* (0% LR). L'adverbe *pas*, en effet, représente 82% des LNR totales.

	pas	très	plus	moins	trop
LR	0%	95%	94%	65%	59%
LNR	100%	5%	6%	35%	41%

TABLE 2 – Résultats des LR et des LNR des séquences Adv+Adj reparties par adverbe

Nous avons enregistré des comportements différents même dans des séquences Adv+Adj où *pas*[11] et *très*[12] étaient associés au même adjectif.

Ces données révèlent une anomalie existant à l'intérieur de la classification sur base morphosyntaxique. En effet, l'un des inconvénients de cette classification réside dans la tendance à regrouper sous une même classe des adverbes qui peuvent se différencier fortement à la fois selon leur distribution et leur fréquence. Les adverbes monosyllabique *pas*, *plus, moins*, *trop* sont des adverbes qui peuvent porter sur le verbe, sur l'adjectif ou sur un adverbe et possèdent des rapports de dépendance différents (DA ou DV) selon leur portée. De plus il faut souligner que dans des séquences Adv+Adj, *plus* (en tant qu'adverbe comparatif et superlatif) se trouve bien plus souvent que l'adverbe *pas* dans un rapport de dépendance de l'adjectif que du verbe. Dans notre échantillon nous avons pu observer que seulement dans 10% des séquences du type *pas*+Adj, l'adverbe *pas* dépend de l'adjectif. Dans ces cas il se rapproche de la fonction de *non*

[11] «Je sais pas+ je sais pas si euh+ c'est vrai que je trouve que le+ le mode de vie à Pris est assez stressant y a+ les gens sont *pas agréables*» (LNR - CFPP2000_Corpus Killian Belamy et Lucas Hermano).
[12] «et puis après rue du Faubourg Saint–Antoine c'est très vivant c'est vraiment *très agréable*» (LR - CFPP2000_Corpus Pierre Beysson).

en tant que préfixe adjectival[13]. En revanche, nous n'avons enregistré aucune séquence susceptible d'être liaisonnée, contenant *plus* régi par le verbe. Concernant l'adverbe *très*, il ne peut porter que sur l'adjectif ou sur l'adverbe. Par conséquent il sera toujours régi par l'adjectif ou l'adverbe duquel il dépend.

Nous pourrions imaginer que la dimension syntaxique permet à elle seule de décrire la variabilité de la liaison associée aux séquences Adv+Adj. Cette différence fonctionnelle et distributionnelle est sans doute l'un des facteurs principaux qui permet d'expliquer l'écart des réalisations de la liaison entre les adverbes monosyllabiques. Toutefois, la variable DCS n'explique ni l'absence totale de LR dans les cas où *pas* est régi par l'adjectif ni l'écart des réalisations entre *très*, *plus*, *moins* et *trop*. En effet, nous supposons que d'autres facteurs jouent un rôle important dans la réalisation de la liaison variable à l'intérieur de la catégorie Adv+Adj. Pour cette raison, dans une deuxième étape de ce travail, nous avons procédé à la construction d'arbres de décision à l'aide des logiciels *Sipina* et *Weka*. Nous avons utilisé les arbres de décision afin d'établir une hiérarchisation des variables explicatives à partir de l'observation de notre échantillonnage. Pendant l'utilisation de *Sipina*, nous avons subdivisé l'ensemble de données en deux échantillons: un échantillon apprentissage correspondant à 338 occurrences (67%) et un échantillon test correspondant à 167 occurrences (33%).

Dès nos premiers résultats sur cet échantillon le descripteur FREQ M1 semble constituer la première variable utilisée dans le partitionnement (variable de segmentation). De fait il possède le meilleur coefficient de Tschuprow[14] T=0,6794 et constitue la variable de segmentation la plus pertinente parmi tous les descripteurs discrets et continus.

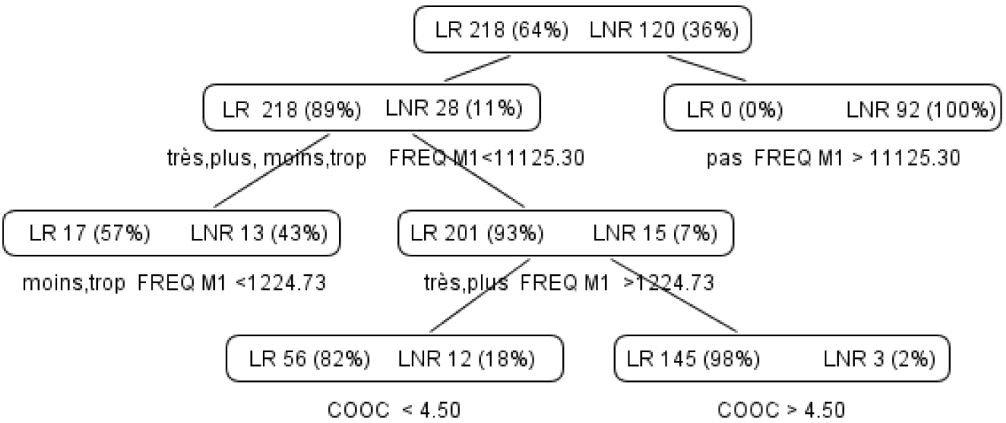

FIGURE 1 – Arbre de décision à partir de l'échantillon d'apprentissage des séquences Adv+Adj

La variable de segmentation FREQ M1 permet de classifier les données relatives à la variable

[13] « ah non ironique j+ *pas* ironique non mais euh j'sais que quand j'étais petit euh+euh j'étais très question réponse quoi » (Corpus CFPP2000_Corpus Mathieu Rosier et Elisa Rysnik).

[14] Indice de corrélation qui permet d'évaluer la qualité de la segmentation et donc de sélectionner le meilleur parmi les descripteurs, évaluant sa pertinence dans le partitionnement de l'arbre de décision.

REALISATION en deux groupes (arêtes). La valeur (seuil de coupure)[15] choisie pour opérer telle segmentation correspond à l'indice de fréquence 11.125,30. Ce point de coupure permet plus précisément de discrétiser, au premier niveau de l'arbre, entre deux groupes d'adverbes et donc de séquences liaisonnées qui leur sont associées. D'une part, l'adverbe pas, dont l'indice de fréquence est supérieur à 11.125,30. D'autre part, les adverbes monosyllabiques dont la fréquence est inférieure à 11.125,30, à savoir très, plus, moins et trop. Cette première classification repose aussi bien sur une dimension syntaxique dépendante de la variable DCS. Bien que celle-ci possède une indice de corrélation très élevé (Tschuprow T=0,6008) elle ne figure pas dans la représentation de l'arbre. En effet, cette variable ne parvient pas entièrement à classer la variable REALISATION. Il n'existe pas de correspondance précise entre la valeur LNR de la variable REALISATION et la valeur DV du descripteur DCS, puisque comme on l'a déjà observé, 10% des LNR associées à l'adverbe pas correspondent aux séquences où pas dépend de l'adjectif suivant (à savoir DA). En outre la variable syntaxique rend difficilement compte des différences de réalisation entre très, plus, moins, trop. En revanche, au deuxième niveau de l'arbre, la variable FREQ M1 permet de segmenter ce dernier ensemble en deux sous-groupes à travers un deuxième seuil de coupure correspondant à l'indice 1224,73. Le premier sous-groupe est constitué par les adverbes monosyllabiques dont l'indice de fréquence est inférieur à 1224,73, notamment moins et trop. Le deuxième est formé par les adverbes monosyllabiques dont la fréquence est supérieure à 1224, à savoir très et plus. Ce deuxième niveau de segmentation représente l'écart des réalisations existant entre d'une part très (95%) et plus (94%) et d'autre part moins (65%) et trop (59%). Autrement dit ce seuil de discrétisation permet de faire la différence parmi les adverbes monosyllabiques possédant la même distribution syntaxique. La variable de segmentation lexicale COOC (au troisième niveau de l'arbre) (Tschuprow T=0,0739) semblerait permettre d'expliquer les LNR des séquences associées aux adverbes très et plus comme possédant une fréquence de cooccurrence inférieure à 4,5 occurrences. Enfin, la variable FREQ M2 (Tschuprow T=0,05) semble exercer une faible influence sur la réalisation de la liaison ainsi que les variables phonologiques (NSYLL M2 et CL) qui ne possèdent presque aucune corrélation[16].

En résumant, la variable lexicale FREQ M1 semble être la variable la plus importante dans la description de la réalisation de la liaison permettant d'opérer des classifications des données à deux niveaux différents de l'arbre. Cependant, il faut préciser que bien que le descripteur FREQ M1 constitue la variable de segmentation la plus pertinente pour notre échantillon, notre classification ne tient compte que des adverbes monosyllabiques. Par conséquent cette variable ne possède que quatre valeurs correspondant aux indices des fréquences de *très*, *plus*, *moins*, *trop*. Il résulte que classifier les réalisations de la liaison sur la base du descripteur FREQ M1 revient à les classer à partir de l'adverbe monosyllabique lui-même. Autrement dit, il semblerait que la réalisation de la liaison des séquence Adv+Adj soit majoritairement influencée par la nature lexicale de l'adverbe même, strictement liée à sa fréquence. Afin d'évaluer la qualité de l'apprentissage, nous avons réalisé le *test set validation*[17] comparant la prédiction du modèle avec les résultats obtenus dans l'échantillon test. Nous constatons un taux d'erreur égal à 3,6%.

[15] Seuil établit par l'apprentissage supervisé à partir duquel la variable de segmentation permet de séparer les données du corpus d'apprentissage en deux arêtes.

[16] Cependant, nous supposons qu'une analyse menée sur un échantillon d'apprentissage plus grand pourrait mettre en évidence des tendances déjà observées par des études précédentes d'après lesquelles la fréquence du M2 peut jouer un rôle dans la réalisation de la liaison.

[17] Méthode de validation croisée utilisée pour estimer la fiabilité d'un modèle fondé sur une technique d'échantillonnage.

Dans un second temps, afin de vérifier la fiabilité de ces résultats, nous avons soumis le même corpus au logiciel *Weka*, utilisant une différente méthode de validation croisée (*K-fold cross-validation*). Cette deuxième analyse a confirmé nos résultats, sélectionnant la variable FREQ M1 comme variable de segmentation dont les meilleurs points de coupure sont les mêmes mis en évidence dans l'analyse précédente. Dans ce deuxième apprentissage, nous constatons un taux d'erreur sur l'ensemble de l'échantillon (505 instances) égal à 8,5% correspondant à 91,48% d'instances classifiées correctement.

6 Conclusions

Dans cette étude, nous avons présenté un premier travail exploratoire mené sur le corpus PFC et sur le corpus CFPP2000 qui nous a permis d'analyser conjointement l'incidence de plusieurs facteurs en jeu dans la production de la liaison. À partir des résultats de l'étude de cet échantillon nous avons relevé que la classe Adv+Adj possède une forte complexité interne liée à l'hétérogénéité qui caractérise les distributions syntaxiques et les indices de fréquence des adverbes monosyllabiques. Dans ce contexte, il nous apparaît nécessaire, tout en gardant l'économicité des classifications existantes, de rendre compte de façon plus précise, (à travers des outils relevant de la linguistique computationnelle), des comportements différents par rapport autres adverbes monosyllabiques. L'ensemble des résultats de cette analyse nous suggère que les classifications établies sur base morphosyntaxique, dépourvues de toute description interne, ne parviennent qu'à élaborer des généralisations qui permettent difficilement de saisir le fonctionnement variable de la liaison.

Dans cette première étude, malgré l'extension limitée de notre échantillon, nous avons essayé de donner un aperçu d'une analyse plus détaillée de la liaison variable en essayant d'établir une connexion entre les multiples facteurs en jeu dans la réalisation de la liaison et en soulevant des problèmes qui suggèrent de nouvelles pistes de recherche. Actuellement nous sommes en train d'élargir cette analyse multidimensionnelle aux autres contextes de liaison du corpus PFC. L'analyse d'un tel ensemble de données nous permettra de valider les considérations faites dans cette première étude et de vérifier l'incidence des facteurs phonologiques, syntaxiques et lexicaux dans la réalisation de la liaison en analysant toute occurrence de séquences liaisonnées à l'intérieur des classifications traditionnelles. Autrement dit, il s'agira de remettre en discussion la légitimité de l'appartenance aux catégories morphosyntaxiques en tant que seul critère de classification. À travers cette multiplication des plans d'analyse nous espérons parvenir à une description de la liaison qui se rapproche le plus possible de la variabilité des usages langagiers. Par conséquent, dans les études futures, il faudra établir une plus grande articulation entres les différentes dimensions, et réfléchir sur les liens existant entre les variables phonologiques et les facteurs lexicaux et entre ces derniers et la dimension syntaxique. De plus, il s'agira d'enrichir la réflexion sur la dimension syntaxique, de dépasser l'analyse superficielle des catégories morphosyntaxiques d'appartenance en étudiant les relations syntaxiques spécifiques à chaque séquence qui établissent les degrés de cohésion interne aux séquences liaisonnées[18].

[18] Laks (Laks, 2005).

Références

BLUMENTHAL, P. (1990). Classement des adverbes: Pas la Couleur, rien que la nuance? *Langue française*, 88(1):41-50.

BRANCA-ROSOFF, S., FLEURY, S., LEFEUVRE, F., PIRES, M. (2012). Discours sur la ville. Présentation du Corpus de Français Parlé Parisien des années 2000 (CFPP2000). http://cfpp2000.univ-paris3.fr/CFPP2000.pdf. [consulté le 27/01/2014].

BYBEE, J. (2001). Frequency effects on French liaison. *In (*Bybee et Hopper, 2001), 337-359.

BYBEE, J. et HOPPER, P., éditeurs (2001). *Frequency and the Emergence of Linguistic Structure*. John Benjamins.

CHRISTODOULIDES, G., AVANZI, M., GOLDMAN, J.P. (2014) DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. *In IX Language Resources and Evaluation* (*LREC 2014),* Reykjavik, Inslande. À paraître.

DELATTRE, P. (1947). La liaison en français, tendances et classification. *The French Review*, 22(2):148-157.

DELATTRE, P. (1966). *Studies in French and comparative Linguistics*. Mouton.

DURAND, J., LAKS, B. et LYCHE, C. (2002). La phonologie du français contemporain: usages, variétés et structure. *In (*Pusch et Raible, 2002), 93-106.

ENCREVÉ, P. (1988). *La liaison avec et sans enchainement*. Seuil

HOLMES, G, DONKIN, A et WITTEN, H.I. (1994). Weka: A machine learning workbench. *In Proceedings of Second Australia and New Zealand Conference on Intelligent Information Systems*. Brisbane, Australie.

LAKS, B. (2005). Phonologie et construction syntaxique: la liaison, un test de figement et de cohésion. *Linx*, 53:155-171.

MALLET, G. (2008). *La liaison en français: description et analyses dans le corpus PFC*. Thèse de Doctorat non publiée. Laboratoire MoDyCo, Université Paris Ouest Nanterre la Défense.

MELIS, L. (1983). *Les circonstants et la phrase*. Presses Universitaires de Louvain.

NEW, B.(2006). Lexique 3: Une nouvelle base de données lexicales. *In Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006)*, Louvain, Belgique.

PUSCH, C. et RAIBLE, W., éditeurs (2002). *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*. Gunter Narr Verlag.

SCHANE, S. A. (1967). La liaison et l'élision en français. *Langages*, 8:37-59.

TRANEL, B. (2000). Aspects de la phonologie du français et la théorie de l'optimalité. *Langue française*, 126:39-72.

WILMET, M. (2010). *Grammaire critique du français*. Duculot.

ZIGHED, D., AURAY, J., DURU, G. (1992). *Sipina: Méthode et Logiciel*. Lacassagne.