

Exploration des segments en périphérie droite en français oral informel

Analyse de corpus et perspectives d'identification automatisée

Mémoire réalisé par
Émile BEGUIN

Promoteur(s)
Liesbeth DEGAND et Thomas FRANÇOIS

Année académique 2016-2017
Master [120] en linguistique, à finalité spécialisée en traitement automatique du langage

INTRODUCTION

Le déploiement temporel est la condition de production des discours oraux : lorsqu'il s'exprime, un locuteur doit nécessairement répondre à des lois d'une succession temporelle. Loin d'être anodine, cette temporalité a parfois été minimisée dans ses effets en-deçà des discours. Si elle structure l'interaction, modèle les discours, agence les informations, peu de travaux ont investigué les effets propres de cette séquentialité sur la morphosyntaxe, sur la sémantique des énoncés et sur les actes de langage. Récemment, cependant, Beeching & Detges (2014) et Beeching (2017) ont pris compte de cette temporalité dans laquelle sont projetées les « unités linguistiques » : une conséquence logique de cette temporalité est ainsi, d'après eux, que les marges de ces unités linguistiques sont contraintes de prendre des fonctions différentes. De là, ils proposent d'investiguer, dans ses aspects fonctionnels, la notion de *périphérie*.

Plusieurs faits conduisent à penser que les périphéries représentent une entité linguistique pertinente, particulièrement dans l'analyse des productions orales. La littérature s'intéresse depuis longtemps à ces éléments, à la frontière du discours et de la syntaxe (Danon-Boileau et al., 1991). De nombreuses recherches se sont penchées sur celles-ci : dislocation à droite et à gauche (Nølke 1996 ; De Cat 2007) ; structure en thème-rhème-postrhème (Danon-Boileau et al. 1991) ; préfixes et postfixes (Blanche-Benveniste, 1990). Les périphéries seraient le lieu de plusieurs fonctions discursives : la littérature leur donne un rôle intersubjectif, en comparaison à une périphérie gauche qui jouerait un rôle subjectif (Traugott, 2012).

Jusqu'ici, toutefois, les périphéries ont été investiguées à partir d'une heuristique *onomasiologique* : les recherches consistaient à épingler une ou plusieurs formes présentant un usage caractérisé en périphérie. Ce n'est que récemment qu'a été proposée une heuristique *sémasiologique* pour les segments en périphérie : avec le projet 'FRFC' (Tanguy, Van Damme, Degand, & Simon, 2012), la périphérie gauche devient ainsi un paradigme observable au sein

duquel de nouvelles formes peuvent être découvertes. Le travail, toutefois, restait à faire pour les périphéries droites.

En réponse à ce manque actuel de la recherche, ce mémoire entend proposer une démarche d'identification des périphéries droites dans les données des transcriptions orales. La trame générale de l'exposé est la suivante : une définition théorique des périphéries droites est proposée, articulant diverses théories grammaticales ; cette définition est ensuite opérationnalisée pour une annotation manuelle exploratoire sur corpus. À la suite de cette annotation manuelle et de ses résultats, des perspectives de traitement automatique sont esquissées afin de proposer un outil performant pour la recherche linguistique sur corpus, que les limites de ce mémoire n'ont pas permis d'implémenter.

Le premier chapitre de ce mémoire s'attache à définir les périphéries droites en tant qu'entité de segmentation de l'oral, afin d'en préparer la description systématique. La question que soulève la notion de périphérie est celle d'une définition d'un « centre » : à cet égard, la notion de « phrase », en tant qu'axiome pertinent pour l'étude et la segmentation de productions authentiques, est présentée comme théoriquement et méthodologiquement problématique. Plusieurs théories répondent à ce problème de segmentation en reconsidérant les relations entre syntaxe et discours ; celui de l'*Approche pronominale* (Blanche-Benveniste, 1990) développé par le *Groupe Aixois de Recherches en Syntaxe* (GARS), est retenu comme socle définitionnel pour les segments périphériques droits. De là, la notion de périphérie droite est définie en fonction de critères à la fois micro- et macrosyntaxiques : elle est située en dehors de la rection du verbe et ne peut former un noyau macrosyntaxique ; elle est composée de l'ensemble des éléments qui, tout en étant non régis et non-noyaux, sont néanmoins des dépendants macrosyntaxiques d'un noyau situé à gauche. Grâce au modèle de l'*Approche pronominale*, plusieurs catégories formelles de périphérie droite sont également envisagées, qui seront investiguées dans le cadre de l'annotation.

Le second chapitre élabore un cadre méthodologique pour l'annotation des segments périphériques droits. Les données considérées pour l'exploration du phénomène sont deux

corpus provenant de la banque de données Valibel (Bachy, et al., 2007). Ces corpus, qui contiennent des conversations informelles enregistrées en Belgique francophone entre 2004 et 2005, offrent l'opportunité d'approcher le phénomène des périphéries droites dans la dynamique des interactions. Compte tenu de ces données, ce mémoire propose une méthode d'annotation opératoire et reproductible, appliquant les principes de l'*Approche pronominale* à des conversations orales transcrites, indépendamment des données d'intonation. La méthode proposée définit ainsi les périphéries droites à partir de deux unités : les *unités de rection* et les *groupes intonatifs*. Les *unités de rection* forment la délimitation gauche, exclusive, des segments périphériques droits ; elles agrègent les propriétés des domaines régis et les propriétés macrosyntaxiques des noyaux. Les *groupes intonatifs* forment la délimitation inclusive de tous les segments périphériques droits, mais aussi gauches ; ils reprennent l'ensemble des affixes macrosyntaxiques dépendants d'un noyau. Au terme du second chapitre sont décrites les spécifications qui permettent de délimiter un segment périphérique droit et d'en discriminer les différents éléments constitutifs. Une annotation de ces éléments est ensuite proposée, relative aux caractéristiques positionnelles de ces éléments, aux catégories morphologiques qu'ils font intervenir, et aux fonctions – textuelles et illocutoires – que ces éléments sont susceptibles de jouer dans l'énonciation.

Le troisième chapitre de ce mémoire analyse les données produites par l'annotation. L'identification des périphéries droites, première étape de l'annotation, montre les régularités d'occurrence de ces segments dans les deux corpus de conversations informelles. Cette partie est également l'occasion de décrire les biais principaux d'identification des périphéries droite en l'absence de données d'intonation. Après avoir analysé les mesures relatives à l'identification, le chapitre considère les données produites par la catégorisation des éléments constitutifs périphéries droites. Les interrelations entre catégories formelles, catégories fonctionnelles et données positionnelles y sont investiguées, avant de proposer une analyse des séquences d'éléments sur base de ces mêmes catégories. Cette dernière analyse, qui repose sur une mesure des fréquences observées de transitions entre catégories, fournit plusieurs données cruciales pour le choix d'un modèle adéquat de reconnaissance automatique.

Prenant appui sur ces périphéries droites désormais identifiées et mesurées, le dernier chapitre de ce mémoire envisage des possibilités d'implémentation à la suite de cette exploration sur corpus. Après avoir décrit le résultat à produire et la tâche à réaliser, deux perspectives sont considérées : une implémentation qui identifierait les périphéries droites à partir d'une analyse automatique des dépendances syntaxiques, et une autre implémentation utilisant les champs conditionnels aléatoires (*'contextual random fields'*, CRF).

Ce mémoire se clôt par une conclusion rétrospective, discutant les résultats pour envisager les perspectives théoriques des phénomènes observés et des méthodes proposées.

I. CADRE THÉORIQUE

Ce chapitre définit la périphérie droite en utilisant le matériel théorique et conceptuel développé dans le cadre de diverses approches de la segmentation du discours et de la grammaire du français en usage. Il propose une définition des périphéries droites et en décrit certaines caractéristiques permettant leur identification dans des données de production authentiques.

La première partie de ce chapitre dresse un tour d'horizon des théories ayant proposé un cadre notionnel et descriptif dans ces domaines. La question centrale de la frontière entre syntaxe et discours est le fil conducteur de l'exposé ; celle-ci est problématisée à partir de la notion de phrase, axiologique dans bien des sous-disciplines des sciences du langage. Plusieurs modèles de description faisant face à cette problématique sont ensuite présentés : le modèle de Genève de hiérarchie de l'interaction, le modèle de Fribourg et la *Grammaire de la période*, et enfin le modèle aixois et l'*Approche pronominale*. Les approches sont ensuite comparées dans leur pertinence descriptive pour capter le phénomène des périphéries droites.

La seconde partie s'intéresse aux différentes mentions de phénomènes de « périphérie droite » dans la littérature linguistique en général. Dans cette partie, je qualifie de « segments postfiniaux » le phénomène empirique renvoyant aux périphéries droites : c'est le paradigme général, quelle que soit la théorie qui l'englobe et la façon dont il est approché par la description linguistique. Ces théories sont abordées dans le but de construire une définition opératoire pour l'annotation ; celle-ci peut être construite à partir de plusieurs théories et qui peut être appliquée de façon déductive.

La dernière partie de ce chapitre propose une typologie des périphéries droites et en identifie certaines caractéristiques formelles. À partir des cadres théoriques considérés, plusieurs formes syntaxiques corrélées avec le concept, désormais défini, de périphérie droite, sont proposées. Dans un deuxième temps, cette partie les caractérise en termes de schèmes

intonatifs en proposant qu'il existe une intonation prototypique et non-marquée des segments périphériques droits.

1. APPROCHES DE LA PÉRIPHÉRIE

Définir et caractériser les périphéries dans le discours demande de se référer à un cadre théorique de description linguistique. Du fait que les périphéries se situent entre syntaxe et discours, plusieurs approches ayant modélisé les relations entre ces deux modules de langue sont proposées dans cette section.

Pour fixer le cadre de cet exposé, je problématise la notion de phrase, pour montrer qu'une description des phénomènes en marge de la syntaxe et du discours – dont les périphéries – doit se poser la question d'un appareil conceptuel spécifique. Ainsi, le domaine de la syntaxe – relations qui se situent « en-deçà de la phrase » – et le domaine du discours – relations qui se situent « au-delà de la phrase » – doivent dès lors être repensés dans leurs interrelations.

En réponse à cette problématique d'une redéfinition des liens entre syntaxe et discours, j'aborde successivement plusieurs orientations théoriques différentes, se différenciant en termes de niveaux d'analyse linguistique privilégiés : le modèle de Genève est orienté vers les structures discursives et pragmatiques ; le modèle du GARS, qui comprend l'« *Approche pronominale* », est orienté vers les structures syntaxiques ; la *Grammaire de la période* est une approche mixte. À travers chaque approche exposée, l'objectif est de préparer les bases d'un modèle descriptif des périphéries qui sera utile à leur définition dans la section 2, et à leur description structurelle et fonctionnelle dans la section 3.

1.1. LIMITES DE LA PHRASE

Comprendre les périphéries droites n'est possible qu'en comprenant l'articulation complexe entre syntaxe et discours. Dans cette partie, je propose, à la suite de Berrendonner & Béguelin (2012), quatre « ingrédients » par lesquels la phrase se voit erronément justifiée comme unité

linguistique qui permettrait de rendre compte des relations entre discours et syntaxe. J'en développe ensuite, au travers d'un exemple, une récusation – parmi plusieurs possibles – qui sera l'occasion de différencier deux visions de la relation entre discours et syntaxe, menant à des paradigmes théoriques distincts.

Berrendonner & Béguelin (2012) avancent quatre « ingrédients » qui conduisent le concept de phrase à être utilisé sans remise en question. C'est l'association de plusieurs de ces ingrédients qui mènera à la perception, tronquée, que la phrase est une unité pertinente en linguistique.

- (α) La *maximalité syntaxique* est un postulat de la grammaire traditionnelle par lequel la phrase est considérée comme « l'unité maximale de la langue, donc comme le terme de la combinatoire syntaxique » (2012:6).
- (β) Le postulat de *complétude sémantique*, quant à lui, repose sur l'idée qu'une phrase véhicule un « sens complet ». Berrendonner et Béguelin (2012:7) considèrent que les modalités de cette idée de complétude en restreignent les critères : l'association d'un acte illocutoire avec un contenu propositionnel chez Searle, par exemple, ou encore le regroupement d'un thème et d'un rhème.
- (γ) La conviction en l'existence de *démarcations prosodiques* de la phrase la considère comme une suite intonative bornée, à droite, par la présence à la fois d'une pause et d'une intonation descendante.
- (δ) L'habitude de la *démarcation graphique*, qui n'investit pas seulement les grammaires traditionnelles, mais conditionne aussi les réels observés par certaines syntaxes formelles ou traités de pragmatiques.

À mon sens, l'ingrédient α est en réalité un double postulat théorique :

- (α -1) La phrase circonscrit les relations entre toutes les catégories grammaticales ;
- (α -2) Toute relation que la phrase circonscrit peut être exprimée en termes grammaticaux.

Le postulat α -1 est falsifié dans le cas où la phrase n'est plus la condition nécessaire des relations syntaxiques ; les relations syntaxiques pourraient donc exister hors des limites fixées par la phrase. Le postulat α -2, quant à lui, est falsifié dès lors que la phrase cesse d'être le lieu *suffisant* d'articulation des relations syntaxiques. Dans ce cas de figure, d'autres relations de nature *non syntaxique* peuvent exister dans ce que l'on avait coutume de nommer « phrases ».

Dans l'exemple (1) cité par Blanche-Benveniste, volontairement non ponctué pour éviter tout parasitage des démarcations graphiques, la structure syntaxique '*tu as gagné une fois*' peut être interprétée comme « projetée » sur le second élément '*moi trois fois*'.

(1) Tu as gagné une fois moi trois fois (Blanche-Benveniste 1990:151)

Comment expliquer l'absence d'un second verbe ? Comment rendre compte de la relation évidente entre ces deux parties ? S'agit-il là d'une phrase ou de deux ? Cet exemple est éloquent dans la mesure où rencontrer une telle projection de structure syntaxique est parfaitement envisageable dans des textes écrits, même formels. Dans le cas écrit, plusieurs ponctuations peuvent être « imaginées » qui ne changeront pas, au demeurant, la relation, à mi-chemin entre syntaxe et discours, qui unit les deux propositions. Le grammairien fait face à une alternative dans la segmentation des deux unités :

(1) [Tu as gagné une fois] [moi trois fois]

(2) [Tu as gagné une fois moi trois fois]

Une possibilité que je qualifierai d'option « syntaxique » est de considérer cet extrait comme constitué de deux ensembles coordonnés ou juxtaposés ; dans ce cas de figure, expliquer la disparition du verbe dans la deuxième partie demande de prendre en compte des relations syntaxiques opérant à un autre niveau, situé au-delà des limites de la phrase, mais en reconsidérant alors le concept de syntaxe et en laissant intacte la nucléarité des propositions/énoncés. Le premier postulat concernant les liens entre phrase et syntaxe est alors falsifié. La seconde possibilité que je qualifierais d'« extensive » est de comprendre l'ensemble comme une seule entité ; des relations non syntaxiques doivent alors être décrites

au sein de la grammaire : des relations en partie logico-sémantiques, mais aussi et surtout pragmatiques intègrent donc le modèle. C'est alors le second postulat concernant la relation entre phrase et syntaxe qui se trouve falsifié.

1.2. NOUVELLES APPROCHES DU DISCOURS

Si j'ai présenté ces deux possibilités d'interprétation face à un exemple concret, c'est parce qu'elles sont l'occasion d'esquisser la distinction entre deux approches : les approches « orientées discours » (Bordería, 2014:5), tournées vers cette vision « extensive » de la grammaire, et de l'autre côté les approches « orientées syntaxe » préférant au contraire une approche « graduelle » de la syntaxe vers le discours. Dans les deux cas, le concept de phrase est considéré comme caduc et la syntaxe, inadaptée, doit être repensée.

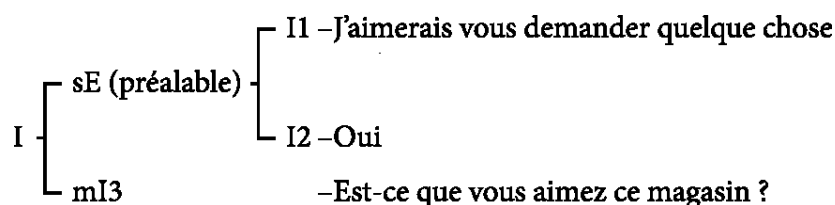
Les approches « orientées discours » proposent de ne plus s'en tenir à une syntaxe descriptive des données observées. L'analyse structurale, héritée de la tradition saussurienne, est abandonnée au profit d'une conception avant tout pragmatique des structures discursives (Bordería, 2014:6). Plusieurs courants d'analyse font ce choix¹, et conçoivent ainsi une quantité de modèles approchant de diverses manières la structuration des textes à partir des usages discursifs. Les différences d'orientation résident principalement dans (1) les unités minimales et les unités maximales d'analyse considérées ; (2) leurs critères de définition fondés sur différents observables langagiers ; (3) l'importance accordée, en contexte interactif, à la coénonciation des unités, et le niveau minimal où celle-ci peut intervenir dans

¹ Le modèle de Bâle (Ferrari 2005, citée par Bordería 2014) repose ainsi sur l'organisation hiérarchique des unités textuelles : les énoncés, subdivisés en unités d'informations : '*frame*', '*nucleus*' et '*appendix*', sont subordonnés à des unités communicatives, elles-mêmes subordonnées à des mouvements textuels. Ce modèle est adapté à une analyse de l'écrit et peut fonctionner en tandem avec une modélisation (micro-)syntaxique. En contraste par rapport à ce premier exemple, le modèle *Val.Es.Co*, développé sur l'espagnol parlé informel (Briz Gomez et al. 2003), n'adopte pas une approche hiérarchique, mais bien séquentielle, reposant sur les actes de langage et l'organisation des tours de parole. Il se fonde sur la notion d'intervention et celle de tour, où un tour est défini comme une intervention contribuant au « développement thématique ou sémantique de la conversation » (Bordería, 2014, p. 12). Il prend en compte la possibilité que des indices faciaux, gesturaux ou paraverbaux constituent des interventions et est donc un outil intéressant pour l'interaction orale.

le discours et enfin (4) l'orientation du modèle vers une organisation séquentielle, relativement hiérarchique ou absolument hiérarchique de chacune des unités reconnues.

1.2.1. LE MODÈLE DE GENÈVE : UNE APPROCHE ORIENTÉE VERS LES DISCOURS

Le modèle de Genève (Roulet et al. 1985) est probablement l'un des plus emblématiques parmi approches « orientées usage ». Développé dans le cadre de l'analyse des langues romanes, il a largement influencé les modèles ultérieurs en introduisant l'idée d'une structure hiérarchique – et générative – opérant à l'échelle des discours. Le modèle est centré sur les interactions – avec la conversation orale informelle comme prototype de celles-ci. Il prend pour unité de base les actes de langage, lesquels peuvent tantôt être isolés et sont alors qualifiés d'*interventions*² ('I' sur le graphique), ou bien être organisés en structures plus complexes qualifiées d'*échanges*. L'exemple suivant, par exemple (tiré de Bordería, 2014:10) fait intervenir deux échanges; les deux premiers éléments sont utilisés pour introduire le troisième.



L'intérêt d'un tel modèle est de présenter la conversation comme bien plus qu'une succession de tours ; il permet d'envisager directement les fonctions jouées par chaque élément dans la dynamique de l'interaction.

L'essence interactive des périphéries droites gagnerait à être étudiée à la lumière de tels modèles. Comprendre quelles structures de discours et quels types d'interventions contiendront des segments périphériques peut se révéler une observation intéressante. Beeching (2017) prend cette direction en proposant une définition des périphéries droites qui se fonde sur la structure des échanges dans l'interaction. Elle reprend pour ce faire le concept

² Les premières publications du groupe de Genève utilisaient le terme de « mouvement ».

de *paires adjacentes* (*question-réponse, requête-confirmation*) développé par Sacks et al. (1974) qui les décrivent comme des éléments structurants de l'interaction, de façon comparables aux *échanges* du modèle de Genève. Beeching (2017:208-209) propose, à partir de ce cadre des *paires adjacentes*, de considérer les périphéries gauches et droites comme les moments de l'interaction dans lesquels se négocient les interventions de chaque locuteur.

Le modèle de Genève, de même que les autres modèles constitués à partir des interactions, prend véritablement son indépendance vis-à-vis des données linguistiques et s'aventure loin dans la prise en compte de phénomènes paralinguistiques³. Un acte de langage, de fait, est entièrement indépendant de sa forme linguistique et peut se composer d'éléments prosodiques, posturaux ou gesturaux. Les données verbales et mêmes paraverbales ne sont dès lors qu'un support parmi d'autres de l'analyse.

En termes d'automatisation, on fait donc face à un champ de la linguistique dont les perspectives d'implémentation ont jusqu'ici été largement sous-explorées pour la raison qu'il serait difficile de rendre compte de données aussi riches, variées et multi-niveaux que celles qui interviennent dans le cadre du modèle de Genève.

1.2.2. LA GRAMMAIRE DE LA PÉRIODE : UNE APPROCHE MIXTE

La *Grammaire de la période*, développée autour de la figure d'Alain Berrendonner à l'université de Fribourg, est à la fois une approche « orientée usage », dans la mesure où elle considère avant tout le niveau pragmatique pour structurer les discours – la notion de « période » est une unité praxéologique –, et à la fois une approche « orientée syntaxe » dès lors qu'une syntaxe descriptive est acceptée au sein du modèle comme « un point de départ adéquat, quoiqu'incomplet, pour le champ de l'analyse du discours » (Bordería 2014⁴).

³ Alors que la prosodie elle-même n'est pas reconnue comme un élément proprement essentiel de l'analyse, dans la mesure où le même modèle peut être utilisé pour interpréter l'écrit.

⁴ Bordería (Bordería, 2014, p. 11) classe pourtant le modèle de Fribourg parmi les approches « orientées usage ».

Cette mixité des approches tient au fait que deux niveaux sont simultanément considérés et cependant rigoureusement dissociés. L'objectif de la grammaire de la période est d'identifier des régularités de dépendance, d'autonomie et de maximalité pour chacun de ces niveaux (Blanche-Benveniste, Postface, 2012:343). D'un côté, la *microsyntaxe* est une proposition de théorie morphosyntaxique ; elle agence les *morphèmes* en unités maximales que sont les *clauses*. De l'autre, la *macrosyntaxe* se laisse comprendre comme une « pragmasyntaxe » et agence les *énonciations* dans des unités maximales que sont les *périodes*. Une position importante de la *Grammaire de la période* est donc qu'il n'existe donc plus d'équivalent à une notion de « phrase » qui serait tour à tour l'unité maximale de la syntaxe et l'unité de base des discours et de la pragmatique ; deux « univers » sémiotiques distincts et possédant « [leur] syntagmatique propre » (Berendonner, 2002:13) sont assimilés par la *Grammaire de la période* à une deuxième et une troisième articulation du langage – selon les termes de Martinet.

(1) LA PÉRIODE

La *période* est l'unité maximale des relations macrosyntaxiques. Elle représente « un programme praxéologique complet » (Avanzi, 2005:3). Au niveau des périodes se traduisent les actes de langage et l'évolution de la « mémoire discursive ». Les périodes enregistrent donc les différents « moments » du discours, lesquels sont modélisés comme des suites séquentielles dans lesquelles les périodes font évoluer la mémoire discursive d'un état *A* à un état *A'*. La période est une unité interactive. Sa clôture exprime la complétude d'une intervention du locuteur, si bien qu'un interlocuteur a virtuellement l'occasion de prendre la parole lorsque celle-ci s'achève.

Une période n'est pas une unité planifiée à l'avance par l'« énonciateur », et n'est pas davantage l'application de règles génératives préexistantes. Le groupe de Fribourg postule néanmoins que les périodes possèdent une structure interne qui est constituée d'un ou plusieurs « macrosyntagmes », dont la relation sémantique est signifiée à partir de schèmes

intonatifs⁵. Les faits d'intonation, toutefois, ne suffisent pas à établir la fin d'une période : c'est la complétude de l'acte de langage qui aura le dernier mot. Ainsi, l'extrait suivant, correspondant à un unique acte de narration, sera analysé comme une seule période :

juste à un moment j'avais un peu peur parce que tu vois on devait descendre comme ça et j'étais assuré et parce que il y avait c'était des enfants alors j'avais peur qu'ils qu'ils me qu'ils me lâchent alors ils me tenaient comme ça après moi je descendais comme ça comme une araignée je me tenais à la corde après je descendais et un moment ils m- ils me lâchaient un peu j'avais peur moi parce que j'avais peur de me blesser [oral, GRE03]

L'ensemble est interprété par Avanzi comme une *période sérielle* (2005:8-9) ; celles-ci sont courantes dans le cas d'une narration ou d'une prescription d'itinéraire.

Il est intéressant de comparer les *périodes* aux *interventions* du modèle de Genève : définies elles aussi par leur valeur pragmatique, les énonciations s'en démarquent à partir de deux aspects : une construction monologique⁶ et la considération de l'évolution graduelle d'un état discursif. En comparaison avec le modèle genevois, le groupe de Fribourg ne postule pas d'une coénonciation⁷ première, et s'oriente davantage vers une modélisation séquentielle du discours.

(2) LES ÉNONCIATIONS

Au sein des périodes, les *énonciations*, que j'ai présentées comme les éléments nucléaires de la macrosyntaxe, possèdent trois propriétés (Avanzi, 2005) :

Premièrement, une *énonciation* est avant tout une unité de *communication* davantage qu'une unité *signification* (Avanzi, 2005:2) : elle se définit par le mouvement qu'elle imprime dans

⁵ Dans son analyse de l'intonation, l'école de Fribourg n'utilise pas de données quantifiées ni de prosogrammes : l'approche est celle d'une interprétation subjective des données enregistrées.

⁶ Dans la grammaire de la période, deux tours de parole contigus ne sont jamais analysés comme une seule et même clause : on leur préfère une interprétation elliptique :

[Si] on a affaire à une suite dialogale, composée de deux répliques ou de tours de parole adjacents, nous considérerons que les fragments produits par chaque énonciateur sont ipso facto des clauses distinctes, quitte à les analyser comme elliptiques (Berrendonner, 2012, p. 61)

⁷ Berrendonner (2002) reprend le concept de *détermination* au sens où l'entend Hjelmslev (1968) dans le cadre de la 'glossématique', le lien de détermination est une relation par laquelle "un terme présuppose l'autre mais non réciproquement".

l'évolution représentations partagées par les interlocuteurs. Dès lors, l'énonciation n'est pas nécessairement faite d'énoncés verbalisés : elle peut s'accompagner ou être remplacée par des signaux gesturaux, faciaux ou paraverbaux, dès lors que ceux-ci acquièrent une valeur pragmatique pertinente dans le discours.

Une seconde propriété des énonciations est de constituer le niveau où se manifestent les *intonèmes*. Dans la *Grammaire de la période*, les intonèmes ont des schèmes intonatifs parfois très complexes, mais possédant le statut d'unités pleinement linguistiques, pourvues d'un sens prototypique et dotées d'une valeur oppositive, ce qui en fait des unités *-émiques* (Avanzi, 2005:2). Sur le plan méthodologique, le groupe de Fribourg s'intéresse aux « effets perçus » de l'intonation, ce qui suppose un biais, assumé par les chercheurs, de « filtrage et de prototypification » (Berrendonner, 2012:94). Si le signifiant intonatif est traité de façon lacunaire, c'est parce que la priorité est donnée à un traitement des signifiés que contiennent les données d'intonation.

Enfin, l'énonciation articule à l'échelle macrosyntaxique des *clauses*, unités maximales de la microsyntaxe. Ce rôle d'articulation est à concevoir selon le même type de relations que celui par lequel les phonèmes articulent des morphèmes. La clause est moins étudiée par la *Grammaire de la période* qui la voit comme une « algèbre de concaténation restreinte » (Berrendonner, 2002).

(3) LES CLAUSES

Au niveau microsyntaxique, les clauses sont à concevoir, dans le flux du langage, comme des « ilots de dépendance syntaxique ». C'est à l'échelle des clauses que peut se concevoir la notion de maximalité syntaxique, dans la mesure où une clause se différencie précisément de ce qui l'entoure par ce lien de dépendance (Berrendonner, 2002:6), qui se traduit par les liens de *rection* au sens où l'entend Hjelmslev.

Berrendonner (2002:8) caractérise la clause à partir de trois traits principaux :

- (i) ses unités minimales sont des segments sonores à fonction significative (les morphèmes)

- (ii) ils sont assemblés par concaténation, de façon à former des unités complexes de même fonction : syntagmes et clauses ne sont que de « grands signes », dont le signifié est un produit « compositionnel » des signifiés véhiculés par leurs constituants.
- (iii) Les règles de la grammaire clausale peuvent être exprimées avec toute la généralité souhaitable en termes de relations formelles (implication, ordre, inclusion) exigibles entre segments de chaîne cooccurents.

Dans la *Grammaire de la période*, la structure de la clause est l'occasion d'une profonde restructuration de la syntaxe. Tout comme la notion de phrase, celle de « mot » est abandonnée au profit d'une unité minimale non équivoque : les morphèmes. La préoccupation principale de la *Grammaire de la période* est de trancher, en microsyntaxe, entre les liens hypotactiques et paratactiques. Il s'agit donc de *délimiter* les clauses plutôt que d'en décrire la structure interne.

1.2.3. L'APPROCHE PRONOMINALE : UNE ORIENTATION SYNTAXIQUE

Le *Groupe Aixoise de Recherche en Syntaxe* (GARS) a été, au milieu des années 1980, le premier groupe de chercheurs à utiliser le concept de *macrosyntaxe* (Blanche-Benveniste, Delofeu, Stefanini, & Van den Eynde, 1984). Dès les premiers temps de la théorie, la préoccupation du GARS est de décrire la *rection* pour en faire un modèle « tout-terrain » pour l'analyse syntaxique de données réelles issues de l'usage. Pour cela, la démarche originale proposée est celle d'une « *Approche pronominale* » : les constituants liés au verbe par rection sont susceptibles d'être proportionnels à des proformes pronominales. Le GARS distingue ensuite une « syntaxe interne » d'une « syntaxe externe » (Blanche-Benveniste, 2012:340), cette dernière ayant pour rôle de traiter des relations apparemment syntaxiques mais néanmoins situées hors du domaine de la rection.

Il s'agit donc bien d'une approche « orientée syntaxe » : les faits de syntaxe ne sont pas rejetés par la théorie, mais considérés comme inadéquats pour permettre une théorie intégrée de la syntaxe dans le discours. La façon dont le GARS entend construire un nouvel outil grammatical est par conséquent bien « orientée syntaxe » et, corollairement, *analytique* : il

s'agit de partir des petits éléments pour ensuite s'acheminer vers des généralisations valables sur les « grandes unités » (Blanche-Benveniste, 2012:341).

(1) LA RECTION

La rection est la relation qui caractérise des constituants dont on peut dire que l'un est dépendant d'un autre. Prototypiquement, la rection est un phénomène qui concerne les éléments verbaux – le cas d'une rection nominale est toutefois également considérée dans le modèle (Blanche-Benveniste, 1990:103). La démarche du GARS consiste à envisager toute forme de rection à partir de la possibilité de réduction de tous les éléments de la construction (verbale ou nominale) à des pronoms. Ce choix théorique n'est pas motivé seulement par le besoin d'objectiver les analyses en utilisant une approche inductive des règles syntaxiques. Il se veut également, pour le GARS, correspondre à « une opération très naturelle, que les locuteurs font constamment dans leurs productions orales » (Blanche-Benveniste, 1990:41).

La liste de proformes d'éléments régis admis dans l'*Approche pronominale* correspond à une classe large qui s'étend au-delà des pronoms clitiques jusqu'à des formes en marge du phénomène de pronominalisation, souvent étiquetées comme des pronoms « démonstratifs », « indéfinis » ou encore « adverbiaux ».

Pour comprendre l'intérêt de cette « *Approche pronominale* » pour opérationnaliser les faits de rection, appliquons-la à des extraits issus de notre corpus⁸. Des productions telles que :

(1) je ne l'ai jamais mise moi à un stage à Fontin (Valibel, BlaNB1l)

(2) apparemment euh c'était quelqu'un de célèbre / dans dans le coin (Valibel, FamRM1r)

sont proportionnelles aux constructions suivantes :

(1) je ne l'ai jamais mise moi là

(2) apparemment euh c'en était un ici

D'emblée, la délimitation d'un domaine régi à partir de l'*Approche pronominale* exclut certains types d'éléments qui seront considérés comme non-régis – mais susceptibles, comme

⁸ Je reporte le lecteur au second chapitre de ce mémoire pour une présentation des données du corpus.

nous le verrons, d'entretenir une relation de dépendance macrosyntaxique et d'être en conséquence considérés comme des éléments « périphériques ». C'est notamment le cas des pronoms toniques – '*moi*', dans l'exemple (1) entre dans une catégorie d'« associés » à la rection – ou encore des marqueurs de discours – '*apparemment*', dans l'exemple (2), opère au niveau discursif.

La première construction, '*je ne l'ai jamais mise moi là*', illustre le caractère parfois discontinu de la rection dans la chaîne parlée. Le pronom tonique '*moi*' forme ainsi une incise traversée par la rection qu'attire le verbe '*mise*' ; le terme '*un stage à Fontin*' peut être substitué par une proforme pronominale sans même qu'un déplacement du substitut soit nécessaire.

Une seconde observation possible à travers l'extrait (1) est la forme '*ne ... jamais*' : outre les éléments réductibles à des pronoms, celle-ci témoigne de l'existence d'autres éléments non-réductibles qui gravitent également autour des éléments recteurs et que le GARS nomme « modalités » syntaxiques. Celles-ci peuvent, dans le modèle du GARS, être de trois types : assertives, négatives ou interrogatives. Corollairement, un élément non recteur se caractérisera par une impossibilité d'accueillir des modalités ; un aspect intéressant dans l'étude des périphéries.

Le second exemple '*apparemment euh c'en était un ici*' est l'occasion de montrer un *dispositif de rection*, c'est-à-dire un agencement de rection des constituants moins stéréotypé que le classique « sujet-verbe » : le dispositif de « support », basé ici sur la formule « c'était ... ». Plusieurs de ces dispositifs sont ainsi possibles qui n'organisent pas de la même façon les constituants : le *dispositif direct* (où la relation de rection ne repose que sur le verbe lui-même et sa position) ; le *dispositif d'extraction* (de type '*c'est ... qui/que ...*') ; le *dispositif pseudo-clivé* (de type '*ce qui/que ... c'est ...*') ; les dispositifs de *support* (de type '*il y a ...*' ou '*c'est ...*') et enfin le *dispositif relatif* (faisant intervenir des propositions relatives). Benveniste insiste sur le fait qu'il n'existe pas de dispositif « canonique » opposé à des dispositifs dérivés (1990:55), mais que chacun, au même niveau, est une « représentation particulière de la construction verbale ».

L'exemple (2) montre également un élément isolé, '*apparemment*', qui se comporte comme un élément périphérique dont le sens porte sur l'ensemble de la construction régie. Certains adverbes dans des positions similaires se comporteront en tant qu'éléments régis et seront substituables à des proformes : '*auparavant euh c'en était un ici*' peut être réduit à '*c'en était alors un ici*'.

(2) LIMITES DE L'APPROCHE PRONOMINALE ET DE LA MICROSyntaxe

Le GARS en vient à déduire l'existence d'une *macrosyntaxe* à partir de son propre modèle de traitement des domaines régis. Tout part d'un constat : au-delà du lien de dépendance entre catégories grammaticales que représente la rection, d'autres relations entre éléments existent qui ne peuvent pas être décrites en termes de rection. Démontrons cet aspect en appliquant l'*approche pronominale* à l'extrait suivant :

(1a) [pourquoi] [on n'en sait rien]

(Valibel, BlaNB1l)

Deux groupes de rection se démarquent dans cet extrait⁹. Entre eux, une relation est perceptible. Encore est-il question de l'expliquer. La réutilisation « en sens inverse » des outils de l'*Approche pronominale* présentés plus haut peut apporter un éclairage : le pronom '*en*' peut ainsi hypothétiquement s'étendre un terme nominal qui serait également sous-tendu par '*pourquoi*' et serait donc un élément contenu dans le discours :

(1b) [pourquoi tout ça] [on ne sait rien de tout ça]

Force est de constater que la « relation » entre les deux groupes de rection, toutefois, n'est pas épuisée par cette explication. Soit en effet la construction suivante :

(1c) [pourquoi] [on ne sait pas]

Ici, aucun pronom n'assure un lien entre les deux éléments régis. Pourtant, peu de choses changent dans les significations et les effets pragmatiques partagées par les deux groupes de rection. Ainsi, que ce soit dans 1a, 1b et 1c, l'un des deux éléments est autonome au niveau

⁹ Le groupe '*pourquoi*' étant un exemple de *recteur averbal*, susceptible par ailleurs d'attirer des éléments régis ('*pourquoi une telle habitude*' → '*pourquoi ça*').

du sens. La partie '*on n'en sait rien*' conserve ainsi l'essentiel du sens si on lui retire l'élément '*pourquoi*', alors que l'inverse n'est pas vrai et produira d'ailleurs, en contexte d'interaction, l'effet illocutoire d'une question.

Le lien qui unit les deux segments peut être décrit, en termes généraux, comme une relation de *dépendance*. Cette dépendance, toutefois, n'est pas réactionnelle et ne fait pas intervenir les constituants syntaxiques, mais serait matérialisée par des contraintes informationnelles – notamment le contraste – et des phénomènes intonatifs (Delofeu, 1991). Il s'agit donc de dépendances d'un autre ordre, qui, pour l'*Approche pronominale*, possèdent une régularité syntagmatique; de nature grammaticale, elles constituent ce que l'*Approche pronominale* qualifiera de *macrosyntaxe*.

(3) LES LIENS MACROSYNTAXIQUES

Plusieurs théories ont voulu plus spécifiquement décrire ces relations comme des liens de nature discursive; c'est le cas, par exemple, de la notion de *togetherness* (Bolinger, 1968). Plus généralement, les concepts « parapluie » de *cohérence* et de *cohésion*, largement utilisés en linguistique du texte, recouvrent en général de tels liens d'interdépendance entre éléments et permettraient d'expliquer les relations observées dans les exemples ci-dessus. Charolles (1995:3-4) distingue par exemple dans ce champ hétérogène situé « hors-phrase »¹⁰ deux catégories de phénomènes. D'une part, des plans d'organisation du discours qui consistent en des « systèmes de marques » (connecteurs, anaphores, cadres de discours, configurations textuelles) et d'autre part des éléments qui « [s'expliquent], au moins partiellement, par des considérations syntaxiques. »

¹⁰ Charolles (1995) défend une démarcation claire entre d'un côté les phrases, qu'il définit à partir de la rection syntaxique, et de l'autre le discours, qui rassemble tous les phénomènes externes à la syntaxe de rection: « Sorti de la phrase il n'y a plus [...] de cadre préconfigurant la distribution des unités verbales, il n'existe pas de structure formelle dans laquelle les unités phrastiques devraient rentrer pour occuper une position prédéfinie. » (p. 3)

L'*approche pronominale* se situe en porte-à-faux de cette conception usuelle des rapports entre syntaxe et discours :

Nous pensons que ces dépendances ont un fondement grammatical, et qu'on peut décrire *grammaticalement* les relations de la [macrosyntaxe], en utilisant d'autres moyens d'analyse que ceux de la rection des catégories. (Blanche-Benveniste, 1990:114) [je souligne]

Pour le GARS, s'il existe une limite à la grammaire des catégories et à la syntaxe de rection, un certain nombre de rapports non-rectionnels fonctionnent néanmoins de manière syntagmatique et peuvent donc faire l'objet d'une description grammaticale.

Contrairement à la *Grammaire de la période* ou au modèle genevois, il ne s'agit pas de substituer à un modèle de cohérence des textes une macrosyntaxe qui ambitionnerait d'analyser l'ensemble des structures textuelles : la macrosyntaxe proposée par le GARS reste un phénomène local qui décrit les relations unissant des groupes de rection. L'unité structurant la macrosyntaxe du GARS est ainsi l'*énoncé*, un ensemble limité dans son extension par un nombre réduit de relations entre groupes de rection.

(4) STRUCTURE DES ÉNONCÉS

Dans la grammaire du GARS, l'*énoncé* est considéré comme l'unité structurante – mais non maximale¹¹ – de la macrosyntaxe. Il est considéré comme un assemblage « de syntaxe, de prosodie, de sémantique et de pragmatique, ainsi que tout un ensemble de routines de discours, certaines étant signalées par des marqueurs morphologiques et d'autres non. » (Blanche-Benveniste & Martin, 2010:159) Au-delà de son aspect composite, l'énoncé est défini par sa propriété de contenir un et un seul *noyau*.

Le *noyau* est une « séquence autonome, caractérisée par un contour mélodique qui en marque la fin, autour de laquelle peuvent se disposer différents éléments. » Il est important de noter que le *noyau n'est pas défini par des caractéristiques grammaticales* ; la notion de rection et la

¹¹ Le GARS ne propose pas d'unité maximale de la macrosyntaxe, pas plus qu'il ne théorise le lien entre macrosyntaxe et structures discursives. En comparaison à la grammaire de la période, la grammaire du GARS se tient de fait du côté de l'*énoncé* en évitant de s'aventurer dans le domaine de l'*énonciation*.

notion de noyau sont disjointes et montrent une séparation de niveau linguistique entre microsyntaxe et macrosyntaxe. D'un côté, les éléments recteurs ne sont pas nécessairement des noyaux. Un élément recteur, s'il correspond souvent, en macrosyntaxe, à un noyau, peut ainsi également se trouver en situation de dépendance vis-à-vis d'un autre constituant. Ainsi, dans l'exemple qui suit, aucun des deux groupes de rection ne peut être réalisé indépendamment de l'autre et ne répond donc au critère d'autonomie. La position du GARS face à ce type de formes sera de considérer que les deux groupes de rection forment un noyau unique (Blanche-Benveniste, 1990:115) :

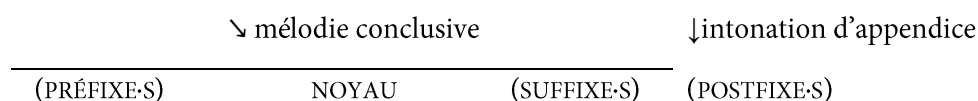
(5) [il y en a qui fument un stick] [d'autres qui roulent comme des tarés] (BlaJD11)

Inversement, un noyau ne correspond pas nécessairement à un groupe de rection ; il peut être constitué de toutes les catégories grammaticales dès le moment qu'elles peuvent constituer un énoncé autonome. Ainsi, des segments comme 'oui', 'non', 'pas question', mais aussi 'au secours', 'café-restaurant' (sur une enseigne) ou encore 'interdit de stationner' (sur un panneau indicateur) seront considérés comme des noyaux, quand bien même ils ne constituent pas des domaines régis.

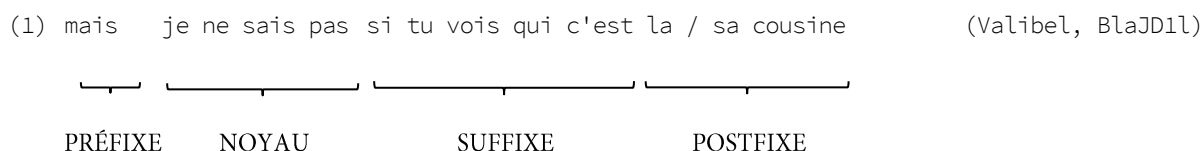
Autour du noyau se structurent des *constituants macrosyntaxiques* qui ont la triple propriété d'être non-autonomes (dépendants), non-obligatoires (optionnels) et éventuellement réitérables dans un même énoncé. De par ces propriétés de dépendance et d'optionnalité, ces dépendants macrosyntaxiques ont été dénommés par le GARS à l'image des structures de dérivation morphologique. Ils sont ainsi de trois types, correspondant chacun à différentes propriétés syntaxiques et différents contours intonatifs : les *préfixes*, qui sont des dépendants précédant le noyau, syntaxiquement ou intonativement marqués comme tels ; les *suffixes*, qui sont des dépendants succédant au noyau et marqués syntaxiquement comme tels ; et enfin les *postfixes* qui sont dépendants du noyau mais possèdent une autonomie syntaxique vis-à-vis de celui-ci¹².

¹² Ils peuvent néanmoins être *associés* à la rection (Blanche-Benveniste, 1990, p. 141), par une relation qui en conditionnera le rattachement syntaxique.

La structure d'un énoncé peut donc se représenter comme suit :



Ce qui produira, dans le cas d'un énoncé comprenant le noyau et ses trois types de dépendants, une analyse macrosyntaxique de ce type :



1.3. COMPARAISON DES APPROCHES GRAMMATICALES

La grammaire élaborée par le GARS est similaire, dans son approche, à la *Grammaire de la période*. Les deux courants adoptent une démarche inductive pour l'élaboration de leur grammaire, en se fondant sur des données tirées de l'usage et en « se méfiant des données fondées seulement sur l'intuition » (Blanche-Benveniste 2012:338). Une grande importance est conférée aux données orales, pourtant plus difficiles à obtenir. Contrairement à une conception courante à l'époque, tous deux rejettent une opposition axiologique entre oral en écrit, soucieux d'élaborer des modèles généralisables aux deux supports d'expression. Également dans une perspective inductive de la grammaire, les deux approches évitent de parler de « structures canoniques » de grammaire (Blanche-Benveniste, 2012:339).

Un point commun important entre les deux théories est la critique – et la réfutation – de la notion de phrase, considérée comme inadéquate en tant qu'unité pour analyser les constructions syntaxiques et leur articulation dans le discours et à travers l'usage. La notion de « rection » est préférée pour décrire la maximalité des relations (micro-)syntaxiques, tandis que les aspects prosodiques et sémantiques qui caractérisaient le tout formé par la phrase canonique sont expliqués dans le cadre du module macrosyntaxique de la grammaire.

Les deux modèles grammaticaux se distinguent malgré tout sur de nombreux aspects. La grammaire du GARS se concentre principalement sur la description de la microsyntaxe. La

façon dont le GARS entend construire un nouvel outil grammatical est en effet *analytique* : il s'agit de partir des petits éléments pour ensuite espérer formuler des hypothèses valides sur les « grandes unités ». La *Grammaire de la période*, en comparaison, observe surtout des phénomènes plus « amples » (Blanche-Benveniste, 2012:341). Elle marque une frontière nette entre une microsyntaxe, de préoccupation plus secondaire, et une macrosyntaxe qui est considérée comme une « troisième articulation du langage », sans « interférence » entre macro- et microsyntaxe (Blanche-Benveniste, 2012:345).

Les deux grammaires se distinguent par leur approche des règles et des modélisations syntaxiques. La *Grammaire de la période* fonde son discours sur une évaluation de la complétude des unités, que ce soit au niveau de la forme ou de la fonction de celles-ci. Dès lors, une question centrale des modèles est la maximalité de chaque unité et les démarcations qui peuvent en être établies. Dans le modèle du GARS, c'est l'axe dépendance-autonomie des unités qui tend à guider les modélisations syntaxiques : cette observation est variable au niveau microsyntaxique pour le phénomène de rection, et au niveau macrosyntaxique pour le concept de noyau : dans les deux cas, il s'agit d'une unité centrale organisant autour d'elle, à divers degrés, des éléments dépendants. À cet égard, une observation importante est que le modèle du GARS ne considère pas une unité maximale de macrosyntaxe : ce qui se rapporte, syntaxiquement, à la notion d'énoncé est l'agencement articulé de plusieurs unités. En comparaison à la *Grammaire de la période*, la grammaire du GARS se tient donc du côté de l'énoncé, en évitant de s'aventurer dans le domaine de l'énonciation. Si la *Grammaire de la période* se risque dans une analyse des actes de langage et de leur enchaînement pour former des discours, l'aspect décontextualisé de l'analyse du GARS l'oblige à se tenir aux frontières de la grammaticalité.

La prise en compte de paramètres discursifs n'est cependant pas exclue du modèle du GARS. Ainsi, la mémoire discursive et son évolution d'énoncé en énoncé est un outil très intéressant qui pourrait venir compléter l'*Approche pronominale*, comme le propose elle-même Claire Blanche-Benveniste (2012:349-350). J'examinerai la pertinence de cette modélisation dans l'étude des périphéries droites.

2. LA PÉRIPHÉRIE DROITE

Dans la partie précédente, j'ai présenté plusieurs outils d'observation de la langue en usage. L'orientation vers une approche grammaticale orientée vers la macrosyntaxe, plutôt qu'une approche orientée vers les structures discursives, s'est présentée comme une nécessité au vu des résultats recherchés.

À partir de ce cadre théorique désormais établi, je définis la périphérie droite en deux temps ; dans un premier temps, je m'intéresse à la notion de « périphérie » pour ensuite, dans un second temps, définir plus spécifiquement celle de « périphérie droite ». Dans la mesure où bien des études abordant la question des périphéries n'ont pas été établies dans une perspective macrosyntaxique, j'aborde à chaque étape la perspective des approches utilisant une syntaxe que je qualifie de « rectionnelle »¹³ – sans théorie macrosyntaxique – avant de m'intéresser de plus près à une définition macrosyntaxique en en précisant les avantages descriptifs.

2.1. LA NOTION DE PÉRIPHÉRIE

Cette partie pose, pour ce mémoire, la définition que je retiendrai de « périphérie » et ensuite celle de « périphérie droite ». Elle contraste les définitions qui peuvent ressortir des approches « traditionnelles » de la syntaxe, dont le modèle grammatical repose sur un niveau d'analyse unique où sont considérés les liens de rection et les relations entre constituants, et les approches « macrosyntaxiques » décrites plus haut, qui différencient quant à elles deux niveaux d'analyse, l'un faisant intervenir les liens de rection, l'autre se basant sur la pragmatique et l'intonation pour la description grammaticale.

Les approches dites « traditionnelles » esquissent une définition générale des « périphéries », gauches ou droites, en les situant dans un cadre terminologique largement débattu, mais

¹³ Ces approches correspondent à des courants très différents : fonctionnalisme, générativisme, distributionnalisme, etc. Elles présentent néanmoins le point commun de ne pas définir de macrosyntaxe : la syntaxe de rection, faisant intervenir des constituants et des dépendances, est donc généralisée à la description de toutes les structures syntagmatiques du langage.

également largement partagé par la communauté linguistique. La périphérie est ainsi définie en relation à la notion de *phrase*, que j'ai problématisée dans la *section 1.1.1.* et ensuite relativement aux *propositions* et aux *énoncés* dans l'acception usuelle de ces termes. Les approches macrosyntaxiques permettent d'affiner la définition de périphérie ; basant leurs modèles sur l'intonation, elles permettent d'aborder les segments postfiniaux comme des éléments effectivement périphériques, dans leur spécificité formelle.

2.1.1. APPROCHES TRADITIONNELLES DE LA SYNTAXE

(1) PÉRIPHÉRIE DES PHRASES

Dans le cadre de grammaires déductives, tant d'une perspective fonctionnaliste que par des approches générativistes, les phrases sont perçues comme des entités dont la délimitation n'est pas équivoque. Inclure la prise en compte de périphéries est ainsi lié à la préoccupation d'une meilleure description de données en marge du modèle. C'est ainsi que la « périphérie des phrases » est généralement envisagée dans le cadre de constructions particulières dans lesquelles des constituants se trouvent « rejetés » hors de la structure canonique d'organisation des constituants.

Des phénomènes tels que la *focalisation*, la *topicalisation*, les *constructions clivées*, etc. ont permis aux grammaires fonctionnelles de montrer l'influence de la structure informationnelle dans la construction syntaxique des énoncés. Les générativistes, de leur côté, ont cherché à comprendre quelles transformations conduisaient à dériver de telles structures : dislocation, grammaticalisation des marqueurs, etc.

D'une manière générale, fonctionnalistes comme générativistes se sont accordés à expliquer le rôle des périphéries par les nécessités de l'organisation informationnelle (Beeching & Detges, 2014:2) : une structure considérée comme initiale ou du moins « non-marquée » se trouve *disloquée* pour ré-agencer les éléments informatifs – topic et focus, thème et rhème – ou pour articuler la coréférence, ou *grammaticalisée* pour agir au niveau du discours. Dans ce contexte, la notion de périphérie désigne ainsi l'ensemble des paradigmes marginaux de la structure

phrastique, corrélés à des fonctions discursives ou pragmatiques, dans lesquels surviennent les éléments disloqués.

(2) PÉRIPHÉRIE DES ÉNONCÉS / DES PROPOSITIONS

Comme je l'ai déjà soulevé, la syntaxe des constituants et la notion de phrase en particulier possèdent leurs limites (cf. 1.1. *Limites de la phrase*). En considérant les périphéries dans leur caractère discursif – et éventuellement conversationnel ancré dans une énonciation – on pourrait également envisager celles-ci comme attachées aux énoncés, ou encore aux tours de parole. La notion de tour de parole peut se comprendre comme « le début et la fin de la phonation d'un locuteur » (Beeching & Detges, 2014:3), tandis qu'un énoncé, à assimiler ici à ce que l'analyse conversationnelle a pris l'habitude de nommer *unité de construction de tour* (UCT) (Sacks, Schegloff, & Jefferson, 1974) constitue un sous-ensemble d'un tour de parole qui représente une complétude interactionnelle ou discursive ; le lieu où un interlocuteur peut prendre la parole sans donner l'impression d'interrompre le locuteur précédent. Un postulat concernant cette complétude est que le locuteur, en produisant cette unité, réalise une unité intonative autonome, ou une proposition autonome correspondant à un sens achevé (Delais-Roussarie & Post, 2008).

Considérer les périphéries à partir du niveau énonciatif ou propositionnel est relativement rare dans la littérature. Tout au plus, certains auteurs considèrent-ils spécifiquement comme éléments périphériques des structures renvoyant à l'énoncé précédent ou suivant, et indexant des modalités vis-à-vis de celui-ci (Beeching & Detges, 2014:4). Traugott (2012), par exemple, identifie des segments qui justifient leur présence par un rôle pragmatique reporté sur une proposition¹⁴ adjacente. Certains segments pourtant syntaxiquement autonomes peuvent hériter de ce trait et devenir de ce fait *pragmatiquement dépendants* : en anglais, c'est le cas de formules interrogatives comme '*isn't it*', '*you know*' ou encore '*no*'. Elle conclut que ces

¹⁴ Traugott assimile les propositions à des prédications qui prennent, syntaxiquement, la forme d'une clause régie : « LP and RP are understood to be external to the proposition, i.e. to the clausal predicate-argument structure » (Traugott, 2012, p. 14)

segments, selon qu'ils sont reportés à gauche ou à droite des énoncés, tendent, respectivement à assigner à ceux-ci une modalité plus *subjective* ou plus *intersubjective*.

2.1.2. LES APPROCHES MACROSYNTAXIQUES

Lorsque, comme dans le cadre des approches macrosyntaxiques, la syntaxe admet deux niveaux séparés d'analyse, la première question qui se pose est celle du niveau pertinent pour décrire les périphéries. Trois définitions sont théoriquement possibles : à partir de la microsyntaxe, à partir de la macrosyntaxe et au croisement des deux modules. J'expose ici chacune des trois perspectives, ses avantages et désavantages en regard du GARS et de la *Grammaire de la période*.

(1) LA PÉRIPHÉRIE EN MICROSYNTAXE

La périphérie, en microsyntaxe, est identifiée à partir des caractéristiques formelles de la rection. Cette approche repose sur l'identification d'« ilots de rection » : certains segments se trouvent dans la rection, tandis que d'autres se trouvent en-dehors de celle-ci. Les périphéries peuvent alors être définies de différentes manières : dans la *Grammaire de la période*, les périphéries sont extérieures aux clauses ; dans l'*Approche pronominale*, les éléments périphériques regroupent tous les éléments hors de la syntaxe de rection, c'est-à-dire tous ceux qui ne peuvent être ramenés à une proforme pronominale.

Intéressons-nous désormais plus précisément au phénomène des « segments postfiniaux ». L'*Approche pronominale* développée par le GARS porte son attention sur les tests permettant d'évaluer les limites de la rection. Comme je l'ai développé plus haut, ces tests reposent sur les possibilités de substitution par une proforme pronominale. Si celle-ci n'est pas une forme clitique, mais bien tonique¹⁵, ou si la forme, malgré sa relation à la rection, n'est pas substituable à une proforme régie, on aura affaire à un associé lexical. Dans le cadre de la grammaire du GARS, l'étude des périphéries en microsyntaxe est bien définie par le concept

¹⁵ Classe revue et étendue dans le cadre l'approche pronominale, où elle a été rebaptisée « élément semi-lexical ».

d'« associé ». Blanche-Benveniste, comme Berrendonner (1990:80-81) distinguent deux catégories d'associés. Dans le premier cas de figure, (1) les « associés avec double marquage » sont représentés, dans la rection, par une proforme régie. Dans le second cas de figure, (2) « on est livré au simple jeu des associations lexicales » (1990:82). L'inclusion est alors sémantique et constitue un lien qui peut aller d'un lien très précis – l'hypothèse d'une simple élision de la forme pronominale du cas 1 – ou très vague et tributaire des conditions d'énonciation (1990:83).

La *Grammaire de la période*, contrairement à l'*Approche pronominale* avec le terme d'« associé », ne construit pas de terme identifiable à des « segments postfiniaux » en microsyntaxe. En macrosyntaxe, en revanche, un lien est possible entre une clause indépendante, sans rection, et une clause régie. Ce lien est exprimé sous la forme de la routine [*action* ▷ *confirmation*] sans envisager une périphéricité du phénomène¹⁶. L'*Approche pronominale*, qui propose la catégorie des *associés*, présente une terminologie et une méthode relativement opératoire d'identification des périphéries. Le concept central est, pour le GARS, celui des rapports de rection : les éléments périphériques peuvent se définir en relation avec les « éléments centraux » que constituent les domaines régis. Il est plus difficile, en revanche, de fixer les bornes extérieures des périphéries ; cela implique de délimiter celles-ci à partir des éléments admis dans leur structure interne, avec un risque de circularité entre l'identification des éléments et la caractérisation de la structure qu'ils représentent.

Qu'il s'agisse de l'*Approche pronominale* du GARS ou de la *Grammaire de la période*, les structures microsyntaxiques s'avèrent intéressantes pour envisager les phénomènes de rection.

¹⁶ Les dislocations à droite, en tant que cas particulier de segments postfiniaux, sont cependant l'objet d'une description intéressante de la part de Berrendonner (2012, pp. 235-252) qui en différencie trois types distincts. La dislocation à droite *de type I* est constituée de deux clauses microsyntaxiquement connexes dont l'une des deux est une *extraposition*, et porte les mêmes marques syntaxiques. Le *type II* concerne deux clauses sans rattachement syntaxique ('*nominativus pendens*') qui seront considérées comme autonomes mais où est redésigné un objet de discours (au niveau macrosyntaxique). Enfin, les dislocations *de type III* concernent les *clauses méta-énonciatives*, portant rétrospectivement sur la précédente. Comme on le voit, seul le premier type de conjonction peut se décrire au travers de la microsyntaxe.

Cette dernière, dans la *Grammaire de la période*, est probablement trop largement définie pour rendre compte de « périphéries » : bien souvent, on aura deux clauses quand on aura un seul groupe régi pour le GARS. D'un point de vue syntaxique, c'est donc l'approche du GARS qui devra être privilégiée dans le cadre de ce mémoire.

(2) LA PÉRIPHÉRIE EN MACROSYNTAXE

En macrosyntaxe, la périphérie est identifiée par des faits prosodiques et au niveau pragmatique. Cette approche est orientée vers les *fonctions* des périphéries. La prosodie est également considérée, mais comme un système séparé des dépendances syntaxiques.

Dans la *Grammaire de la période*, un grand nombre d'éléments que la microsyntaxe considère comme « disloqués » résultent d'une action communicative¹⁷ qui s'est muée dans le langage en « routine macrosyntaxique » de type [*action* ▷ *confirmation*]. Non liée aux formes linguistiques, cette routine se retrouve également dans d'autres formes structures que celles des dislocations¹⁸. La routine [*action* ▷ *confirmation*] correspond à un procédé cognitif et informationnel fondé sur une récurrence d'actions communicatives ne sont pas toujours présentées selon une structure *déductive* mais parfois *abductive* (Berendonner, 2012:240-245). Elle suggère qu'un « calcul probabiliste » est opéré à la réception à partir d'un ou de plusieurs éléments, qui permet d'inférer une confirmation d'un « objet-de-discours » ou un commentaire méta-énonciatif à valeur rétrospective (Berendonner, 2012:238). La routine macrosyntaxique [*action* ▷ *confirmation*] est représentée au niveau de l'intonation par le prototype de l'*intonème réduit* (cf. 3.1.2. *Schémes atypiques*).

Les éléments macrosyntaxiques du GARS n'ont aucun lien avec les catégories grammaticales. En conséquence, les définitions traditionnelles de la périphérie, reposant sur la syntaxe des

¹⁷ Pour Berendonner (2003, p. 98), « la notion d'action communicative sert à figurer des opérations implicites, de niveau cognitif, et non des actes qui seraient inscrits dans le signifié des clauses et typés en langue, voire codifiés dans le lexique. En d'autres termes, elle décrit ce que l'énonciation fait et non ce que la clause énoncée dit qu'elle dit qu'elle fait. [...], elle permet en outre de figurer des opérations qui ne s'appliquent pas à un contenu propositionnel. »

¹⁸ Catégories non-lexicales telles que les recteurs faibles, notamment, ou encore des structures que je qualifie ci-après de *verbes défocalisés*.

constituants, sont difficiles à concilier avec la structure des énoncés macrosyntaxiques, telle que proposée par le GARS. Dans ce modèle, en effet, tous les *affixes* gravitent autour d'un *noyau* et pourraient donc être considérés comme des périphéries de celui-ci, dans une compréhension extensive de la notion. Considérer les périphéries droites permet d'exclure *a priori* les préfixes, situés à gauche du noyau. Les suffixes, dotés d'un contour mélodique riche qui marque leur subordination directe au noyau, et fréquemment introduits par un suffixeur ('*car*', '*tandis que*', '*de telle sorte que*', '*puisque*') ne correspondent pas non plus à la définition générale des périphéries que j'ai esquissée précédemment et coïncident très souvent avec l'occurrence de deux groupes de rection distincts. Dès lors, une périphérie droite peut être caractérisée en macrosyntaxe dans le modèle du GARS à partir du regroupement entre un *noyau* et un *postfixe* pour former une unité de macrosyntaxe. Cette association se traduit par une « intonation remarquable » (Blanche-Benvenise, 1990:140) qui est celle de l'*appendice* (cf. 3.1.2. *Le schème prototypique*).

Tant dans l'approche fribourgeoise que dans celle du GARS, une définition de la périphérie droite par les faits de macrosyntaxe est donc envisageable. Elle a l'avantage de pouvoir rendre compte de nombreux phénomènes discursifs en les relatant à des données factuelles : les schèmes intonatifs. Ce point fort, cependant, est également un point faible pour l'élaboration d'une définition opérationnelle : si l'intonation est le seul indice fiable, il devient difficile d'observer des régularités dans les faits de nature verbale tels que la syntaxe de rection et l'agencement des constituants.

(3) LA PÉRIPHÉRIE ENTRE MICRO- ET MACROSYNTAXE

Au-delà d'une définition purement micro- ou macrosyntaxique, une caractérisation de la périphérie comme un phénomène intermédiaire, en marge des deux modules grammaticaux, se présente comme une troisième voie. Dans ce cadre, la définition devient hybride : elle reprend en partie des faits de microsyntaxe, et en partie des faits de macrosyntaxe. La *Grammaire de la période*, qui conserve une séparation stricte entre les deux niveaux d'analyse, conviendrait difficilement à cette « hybridation » conceptuelle. Le modèle du GARS envisage

en revanche une continuité entre les niveaux micro- et macrosyntaxiques dans laquelle la macrosyntaxe commence là où la syntaxe de rection s'arrête : cette conception théorique pourra davantage convenir à une définition mixte. Blanche-Benveniste identifie ainsi des structures morphosyntaxiques corrélées aux postfixes, dont elle dégage cinq « classes » : [I] le groupement « verbe + associé » ; [II] la suite « rection + verbe » ; [III] l'association de deux éléments verbaux ; [IV] le « type "dit-il" » et [V] le « type "bien sûr" » (Blanche-Benveniste, 1990:141-143).

Le problème d'une définition fondée sur des corrélats est qu'elle ne capte pas l'objet considéré dans sa régularité et sa systématité. De fait, certains phénomènes qui correspondent bel et bien à une intonation postfixe ou à une routine [*action* ▷ *confirmation*] échapperont pourtant aux domaines d'inclusion de cette définition.

1.1. DÉFINITION PROPOSÉE

Suite à ce que j'ai exposé dans les parties précédentes, il ressort que les périphéries se définissent en *essence* par des caractéristiques macrosyntaxiques et des caractéristiques intonatives. Formellement, le GARS identifie ainsi les périphéries droites aux structures *postfixes*, tandis que la *Grammaire de la période* les incorporerait à la routine [*action* ▷ *confirmation*]. En termes d'intonation, ces structures se traduisent dans la *Grammaire de la période* l'intonème « réduit », là où le GARS décrit une intonation d'« appendice » pour manifester, à l'oral, les structures postfixes (cf. 3.1. *Le schème prototypique*). Les deux approches soulignent que ce paradigme de macrosyntaxe n'est pas lié aux structures microsyntaxiques ni à la rection.

Cette définition par la macrosyntaxe est difficilement praticable sans la prise en compte de faits de *surface* qui corroborent ces structures, et sont mieux caractérisés à partir du module descriptif de la microsyntaxe. Par exemple, les phénomènes de dislocation relevés par les approches traditionnelles peuvent être évalués à partir des phénomènes de rection ; dans ce but précis de descriptions, la terminologie des *associés* proposée par le GARS se présente comme la plus opératoire. D'autres formes recensées par Blanche-Benveniste (1990:141-143)

comme typiquement postfixes ont largement été observés dans d'autres cadres théoriques : les dislocations et « pseudo-dislocations » à droite ; les focalisations d'éléments averbaux ; les marqueurs de discours placé en finale d'énoncé ; les verbes « parenthétiques » et les « recteurs faibles ».

Comme je l'ai évoqué dans la sous-partie précédente, une définition opératoire qui, empiriquement, repose sur des formes microsyntaxiques pour caractériser des phénomènes macrosyntaxiques, passera nécessairement à côté d'une partie de son objet. La seconde partie de ce mémoire, par laquelle seront annotés des extraits de corpus, aura donc le double objectif de valider cette définition et de rendre compte de ces éléments qui, actuellement, échappent toujours aux radars de la théorisation.

3. CARACTÉRISTIQUES DES PÉRIPHÉRIES DROITES

À partir de la définition proposée, cette section passe en revue les observations et formalisations d'éléments caractérisables comme des « périphéries droites ». L'objectif est de donner le meilleur aperçu des structures corrélées à la notion, désormais définie, de périphérie droite. Deux types de structures sont pertinents pour la définition que j'ai donnée : d'une part, les structures de microsyntaxe, liées aux phénomènes de rection et corrélées, en macrosyntaxe, à l'occurrence de *postfixes* (Blanche-Benveniste, 1990:140-144); d'autre part, les schèmes intonatifs particuliers qui indexent les structures macrosyntaxiques.

3.1. CORRÉLATS MICROSYNTAXIQUES

Dans la section précédente, j'ai proposé, à partir des observations de Blanche-Benveniste (1990:140-144), une définition opératoire de la périphérie droite qui repose sur l'identification de structures caractéristiques de microsyntaxe. La structure pragmatique n'étant pas accessible, identifier les formes ne pourra se faire qu'à partir de corrélats syntaxiques. À partir des traitements divers qu'en a effectués la littérature à ce jour, je distingue, à la suite de Blanche-Benveniste (1990:140-144) quatre familles de faits linguistiques apparentés à des structures postfixes. D'une part, les structures associées à la rection que sont les associés lexicaux et les marqueurs discursifs et pragmatiques ; d'autre part, les groupes de rection subordonnés par un noyau averbal qui sont de deux types : les recteurs faibles et les structures verbales défocalisées.

3.1.1. ASSOCIÉS À LA RECTION

À partir de la grammaire du GARS, une première catégorie d'éléments postfixes se dégage, qui correspond aux associés de « type 'bien sûr' » et aux « associés lexicaux ». Au niveau microsyntaxique, ces éléments présentent en effet la caractéristique de compter parmi les éléments *associés* à la rection – et donc d'être « périphériques » également du point de vue de la microsyntaxe. Blanche-Benveniste (1990:77) les définit comme suit :

Nous appelons « associés » les éléments qui ont parfois l'apparence de compléments régis par le verbe, mais qui n'en sont pas. Ils n'ont aucune propriété des éléments régis ; on les définit, en premier lieu, par l'absence de ces propriétés.

Si les associés ne sont pas membres de la rection, ils entretiennent néanmoins un lien avec le terme recteur – que le GARS décrit grâce aux relations macrosyntaxiques. La famille des associés est très vaste : elle regroupe des éléments lexicaux, des syntagmes introduits par des conjonctions, ou encore des marqueurs discursifs et pragmatiques. Les associés qui nous intéressent ne sont qu'un sous-ensemble de cette vaste catégorie, dont la particularité est de pouvoir survenir en position postfixe.

(1) DISLOCATIONS À DROITE ET ASSOCIÉS LEXICAUX

La dislocation a été largement étudiée, notamment par la grammaire fonctionnelle et par la grammaire générative qui ont cherché à expliquer le procédé syntaxique en jeu (Ashby, 1988 ; Nølke, 1996 ; De Cat, 2007) et ses motivations discursives et pragmatiques. Le phénomène est un procédé syntaxique impliquant « le détachement d'un constituant de la phrase à son début ou à sa fin » (Nølke, 1996:385). Le terme comprend des phénomènes apparentés, mais néanmoins très variés, comme le montre Berrendonner (2012:236) au travers plusieurs exemples où les structures disloquées à droite apparaissent en italique¹⁹. Trois groupes peuvent se distinguer :

(1a) vous vous en amusez hein *des chapeaux*

(1b) bon je vais lui téléphoner *à ce type*

(2a) est-ce que vous êtes content *de la marionnette*

(2b) à Marseille nous avons / *du soleil*

(3a) je mets huit heures de cours le samedi et à la fin on va boire un verre *ceux qui restent*

(3b) il est incroyable *cette bête*

(Berrendonner, 2012:236)

¹⁹ Exemples adaptés à la notation Valibel.

La littérature différencie généralement deux dislocations^{20, 21}: (1a et 1b) une *dislocation au sens strict* ou « double marquage » dans laquelle le constituant est préalablement instancié par une proforme pronominale et ensuite sous la forme d'un associé lexical, et (2a et 2b) une *pseudo-dislocation* dans laquelle le constituant n'est pas anticipé par un pronom, mais possède néanmoins un rôle syntaxique attribuable. À ces deux dislocations doit être associé un troisième phénomène connexe : (3a et 3b) les *associés libres* auxquels ne peut pas être attribué un rôle syntaxique, mais qui modifient néanmoins le sens propositionnel du groupe de rection qui précède.

Pour le GARS, c'est le caractère d'associé qui caractérise le phénomène de dislocation. Un point important est la proportionnalité de l'élément disloqué non pas à un pronom clitique – régi –, mais bien à un pronom tonique – non régi. Les pronoms toniques sont, dans la théorie, qualifiés de « quasi-lexicaux » : ils interviennent de la même façon que tout élément lexical sur une structure régie. L'analyse de leur comportement d'associé permet des inférences intéressantes sur les fonctionnements et les raisons d'être de la dislocation à droite et sa distinction de la dislocation à gauche.

Par l'analyse de l'opposition clitique-tonique et par le contraste de leur emploi entre périphéries gauche et droite, Detges & Waltereit (2014)²² montrent que la périphérie droite des énoncés remplit essentiellement deux rôles : d'une part, un rôle, relativement peu courant dans le discours, de maintien de la cohérence (usage qu'ils qualifient d'*anti-topic*), et d'autre part, le rôle de l'expression d'une attitude du locuteur, autrement dit un rôle *modal* : fonction émotive, mitigation de point de vue, etc. Horlacher et Müller (2005) concluent également de

²⁰ La même distinction existe pour les dislocations à gauche.

²¹ Nølle (1996) présente un possible troisième groupe de dislocations, rarement attesté par ailleurs. Introduite par « que » ou « comme », elles ne sont pas substituables par des pronoms sans perdre leur détermination « indéfinie » :

C'est une belle fleur *que la rose* / *une rose* est une belle fleur / **elle* est une belle fleur

Elle est charmante *comme fille* / elle est *une fille* charmante / **elle* est charmante

²² À noter que Detges & Waltereit (2014) reprennent également, dans la même recherche, une démonstration de la distribution complémentaire des pronoms clitiques vs. toniques entre le groupe de rection et les éléments périphériques.

leurs analyses que le statut informationnel de la dislocation à droite est surestimé, alors que le rôle de celles-ci dans la gestion des topics conversationnels ou l'expression du désaccord est largement sous-estimée. L'analyse des données d'annotation sera l'occasion d'approfondir ces fonctions jouées par les segments disloqués.

(2) MARQUEURS DISCURSIFS ET PRAGMATIQUES

Le type « bien sûr » est une catégorie de postfixes proposée de façon succincte par Blanche-Benveniste (1990:143-144) qui se base sur le marqueur « bien sûr » (« je viendrai / **bien sûr** ») et son paradigme, mais ne qualifie pas autrement cette catégorie qu'en parlant tout simplement du « type "bien sûr" », sans le détailler²³.

À ce paradigme peuvent être intégrés de nombreux éléments correspondant à différentes catégories de la microsyntaxe (adverbes, prépositions, interjections mais aussi noms, verbes, etc.) que la littérature qualifie de diverses manières : marqueurs de discours, marqueurs pragmatiques, ponctuant,... Leur principale caractéristique commune est somme toute d'être « pragmatiques » pour intervenir au niveau du discours ou de l'interaction.

Les *marqueurs de discours* sont des formes linguistiques susceptibles d'émaner de toutes les catégories grammaticales, et spécialisées pour véhiculer un sens principalement discursif et/ou pragmatique ; ils interviennent peu dans la sémantique des propositions et dans la syntaxe de rection, et sont prosodiquement indépendants. Schourup (1999) insiste sur le rôle de connexion qui caractérise toujours les marqueurs de discours, en précisant que cette connexion peut opérer dans « un contexte plus large » couvrant de multiples « plans du discours ».

De la catégorie des marqueurs de discours proprement dits sont parfois distingués les *marqueurs pragmatiques* (MPs) qui s'en différencient par trois caractéristiques principales (Aijmer & Simon-Vandenberghe, 2008) : (1) l'indexicalité – le fait de pointer vers

²³ Du point de vue microsyntaxique, le type « bien sûr » correspond pour le GARS à une sous-catégorie du large ensemble des *associés* (Blanche-Benveniste, 1990, p. 77), tout comme les phénomènes de dislocation.

des éléments du contexte ; (2) la réflexivité – la capacité de jouer un rôle métapragmatique de commentaire de discours et de servir de support à l'interprétation ; (3) l'hétéroglossie – le fait de marquer explicitement un point de vue ou une posture énonciative.

Si l'étude de marqueurs en périphérie gauche a été l'objet de nombreuses publications, la périphérie droite a jusqu'à ce jour été relativement délaissée en matière d'étude de marqueurs de discours. On peut citer, pour l'anglais, les recherches de Haselow (2011) pour le marqueur '*then*' en fin d'énoncé, ou Kim & Jahnke (2011) pour le marqueur '*even*' dans le même paradigme ; pour le français, Degand & Fagard (2011) et Degand (2014) contrastent les usages de plusieurs marqueurs en initiale et en finale d'énoncés et de tours de parole. Degand (2014) montre que les marqueurs discursifs '*alors*' et '*donc*', en périphérie droite, tendent à remplir une fonction interpersonnelle ou modale, plutôt que structurante de l'information comme cela en est la tendance en périphérie droite. Il est en outre plus courant que ces deux marqueurs y jouent un rôle davantage intersubjectif.

3.1.2. ÉLÉMENTS RECTEURS

En macrosyntaxe, les groupes régis ne sont pas systématiquement des noyaux : ils peuvent se retrouver dans tous les composants de la macrosyntaxe, noyaux comme affixes. Le GARS recense deux situations courantes dans lesquelles des groupes régis se trouvent postfixés : le « type *dit-il* » (Blanche-Benvenise, 1990:143) et le type « rection + verbe » (Blanche-Benvenise, 1990:142). La littérature a relativement bien documenté ces deux phénomènes. Dans un souci de généralité, je les rebaptise ici respectivement « recteurs faibles » et « verbes défocalisés ».

(3) RECTEURS FAIBLES

Blanche-Benveniste (1990:143) donne de cette catégorie l'exemple de « *dit-il, croit-il, trouve-t-il, etc.* » mais admet qu'un élargissement au-delà de la catégorie des verbes de « discours direct » est souhaitable. Je propose de généraliser cette catégorie à l'ensemble des « recteurs

faibles », lequel recoupe en partie celui des « parenthétiques » (*parentheticals*)²⁴ dans d'autres traditions de recherche – notamment pour ce qui concerne l'anglais. Le point commun entre ces formes est d'afficher un certain degré de figement grammatical qui réduit les possibilités de manipulations syntaxiques. Ce figement grammatical va de pair avec une spécialisation vers un sens discursif ou pragmatique, plutôt que propositionnel. Le procédé est alors assimilable à celui qui caractérise les marqueurs de discours et les marqueurs pragmatiques, abordés dans la section précédente, de même que le paradigme de leur usage : les emplois sont attestés en périphérie gauche, en périphérie droite et en incise dans un groupe régi.

La palette de verbes potentiellement concernés par la rection faible provient, d'après Apothéloz (2003:242), de deux catégories lexicales : d'une part, (1) certains verbes « désignant une attitude propositionnelle » (*'penser', 'croire'*) et d'autre part (2) des verbes usuellement « introducteurs de discours direct » (*'dire', 'expliquer', 'être là'*). Je propose d'ajouter l'ensemble des verbes renvoyant à une attitude perceptive, épistémique ou cognitive (*'voir', 'savoir'*).

L'*Approche pronominale* offre des perspectives intéressantes pour catégoriser les recteurs faibles. Ainsi, l'« objet » d'un recteur faible ne pourra pas être pronominalisé (Blanche-Benveniste, 1989). L'exemple suivant (tiré d'Andersen 2007) montre l'impossibilité, dans une réponse, de substituer une proforme pronominale à la complétive introduite par *'je crois'*. Cela indique, dans le cadre de l'*Approche pronominale*, qu'on ne se trouve plus dans une relation de rection.

euh je sais je crois j'avais treize ans [...]

(1) — tu crois ? oui je crois /

(2) — *tu le crois ? *oui je le crois

(Andersen, 2007:311)

La non-pronominalisation de la complétive n'est pas la seule indication d'un recteur faible : d'autres particularités morphosyntaxiques sont perceptibles (Apothéloz, 2003:243-249) : les modalités interrogatives ou négatives ne sont plus applicables, ou produisent une intonation ou

²⁴ Une distinction entre les recteurs faibles et les parenthétiques est qu'une intonation accentuée ou noyau reste possible pour ces derniers. Detges & Waltereit donnent par ailleurs des exemples de parenthétiques accueillant eux-mêmes – d'après leur analyse – des périphéries : « moi je sais pas » ou encore « je sais pas moi ».

un sens nécessairement différents ; les clitiques employés dans l'environnement de rection d'un recteur faible sont eux aussi relativement figés ; il en va de même pour le temps grammatical.

(4) VERBES DÉFOCALISÉS

Blanche-Benveniste (1990:142 ; 2010:171) recense un second cas de groupes de rection survenant en position postfixe. Dans ce cas de figure, un élément régi devient noyau, ce qui va projeter le reste de la rection dans une position d'affixe. Plus spécifiquement, cette structure habituellement noyau devient postfixe lorsque l'élément projeté en noyau est par la même occasion mis en évidence par antéposition. Blanche-Benveniste donne ainsi les exemples suivants, où l'élément en italique est prononcé avec une intonation autonome de noyau :

du nom / je me rappelle plus

plus que deux / il t'en reste

que de la tendresse *tu m'inspires*

(Blanche-Benveniste, 1990:142)

(Sabio 2006, cité par Tanguy 2010)

Lorsque ce type de structure binaire survient, c'est désormais l'élément antéposé qui reçoit les modalités syntaxiques, plutôt que le verbe recteur (« *rien que de la tendresse tu m'inspires* » / « *pas que de la tendresse tu m'inspires* » [Tanguy 2006:31]). J'ajouterai que interprétation peut être vérifiée à partir des constructions clivées (« *c'est que de la tendresse que tu m'inspires* »). D'après Tanguy (2010:34-35), ces structures ont pour rôle de reporter la prédication sur le syntagme antéposé. Au niveau de l'organisation informationnelle de l'énoncé, il ne s'agit donc plus d'un procédé de *topicalisation* – auquel cas le syntagme averbal antéposé aurait formé une périphérie gauche – mais bien un procédé de *focalisation* sur le complément.

Ce type de structures à constituant focalisé – ou à verbe défocalisé, selon l'élément sur lequel est portée l'attention descriptive – demeure toutefois syntaxiquement équivoque. Blanche-Benveniste (1990:142) et Sabio (2006) font en effet remarquer que dans tous les cas considérés, ces structures connaissent potentiellement deux intonations : dans un cas, le verbe est focalisé, tandis que dans l'autre, c'est le constituant averbal qui bénéficiera du focus informationnel²⁵ :

du nom / je me rappelle plus vs. *du nom / je me rappelle plus* [l'italique marquant l'intonation noyau]

Dans la perspective de l'étude des périphéries droites, cette ambivalence de structure nécessite de se concentrer sur les données d'intonation, afin d'établir de façon claire l'élément focalisé. Les implications des données prosodiques pour l'annotation seront évoquées juste après.

3.2. SCHÈMES INTONATIFS

3.1.1. LE SCHÈME PROTOTYPIQUE : DÉCROCHAGE BRUTAL ET PLAGE BASSE

Le fait que les segments postfinaux relèvent d'une intonation caractéristique est un élément largement attesté dans la littérature. Un accord se dégage pour les décrire comme des segments en décrochage par rapport à la mélodie du groupe intonatif précédent : le ton est maintenu en bas du spectre, non modulé et non accentué. Quel que soit le modèle théorique considéré, cette caractéristique est reconnue comme un indice relativement univoque des segments postfinaux qui permet de constituer des regroupements intonatifs empiriques.

Plusieurs auteurs identifient les segments postfinaux des énoncés à partir de ce prototype de chute mélodique. Danon-Boileau et al. (1991:111), avec Morel (2007) développent ainsi une

²⁵ Certaines observations invitent à privilégier, dans les cas ambigus, les focalisations verbales. Berrendonner (2012, pp. 256-258) aborde ainsi également ces « verbes défocalisés » en les qualifiant d'« extrapositions frontales ». Il en démontre le caractère marqué à partir des modèles de la GP : les macrosyntagmes [*action* ▷ *confirmation*], qui impliquent une double instanciation des constituants, sont en effet moins adaptés à l'introduction de constituants de rang *P* (propositions), ce qui en conditionne la rareté d'occurrence.

approche de la cohésion discursive spécialisée pour l'oral²⁶, orientée vers la description des structures informationnelles à partir de l'intonation. Ils différencient trois ensembles informationnels formant les énoncés : le *thème*, le *rhème* et le *postrhème*. Le *rhème*, caractérisé par une intonation « variable » (p. 115) constitue l'information nouvelle et est le seul élément obligatoire. Le *thème*, d'intonation « montante », comme le *postrhème*, d'intonation « basse » sont des structures d'intégration discursive : ils « importent-exportent » des informations du discours pour les relater, au niveau local-syntaxique, au rhème. Le postrhème a pour fonction de « boucler sur elle-même la prédication » (Morel, 2007:41). Son intonation particulière provient dès lors de la nécessité de marquer la cohésion avec ce qui précède (Danon-Boileau et al., 1991:111) :

Quand l'oral met en place une relation avec ce qui précède sans utiliser de marques syntaxiques, il doit nécessairement opérer un décrochage brutal (intonation basse) qui interrompt en fait la trame mélodique et forme incise.

Danon-Boileau et al. expliquent donc que ce contour intonatif prototypiquement bas, en « décrochage brutal », exprime un rattachement à l'unité précédente, qui, traduite dans les termes d'une approche macrosyntaxique, s'apparente au marquage régressif d'une dépendance. Morel ajoute que cette caractéristique intonative permet de différencier le postrhème d'autres structures d'information caractérisées par une indépendance intonative : les *afterthoughts*, les *appositions de précision* et les *compléments d'information*.

Dans le cadre du modèle du GARS, Mertens élabore un modèle des structures intonatives. Il conceptualise une « dominance » entre groupes intonatifs : ceux-ci peuvent être regroupés et articulés sous forme de dépendances, d'après une structure qui, comme le lecteur pourra le constater, n'est pas sans rappeler l'organisation macrosyntaxique des énoncés. Cette similitude

²⁶ Cette approche est très différente des théories de macrosyntaxe que j'ai présentées plus haut. L'approche de Danon-Boileau et al. (1991) tout comme celle de Morel (2007) s'attache à décrire des phénomènes de cohésion discursive là où le GARS ou la GP verront des structures macrosyntaxiques. Leur analyse repose sur une distinction bien établie entre l'oral et l'écrit qui induisent deux cohésions distinctes : à l'écrit prévaut une cohésion hypotactique marquée par la syntaxe et les dépendances de propositions, tandis qu'à l'oral prévaut une cohésion paratactique indiquée par l'intonation.

est précisément due au fait qu'un lien *indexical* existe entre les deux structures. Le locuteur utilise l'intonation pour donner des indices de regroupements de macrosyntaxe, et c'est à partir de l'intonation que l'allocutaire reconstitue la macrosyntaxe en inférant des regroupements²⁷.

Mertens modélise l'énoncé intonatif par le schéma suivant (1990:167), où le symbole « % » représente les frontières du groupe intonatif ; les unités entre parenthèses sont facultatives ; la notation 'AI' représente un accent *initial* et la notation 'AF' un accent *final* – ou principal.

% ((non accentuée) AI) (non accentuée) AF (appendice) % [je souligne]

Le regroupement correspondant aux postfixes – et par conséquent aux périphéries droites d'énoncés – est celui de l'intonation d'*appendice*. Malgré des « relations [micro-]syntaxiques diversifiées »²⁸, les appendices se présentent *tous* sous la forme d'une « intonation remarquable » (Mertens, 1990:166), caractérisée par un « contour plat, [...] le manque d'accentuation et [...] une diminution du niveau sonore » (Mertens, 2009:5).

Le prosogramme ci-dessous représente un extrait enregistré dans lequel la séquence « qui pourra vous répondre » concorde avec cette description de l'appendice. La hauteur vocale, mesurée par la courbe bleue de la tire de description prosodique, chute brutalement au début de la séquence. Les traits noirs, qui synthétisent les variations mélodiques, sont à ce moment précis alignés sur la même fréquence de 150 hertz, montrant que l'intonation cesse d'être modulée par des variations accentuelles.

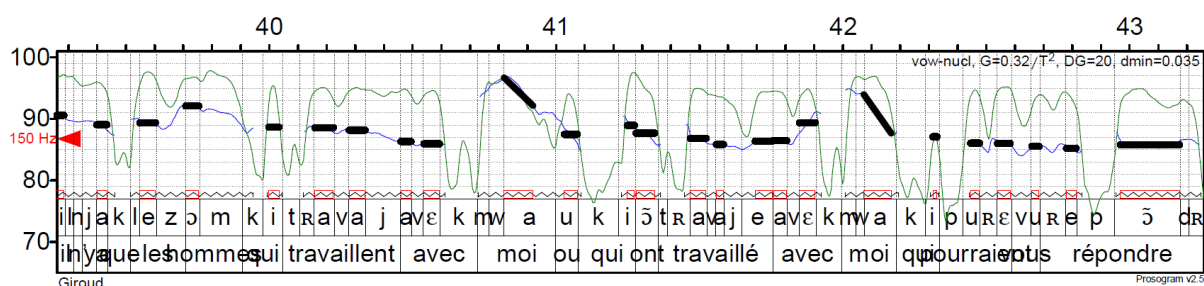
²⁷ Le GARS explicite le lien entre intonation et syntaxe de façon équivoque. Blanche-Benveniste caractérise ainsi les postfixes à partir de l'intonation (1990, p. 168) :

Le regroupement entre noyau et postfixe est caractérisé par une intonation remarquable [d'appendice].

Tandis que Mertens (1990, p. 168), en présentant l'intonation d'appendice, raisonne dans l'autre sens et indiquant que c'est la structure macrosyntaxique de postfixe qui détermine ce schème intonatif :

Dans l'étude sur la macrosyntaxe, tout ce qui est présenté comme postfixe a cette intonation particulière d'appendice.

²⁸ Piet Mertens ne rattache pas de catégorie particulière à cette intonation d'appendice ; il indique même des cas pour lesquels cette intonation d'appendice ne s'applique même pas à un constituant complet (Mertens, 1990, p. 169).



(Tiré de Mertens, 2009:5)

Il convient de mentionner que dans l'extrait ci-dessus, c'est un dispositif en relative de type 'qui' qui prend une forme d'appendice. Cet exemple précis a été choisi pour illustrer la diversité des segments susceptibles de correspondre à ce contour intonatif. De fait, dans les « corrélats syntaxiques » que j'ai décrits plus haut, aucun n'englobe les constructions relatives : cette réalisation intonative particulière correspond à un « marquage »²⁹ intenté spécifiquement par le locuteur.

Pour Mertens, l'intonation d'appendice est un « type particulier de contour intonatif », qui se différencie de tous les autres par cette absence d'accentuation (Mertens, 2009:5). Fonctionnellement, cette caractéristique crée d'après lui un phénomène d'arrière-plan vis-à-vis des éléments précédents, quant à eux accentués : de la sorte, l'appendice se dégage de la mélodie générale pour représenter un « élément optionnel après l'accent final » (Mertens, 1990:167) ; il fonctionne de manière rétroactive pour rattacher le segment au précédent et, dans le même temps, le discriminer du segment suivant, qui se trouve souvent être non accentué lui aussi (Mertens, 1990:167).

²⁹ Ici, mon objectif sera d'identifier des périphéries droites non marquées. J'entends le concept de « non marquage » au sens qu'en propose Mertens (2009) : l'intonation non-marquée est l'intonation par défaut d'une séquence. C'est l'écart entre cette intonation non marquée et l'intonation réellement utilisée par le locuteur qui sera productrice de sens (Mertens, 2009, p. 2). Autrement dit, les formes marquées sont significatives par leur présence, mais également par leur absence. Dans l'extrait ci-dessus, la prononciation en appendice d'une construction relative, de par son caractère marqué, témoigne d'une volonté du locuteur de placer le focus informationnel sur la séquence précédent l'élément en appendice, « les hommes qui ont travaillé avec moi ».

3.1.2. SCHÈMES ATYPIQUES

Si la reconnaissance d'un contour intonatif caractéristique des périphéries droites se dégage, cette observation est cependant nuancée par certains auteurs. Morel présente ainsi le cas des « questions » et des « suspensions de validation » comme deux exceptions à cette modulation en plage basse. Ces deux extraits ont été tirés de Morel (2007:44) [adaptation à la transcription Valibel]

il est gentil ton maitre ? allez parle-moi un peu de ton maitre comment il est
je sais pas si elle est vraie cette histoire parce qu'elle paraît complètement euh

Dans les deux extraits, les structures « postrhèmes » qu'a identifiées Morel, ici soulignées, ont été produites avec une intonation montante. Pour Morel, cette particularité intonative trouve sens dans l'interaction : elle traduit en fait un appel lancé au coénonciateur, à qui il incombe de valider la relation entre le rhème et le postrhème.

La *Grammaire de la période* s'est aussi intéressée à ces contre-exemples d'intonation dans la routine [*action* ▷ *confirmation*]. Arrêtons-nous un instant sur la façon dont la *Grammaire de la période* traite les données d'intonation. Dans le cadre de la GP, c'est la structure des énonciations qui détermine l'intonation, laquelle est ainsi le reflet des intentions pragmatiques de l'énonciateur³⁰. Les chercheurs du groupe de Fribourg s'attachent dès lors à identifier les routines de la pragmasyntaxe, traduites par des prototypes d'*intonèmes*. Les intonèmes ont trois fonctions liées à la macrosyntaxe³¹ : la démarcation des énonciations ; l'indication de relations de dépendance/dominance/juxtaposition ; l'indication d'une modalité communicative – notamment la question, le doute ou l'antagonisme.

Dans le cadre de la *Grammaire de la période*, les segments postfiniaux ne sont pas « en plage basse » mais se présentent comme un *intonème réduit*. La réduction d'intonème est un

³⁰ L'intonation est considérée par la GP comme un ensemble de « signes de nature mimo-gestuelle incorporés à l'énonciation » (Berendonner, 2012, p. 95) ; de la même manière que les signes gesturaux, celle-ci est donc mise en relation avec l'acte de langage intenté.

³¹ Par opposition à d'autres qui ont une prennent sens pour donner des indications morphosyntaxiques, ou encore obéissent à des régularités rythmiques.

procédé qui comporte deux aspects : d'une part, l'intonème constitue (1) la « copie réduite » de la suite intonative précédente (voir Rossi 1999:85), tandis que d'autre part (2) la hauteur de départ de l'intonème est déterminée par la hauteur finale de ce qui précède. Ces deux aspects permettent, en pragmasyntaxe, de signifier l'annexion du segment à son prédécesseur direct. L'exemple de Morel « il est gentil *ton maitre* ? » trouve dès lors une autre explication : l'intonème représentant la modalité communicative de question, dont la finale est montante, sera en fait copié sur les deux syllabes du segment postfinal, ce qui produira une nouvelle intonation montante, mais dont le rattachement à la clause précédente est sans équivoque.

II. DONNÉES ET ANNOTATION

1. LA BANQUE DE DONNÉES VALIBEL

1.1. OBJECTIFS ET CONSTITUTION

La banque de données textuelles orales Valibel³² s'est constituée progressivement à partir de 1989 avec le souci de renouveler les études sur le français de Belgique en rassemblant une quantité significative de données authentiques, issues de leur contexte effectif de production. L'approche est revendiquée comme à la filiation de deux courants de recherches concernant le « français parlé » : le *Groupe Aixois de Recherches en Syntaxe* (GARS), mené par la figure de Claire Blanche-Benveniste, en particulier pour sa méthodologie de transcription orientée vers les phénomènes morphologiques et syntaxiques, et d'autre part les sociolinguistes québécois s'intéressant à la variation du français oral (Cedergren, Deshaies, Sankoff, Thibault, Vincent), pour leur démarche variationniste et leur approche des données sociolinguistiques et situationnelles (Bardiaux 2012).

La banque de données textuelles orales Valibel n'est pas à proprement parler un corpus, et encore moins un corpus de référence pour le français parlé en Belgique. La construction d'un corpus de référence, en effet, demande d'effectuer des choix méthodologiques à propos de ce qui constitue la variété ou la langue considéré, des discours et des registres susceptibles de la représenter, et avec quel équilibre. Lors de chaque analyse des données, il serait en outre nécessaire d'assumer les artefacts heuristiques d'un tel choix. Dans l'objectif de « documenter la diversité des pratiques observables », la constitution de Valibel s'oriente donc vers la constitution d'un « réservoir » de données (Dister, Francard, Hambye, & Simon 2009: 9) – dont certaines composantes peuvent être équilibrées.

³² Pour « *V*ariation *L*inguistique du français en *B*ELgique. »

Le développement des supports informatiques et les nouveaux champs de recherche conduisent progressivement à la construction d'une banque de données textuelles orales de grande envergure : le corpus Valibel est ainsi actuellement constitué d'environ 3.900.000 mots, pour un total de 373 heures d'enregistrement (Valibel 2017). L'évolution des supports informatiques conduit à un enrichissement des données accessibles.

1.2. ASPECTS DE LA TRANSCRIPTION

Le cœur du projet Valibel a consisté à élaborer des conventions de transcription et de métadonnées susceptibles d'être ouvertes à des contributeurs hors du centre de recherche (mémorants, doctorants, etc.). Un autre enjeu pour la mise en forme des données était de les rendre accessibles à des non spécialistes ou des chercheurs tiers, sans surcharge d'information. Les fichiers du corpus Valibel se présentent ainsi sous la forme de textes lisibles en orthographe standard (cf. 1.2.1 *Conventions graphiques, ci-après*). Dès le début des années 2000, les outils de traitement des données orales se généralisent et vont mener à intégrer, dans la banque de données, des fichiers alignant ces transcriptions orthographiques avec les fichiers sonores des enregistrements (Dister, Francard, Hambye, & Simon 2009: 6).

La transposition du signifiant de parole vers un support lisible et réutilisable – qu'il soit seulement écrit ou multicanal – n'est pas un acte neutre, mais occasionne un grand nombre de choix méthodologiques. La démarche d'annotation développée dans ce chapitre, consistant à identifier un type particulier de segments (macro-)syntaxiques uniquement à partir des données écrites, est dès lors tributaire de ces choix de transcription. Plusieurs aspects nécessitent un questionnement : les informations concernant la prosodie seront-elles réutilisables, ou bien, au contraire, parasiteront-elles l'analyse en raison de la trop grande disparité ou la contingence des phénomènes qu'elles décrivent ? Les conventions graphiques seront-elles pertinentes pour rendre compte de régularités morphosyntaxiques comme les récurrences de certains marqueurs, malgré des prononciations distinctes ? Jusqu'où sera-t-il

possible d'interpréter une énonciation et d'inférer le déroulement pragmatique de l'interaction ?

Dans la suite de cette question, j'aborde succinctement trois aspects de la transcription dans la banque de données Valibel : les conventions de graphie, les informations concernant la prosodie et la représentation linéaire des dynamiques multidimensionnelles de l'interaction.

1.2.1. CONVENTIONS GRAPHIQUES

Un principe des conventions de transcription Valibel est de suivre une orthographe conventionnelle calquée sur les standards du français écrit, afin d'en faire un support aisément consultable. Cette transcription, cependant, s'abstrait de la ponctuation et renonce à utiliser des « trucages » graphiques dont le rôle serait de rendre compte de certaines réalisations phonétiques propres à l'oral en modifiant la graphie standard (Dister, Francard, Hambye, & Simon 2009: 6).

Ainsi, différentes réalisations phonologiques d'un même morphème n'apparaissent pas dans la transcription orthographique : la forme '*enfin*', par exemple, ne sera pas orthographiée différemment si elle est réalisée sous sa forme complète /ɑ̃.fɛ̃/ ou sous sa forme tronquée /fɛ̃/, et cela même si cette différence correspond à des régularités³³. De la même façon, le

³³ Le cas de '*enfin*' n'est pas isolé : la banque de données Valibel oriente ses conventions vers une uniformisation graphique des « particules de l'oral », qui seront fréquemment rencontrées dans l'identification des périphéries. Une liste indicative en est donnée dans les conventions de transcription du groupe Valibel (Bachy, et al., 2007, p. 12) :

ah	oh
aïe	oh la la, oh là là
bah	ok
bè [bE]	ouais
ben [be~]	ouille
eh [E]	oula, ouh là
enfin [a~fe~, fe~]	p [p]
euh	pf [pf]
gnagnagna	pff [pf:]
hein	pt
hum	t
m [m]	wouf, waf
mm [mm]	
moui	
mouais	

morphème ‘*il y a*’, dont on peut trouver quatre réalisations /i.li.a/ /il.ja/ /i.ja/ ou /ja/), est considéré, malgré trois mots graphiques, comme une unité et n’est donc pas différencié dans la graphie (Bachy, et al. 2007). La réalisation effective de ‘*il y a*’ ou de ‘*enfin*’ ne sera dès lors accessible qu’à partir des enregistrements alignés à la transcription.

Ce choix méthodologique des graphies unifiées a pour avantage de séparer les niveaux d’articulation de la langue : une perte d’information (morpho)phonologique se fera aux avantages d’une prise en compte des régularités de la morphosyntaxe ; un avantage pour l’identification de segments tels que les périphéries droites.

1.2.2. INFORMATIONS PROSODIQUES

Seuls trois symboles notent, dans les conventions Valibel, des caractéristiques prosodiques de l’énonciation. Deux d’entre eux notent les pauses : les pauses silencieuses courtes (‘/’), les pauses silencieuses longues (‘//’) ; un autre marque une intonation montante caractéristique des questions (‘?’). Les pauses pleines³⁴ ne sont pas « symbolisées », mais sont représentées par des graphies homogénéisées : ‘*eu*h’, ‘*hm*’, ‘*hum*’. Les différences de longueur ne sont pas représentées. Les autres phénomènes paraverbaux, comme les silences ou les inspirations, ne sont pas directement transcrits, mais peuvent être signifiés par des didascalies.

Les notations des pauses, que j’ai mentionnées plus haut, ne reflètent pas des informations factuelles liées aux caractéristiques acoustiques du signal, mais bien le jugement intuitif du transcripteur à l’écoute des enregistrements. Cela implique une variation importante, laquelle a déjà fait l’objet d’études approfondies qui révélant la prévalence de certains biais de transcription. Un facteur clé serait ainsi la perception relative de la vitesse d’articulation locale (Grosjean & Deschamps 1975: 156). Du côté de la production, pourtant, Vercruyssen, Lacheret & Fossard (2014) constatent une corrélation faible entre longueur des pauses et vitesse d’élocution. Ce sont bien davantage les conditions cognitives de production et de

m’enfin

³⁴ Les « pauses pleines » peuvent être définies comme « toute vocalisation qui maintient le signal dans une position neutre » (Crible, Dumont, Grosman, & Notarrigo, 2015).

planification discursive, tels que la complexité référentielle et les « scripts d'arrière-plan » qui influenceraient l'apparition et la longueur des pauses chez un locuteur. La notation des intonations montantes ('?'), dans les conventions Valibel, est elle aussi tributaire d'un jugement intuitif. De fait, il ne s'agit pas d'identifier chaque intonation finale montante, mais bien les « questions à forme déclarative » (Bachy, et al. 2007: 4). L'utilité principale de la notation est donc d'indiquer une opposition entre l'affirmation et la question lorsque d'autres marques linguistiques ne le permettent pas. Cette caractéristique en fait une donnée délicate pour l'identification certaines démarcations prosodiques.

L'annotation systématique des périphéries droites que j'effectue ci-après se doit de trancher face à cette variabilité des notations prosodiques dans les transcriptions écrites. Les choix d'analyse ont consisté à écarter la notation des intonations montantes ('?') des analyses. Les pauses longues ('//') sont quant à elles systématiquement considérées comme des démarcations entre unités intonatives. Les pauses courtes ('/') tout comme les pauses pleines – indépendamment de leur longueur, non notée – ne sont pas des démarcations pertinentes ; elles peuvent en revanche le devenir par combinaison ('euh /' ; '/ euh'). L'élément 'euh', pour sa part, n'est pas assimilable à la fonction de marqueur de discours ou de marqueur pragmatique, que ce soit en usage isolé, ou en cooccurrence avec d'autres marqueurs ('donc euh' 'mais euh').

1.2.3. INFORMATIONS CONCERNANT LES TOURS DE PAROLE

La banque de données Valibel appréhende la notion de tour de parole comme « l'occupation matérielle du canal par un locuteur » (Bachy, et al. 2007: 7). De manière liée, deux paroles se chevauchent quand deux locuteurs occupent simultanément une partie du canal. Cette définition physique, qui présente l'avantage de trancher de façon nette et d'éviter une trop grande variabilité dans la transcription des interactions, présente toutefois des désavantages lorsqu'il s'agit de prendre en compte (1) des interventions, i.e. des continuités énonciatives au-delà de ces « tours de parole », ainsi que (2) des discontinuités en deçà de ceux-ci.

Dans le premier cas (1), des interventions à part entière ne seront pas considérées comme des "tours de parole" autonomes. Un locuteur A et un locuteur B peuvent ainsi s'exprimer simultanément, sans que la simultanéité n'implique nécessairement un conflit de prise de parole dont la résolution serait une conservation de la parole de A ; pas plus qu'elle n'implique une non prise en compte de l'intervention du locuteur B. Dans cet extrait, on perçoit ainsi qu'un échange a bien cours alors qu'un des deux locuteurs est entièrement recouvert par l'autre.

blaAT1 ben tu |- vois en général <blaND1> c'est long hein -| oui c'est long
(blaAT1l:69)

La représentation des contributions en deux tires parallèles rend l'échange plus visible :

blaAT1 ben tu vois en général	oui c'est long	
blaND1	c'est long hein	(blaAT1l:69)

Dans ce passage, par exemple, la réponse « écholalique » de la locutrice blaAT1 indique qu'elle a parfaitement considéré la contribution de blaND1 et qu'il s'agissait donc dans ce cas-ci d'une intervention à part entière, au sens où un acte de langage est bien exécuté.

Des cas (2) peuvent survenir dans lesquels, à l'inverse de ce que je viens d'exposer, une énonciation est continuée au-delà des tours de parole transcrits. Ici, la production d'un locuteur B est considérée comme un tour de parole pertinent dans l'annotation, alors que le locuteur A n'est interrompu ni dans son élocution, ni dans son énonciation. Dans l'extrait suivant, le problème est représenté par un marqueur de type « régulateur », 'm', lequel est rarement considéré comme un acte de langage à part entière puisqu'il soutient l'énonciation du colocuteur :

famHC1	donc moi j'ai
famED1	m
famHC1	très bien mangé et le dessert fort fin ou / euh

La continuité syntaxique, ici, n'est même pas interrompue par l'allocutaire. En tout état de cause, toute forme de pause silencieuse, fût-elle brève, est susceptible d'être remplie par une

contribution de l'allocutaire. Alors qu'on se trouve face à une continuité de l'énonciation, le résultat en sera la représentation, dans la transcription, de trois tours de parole au lieu d'un.

Au niveau de l'annotation, cette caractéristique des transcriptions Valibel pose deux types de problèmes. D'une part, les « fins de tour », en tant que position pertinentes pour la typologie de certaines périphéries droites, ne sont pas données par la transcription et doivent faire intervenir le jugement de l'annotateur. D'autre part, ce choix de transcription posera problème lorsqu'il s'agira d'identifier des segments caractérisés par une continuité intonative. Les relations macrosyntaxiques et parmi elles les périphéries droites, qui dépendent de ces continuités, appellent donc la mise en place d'une notation spécifique des « macro-unités » ne correspondant pas aux tours de parole donnés par la transcription.

1.2.4. MISE EN SCÈNE DE L'INTERACTION

La transcription de certains phénomènes n'est tout simplement pas normalisée d'un transcripneur à l'autre : les « indications du non-verbal », notées entre parenthèses, sont ainsi utilisées de façon très différente en fonction des transcripneurs. Certains se contenteront de la notation de phénomènes paraverbaux ou non-verbaux bien identifiés – rire, silence, interférences, expiration, souffle, toux – tandis que d'autres transcripneurs ajouteront des didascalies narrant l'action (« s'adressant à une tierce personne » « bruit d'une porte qui s'ouvre ») voire des commentaires méta-transcription (« on ne sait pas si laquelle des deux locutrices parle »). Les informations concernant ces phénomènes sont à considérer comme un support de compréhension, mais ne peuvent, de par leur variabilité de transcription, soutenir une analyse systématique. Ainsi, l'annotation des périphéries que je détaille ci-après ne considérera pas la notation '(silences)' comme une marque fiable d'identification.

1.3. CORPUS ÉTUDIÉS

L'analyse réalisée dans le cadre de ce mémoire est résolument orientée vers les conversations informelles ; le postulat de recherche est en effet que la dynamique de l'interaction et un faible

degré de formalité sont deux facteurs variables facilitant l'apparition de segments périphériques droits ; le phénomène, de ce fait, pourra être observé une plus grande fréquence, laquelle est susceptible d'entraîner une plus grande richesse des formes rencontrées.

Dans cet objectif, ce mémoire se concentre donc sur deux sous-corpus de la banque de données Valibel : la composante *fam* et la composante *bla*.

- La composante '*fam*' (Valibel 2003) reprend des conversations informelles, recueillies *in vivo* dans l'activité familiale. Chaque enregistrement provient de différentes régions de Belgique francophone.
- La composante '*bla*' (Valibel 2004) reprend des échanges informels entre deux locuteurs. Un observateur peut éventuellement jouer un rôle de troisième participant. Tous les extraits analysés dans le cadre de ce mémoire ont été enregistrés dans la province de Liège.

2. IDENTIFICATION MANUELLE DES PÉRIPHÉRIES DROITES

Dans cette partie, je propose une méthode d'annotation opératoire et reproductible adaptant la grammaire du GARS à un corpus de transcriptions écrites, en prenant pour base les transcriptions de la banque de données Valibel. Une application « au premier degré » de la grammaire du GARS consisterait à annoter, d'une part, la microsyntaxe, en faisant la différence entre les domaines régis et les éléments hors de la rection, et d'autre part la macrosyntaxe en délimitant les énoncés et en les décomposant en préfixes, noyaux, suffixes et postfixes. Toutefois, comme je l'ai soulevé, cette démarche d'application présente l'inconvénient de reposer sur des données d'intonation, non présentes dans la transcription orthographique.

La démarche que je propose a pour but d'adapter les critères du GARS à une limitation aux transcriptions écrites. La macrosyntaxe, en particulier, n'est plus observée ni annotée directement : elle est identifiée à partir des unités microsyntaxiques et des indices linguistiques qui corroborent ces structures, et qui sont directement observables dans la transcription.

Les périphéries droites sont l'élément cible de cette démarche d'annotation. Elles sont caractérisées comme l'ensemble des morphèmes dépendants, en termes macrosyntaxiques et intonatifs, d'un élément recteur, mais néanmoins indépendants de la rection de ce dernier. L'identification des *périphéries droites* proprement dites suppose donc d'établir, en amont, les limites de deux unités : l'*unité de rection*, qui opérationnalise les regroupements de rection théorisés par l'approche pronominale, et le *groupe intonatif*, qui se fonde sur des indices textuels et syntaxiques pour établir des liens de dépendance macrosyntaxiques.

Dans cette partie sont développés les choix méthodologiques généraux pour l'identification de ces trois unités ; une grille d'annotation est donnée en annexe de ce mémoire (*Annexe II*), qui reprend les codes d'annotation utilisés et décrit l'étiquetage d'annotation typologique réalisé

en complément, sur lequel reposent les observations de la partie 3. *Observations* de ce chapitre.

2.1. L'UNITÉ DE RECTION : ANNOTATION DES DÉPENDANCES MICROSYNTAXIQUES

L'identification de la *rection* en microsyntaxe revient à identifier les éléments recteurs et de délimiter ensuite la portée de leurs dépendances syntaxiques. Comme je l'ai développé dans le chapitre précédent, le GARS opère une distinction claire entre le phénomène de *rection*, marquant la microsyntaxe, et le *noyau* d'un énoncé, qui structure la macrosyntaxe. La *rection* caractérise la relation qu'entretient un élément *recteur* – le plus souvent un verbe, parfois un nom – avec des dépendants syntaxiques qui forment son *domaine régi*, tandis que le *noyau* structure un autre type de dépendances arrangeant les *énoncés* de façon tout aussi grammaticales, mais non qualifiables en termes de relations de rection.

Cette partie propose une méthode opératoire d'identification des unités de rection, orientée spécialement vers le repérage des périphéries droites. L'objectif n'est pas d'en dresser une typologie, mais de les délimiter de façon tranchée : discriminer, dans l'annotation, ce qui est *régi* et ce qui ne l'est pas. En extension de la définition donnée dans le chapitre précédent, je propose ici d'utiliser le concept d'« unité de rection » développé dans le modèle BDU (Degand & Simon 2009 ; Simon & Degand 2011), en l'orientant spécialement vers le repérage des périphéries droites. L'objectif n'est pas d'en dresser une typologie, mais de les délimiter de façon tranchée : discriminer, dans l'annotation, ce qui est *régi* et ce qui ne l'est pas. Une démarche d'identification en deux temps est ensuite présentée : dans un premier temps, il s'agit de repérer l'élément recteur articulant l'unité de rection ; dans un second temps, d'opérer une batterie de tests syntaxiques visant à délimiter les frontières de cette unité.

2.1.1. L'UNITÉ DE RECTION : OPÉRATIONNALISATION À TRAVERS LE MODÈLE « BDU »

En dehors du GARS, d'autres équipes et chercheurs ont utilisé la notion de rection pour séquencer des corpus de français parlé. Le modèle *BDU* (pour « basic discourse units » ; Degand & Simon 2009 ; Simon & Degand 2011) détermine ainsi des « unités discursives de bases » qui sont considérés comme autonomes dans la segmentation du discours. Les BDUs sont délimitées à partir de l'interrelation entre prosodie et (micro-)syntaxe : les « unités prosodiques majeures » et les « unités de rection » segmentent ainsi le discours à tour de rôle, et parfois de façon congruente. Ces dernières « unités de rection » sont définies à partir de l'approche pronominale du GARS, dans sa version étendue aux tests d'extraction de constituants (cf. Blanche-Benveniste 2002). Comme pour l'approche pronominale, toute unité de rection est constituée de formes dépendant d'un élément recteur identifié. À la différence du GARS, Degand, Martin & Simon (2012: 2614) ont une approche moins déductive et plus inclusive de la notion de rection, à travers laquelle elles proposent d'étendre la gamme des éléments potentiellement recteurs :

Une unité rectionnelle s'articule autour d'un noyau³⁵, c'est-à-dire un élément recteur. Généralement, L'élément recteur est un verbe tensé, mais il peut être également un constituant averbal, tel qu'un nom, un pronom, un adverbe ou un groupe prépositionnel prédicatif.

Là où Blanche-Benveniste décrivait des noyaux verbaux et des noyaux averbaux (1990: 126), Degand, Martin & Simon décrivent trois types d'unités de rection : les unités de rection *verbales* (1) sont organisées autour d'un « verbe tensé » et apparaissent comme « achevée[s], syntaxiquement et sémantiquement » (Degand & Simon 2011: 47) ; les unités de rections *averbales* (2) sont un regroupement syntaxique dans lequel un autre constituant sert d'élément central – nom, adjectif, pronom, adverbe, préposition, interjection ; enfin, les unités de rection *elliptiques* (3) se caractérisent par leur incomplétude syntaxique, mais leur complétude fonctionnelle : la séquence qui est omise mobilise un contexte discursif. D'autres publications (Degand & Simon 2009 ; Simon & Degand 2011) ajoutent deux autres types

³⁵ Notons que la notion de « noyau » ne recouvre pas le terme utilisé par le GARS en macrosyntaxe.

d'unités de rection : (4) les unités de rection *inachevées*, auxquelles manque au moins un complément obligatoire ou bien pour lesquelles un syntagme est tronqué ; et (5) les unités de rection « *plus* », qui incluent dans l'unité de rection les éléments *associés* à la rection. Ces dernières, qui incluent les périphéries droites, seront les seules à ne pas être retenues dans l'annotation.

2.1.2. ÉLÉMENTS POTENTIELLEMENT RECTEURS ET DISPOSITIFS DE RECTION

Une première étape de l'identification des unités de rection consiste à évaluer les éléments potentiellement recteurs, autour desquels vont être structurés les domaines régis.

(1) Rection verbale et ses dispositifs

L'élément recteur le plus courant est le *verbe*, qui caractérise les unités de rection verbale. Des cas de rection nominale sont également possibles. Plusieurs *dispositifs* de rection verbale sont envisageables au-delà du prototypique « dispositif direct », que le GARS ne considère pas comme « canonique » (1990: 55); certaines constructions font ainsi intervenir des « auxiliaires de dispositif » qui ne sont pas considérés comme le centre d'une rection. Dans tous les cas de dispositif, l'annotation retient une seule unité de rection ; les parties soulignées représentent le « foyer de rection » de chaque dispositif.

DISPOSITIF DIRECT	[elle <u>va</u> à la pharmacie chercher des petits] (silence) [*de de de quart de l/ de l/ pas même un quart de litre des petits packs] là (blaJD1l:124)
RELATIF	mais enfin [qui n' <u>est</u> pas si fictif que ça] mais (blaGD1l:103)
SUPPORT EN « IL Y A »	oh [<u>il doit encore y avoir</u> de ce vieux gâteau avec des petites croix dessus aux cerises] [non] (famES1r:884)
SUPPORT EN « C'EST »	[<u>c'est</u> vrai] (famRM1r:249)
INTERROGATIF	[est-ce que tu imagines] ? [dix-huit millions]
CLIVÉ / EXTRACTION	[<u>c'est</u> lui <u>qui est</u> le meilleur là-dessus] ?
PSEUDO-CLIVÉ	[<u>ce que</u> je fais aussi <u>c'est</u>] par exemple [*je vois des points de grammaire]

(2) Recteurs averbaux

Plusieurs constituants sont considérés comme des recteurs potentiels. Certains forment, à la manière du modèle de Tanguy et al. (2012: 14), des « séquences fonctionnelles autonomes ». Ces séquences sont indépendantes sur le plan macrosyntaxique ; en microsyntaxe, elles sont constituées d'un élément central qui attire à lui des constituants grammaticaux.

NOMS	quelle <u>affaire</u> (blaGD11:11)
ADJECTIFS	ah [pas <u>mal</u>] (famRM1r:171) [très <u>sympa</u>] // [on a bien papoté] elle
ADVERBES	[<u>effectivement</u>] (famRM1r:171) [<u>plus</u> que jamais] quoi [elle est enceinte] (blaJV11:153) [<u>peut-être</u>] qu'[elle voulait] que [ce soit une surprise] (famRM1r:224)
PRONOMS	[elle disait] que // [<u>quelqu'un</u> de très chouette] (blaJV11:16) [<u>moi</u>] [qui suis pas très forte en français] (famRM1r:269) [même pas aller les promener] / [<u>rien</u>] (blaNB11:269)
GROUPES PRÉPOSITIONNELS PRÉDICATIFS	[non non] [dist-elle pas] / [<u>dans</u> quinze jours] (blaJD11:58)

(3) Formes nucléaires (noyaux spécialisés)

L'annotation code également comme *unités de rection* une série d'unités que le GARS qualifie de « noyaux spécialisés » (Blanche-Benveniste, 1990: 127), selon une définition plus large que les « éléments recteurs » proposés pour le modèle BDU. La caractéristique de ces segments est d'être indépendants sur le plan intonatif et de pouvoir former des énoncés autonomes. S'ils convoquent des dépendants, le lien de dépendance ne sera plus qualifiable en termes de rection ; les tests de substitution pronominal et d'extraction sont tous deux bloqués. Ces éléments formeront donc des unités de rection à part entière, *a priori* sans possibilité de sélectionner des compléments régis.

ADVERBES	mais mais [<u>quoi</u>] mais [ç/ ça provoque le cancer] (famES1r:1185)
INTERROGATIFS	[comment] ça ? (blaGD11:178)

ADVERBOÏDES	[oufti] / [ouais] [d'accord] (blaGD1l:126)
PRÉDICATIFS	[dommage] qu'[il n'y a pas une caméra] (blaAT1l:256)
ADVERBOÏDES	[nom d'une pipe] // [je n'ai rien à dire dessus]
EXPRESSIFS	(blaJD1l:06)
(INTERJECTIONS)	m'enfin euh pf [merde] quoi tu vois (blaAT1l:145)
	[chapeau] hein (blaJD1l:180)
FORMULES	[allô] Mick (blaAT1l:85)
PERLOCUTOIRES	[pardon] (blaAT1l:94)
	[tantôt] hein [bisou] hein (blaAT1l:85)
ONOMATOPÉES	puis [zouf] (blaJD1l:180)
	[pouf] [on remet les pendules à l'heure] (blaAT1l:267)

La grille d'annotation liée à ce mémoire propose une liste de travail, non-exhaustive, de ces *formules nucléaires*. L'identification des dépendances macrosyntaxiques, dans la seconde passe d'identification, offre la possibilité d'en identifier de nouvelles.

Précisons encore que si l'annotation considérera ces éléments comme des unités de rection, tel n'est pas le propos initial de la théorie grammaticale aixoise, qui ne considère de ces éléments que leur rôle macrosyntaxique de « noyaux ». Au risque de perdre certaines informations que seule la prosodie serait susceptible de délivrer, l'objectif de l'annotation présentée ici n'est pas de contester cette description, mais de proposer une démarche d'identification uniforme, dans laquelle la notion de rection et la notion de noyau sont, de par leur corrélation empirique, assimilés sous la même notation.

2.1.3. IDENTIFICATION DES UNITÉS DE RECTION

À partir de ce que j'ai décrit comme *potentiellement recteur*, deux tests sont appliqués pour évaluer si un complément est régi ou non : la substitution par une proforme pronominale³⁶ et l'extraction dans un dispositif clivé '*c'est ... qu-*'. Les exemples suivants montrent comment la réduction à des proformes pronominales et l'extraction dans des clivées permettent d'établir les frontières de l'unité de rection :

(1)	il est trop fort quoi franchement	(blaJV11:21)
Réduction pronominale	il l' est quoi franchement	
Extraction en clivée	c'est trop fort qu'il est quoi franchement *c'est franchement qu'il est trop fort quoi *c'est quoi qu'il est trop fort franchement	
Frontières de l'UR	[il est trop fort] quoi franchement	
(2)	il faudrait peut-être en parler / franchement avec eux tu vois	(blaAT11:47)
Réduction pronominale	il faudrait peut-être leur en parler / ainsi tu vois	
Extraction en clivée	c'est franchement qu'il faudrait peut-être en parler avec eux tu vois c'est avec eux qu'il faudrait peut-être en parler franchement tu vois ?c'est peut-être qu'il faudrait en parler franchement avec eux tu vois *c'est tu vois qu'il faudrait peut-être en parler franchement avec eux	
Frontières de l'UR	[il faudrait peut-être en parler / franchement avec eux] tu vois	

L'approche pronominale et l'extraction en clivée sont utilisées de manière complémentaire pour trancher des cas ambigus. Ainsi, en (2), on constate le caractère ambivalent de '*peut-être*', qui répond positivement à l'extraction, mais pas à la substitution pronominale. L'orientation

³⁶ Blanche-Benveniste donne une liste des proformes considérées comme « pronominales » dans le cadre des analyses du GARS :

<i>je</i>	<i>me</i>	<i>y</i>	<i>là [ici]</i>	<i>ça</i>	<i>qui</i>	<i>quelqu'un</i>	<i>personne</i>
<i>tu</i>	<i>te</i>	<i>en</i>	<i>ainsi</i>	<i>ceci</i>	<i>que</i>	<i>quelque chose</i>	<i>rien</i>
<i>il</i>	<i>se</i>		<i>alors</i>	<i>cela</i>	<i>quoi</i>	<i>quelque part</i>	<i>nulle part</i>
<i>elle</i>	<i>le</i>			<i>celui-ci</i>	<i>où</i>		
<i>on</i>	<i>la</i>			<i>celui-là</i>	<i>comment</i>		
<i>nous</i>	<i>les</i>						
<i>vous</i>	<i>lui</i>						
<i>ils</i>	<i>leur</i>						
<i>elles</i>							

Tiré de Blanche-Benveniste (1990, p. 42)

de l'annotation sera, dans ce type de cas, inclusive : un seul test positif valide l'inclusion de l'élément dans l'unité de rection. Une exception a posteriori est celle des cas où une périphérie droite potentielle se présente, auquel cas l'hypothèse de la périphérie droite potentielle sera privilégiée afin de maximiser le rappel des formes.

Contrairement à ce que propose le modèle BDU, les unités de rection ne sont pas considérées comme extensibles à leurs associés³⁷. Les proformes pronominales retenues pour les tests de substitution ne s'étendent pas aux pronoms « semi-lexicaux » ; les règles qui définissent la rection des verbes et des noms respectent par ce fait les propositions de l'approche pronominale pour le traitement des domaines régis.

2.1.4. RÉPÉTITIONS, REFORMULATIONS, REPRISES ET UNITÉS INACHEVÉES

Dans certains cas, la réduction pronominale et l'extraction en clivée ne suffiront pas à délimiter une unité de rection, certains constituants pourtant obligatoires dans la valence du verbe seront lacunaires ; d'autres seront excédentaires. Ces irrégularités de production sont à comprendre dans le flux des conversations : certains éléments sont répétés ou interchangeés par le locuteur en cours de production, et sont la trace du travail cognitif et de la replanification permanente du discours.

Les (1) *répétitions* d'éléments, à l'identique, de même que et les (2) *reprises* substituant un constituant par un autre à la même position syntaxique, ne sont pas l'objet d'une annotation particulière. Le raisonnement est le suivant : si un constituant dépendant est substitué, la même unité de rection est toujours en cours d'élaboration – et ce, même si l'élément recteur n'a pas encore été exprimé. Ainsi, si une unité de rection est syntaxiquement limitée dans le

³⁷ Simon et Degand (2009:7) proposent la catégorie des *unités de rection* « plus » ('URC+') au travers desquelles elles étendent la notion de dépendance aux relations macrosyntaxiques :

This micro-syntactic analysis is then expanded to the macro-syntactic level which includes so-called 'associés' ('adjuncts') which are not governed by the main verb (hence offering no possibility for clefting), but are semantically or pragmatically linked to the whole dependency clause (in a 'préfixe' or 'postfixe' establishing a pragmatic relationship to the main clause).

Comme l'illustre cet extrait, la prise en compte des 'URC+' consiste à intégrer dans la rection les 'associés' préfixés et postfixés, quand ils sont définis ici comme périphériques – i.e. extérieurs – au domaine de la rection.

nombre de constituants qu'elle peut contenir, ces constituants peuvent quant à eux être virtuellement répétés et reformulés à l'infini³⁸. Ainsi, cette unité de rection, en dispositif clivé, formera une seule unité de rection :

ben [c'est toi qui m'as / qui m'as / qui m'as / qui m'as montré comment on devait faire]
(rires) (blaAT11:95)

Les répétitions et les reprises sont à différencier des *constructions reformulées* et des *constructions inachevées*, dans lesquelles l'unité de rection est abandonnée en cours de construction par le locuteur. Dans certains cas, on n'aura que l'amorce d'un morphème ; dans d'autres, c'est un constituant obligatoire de la valence qui sera lacunaire. Dans le cas d'une reformulation, une nouvelle unité de rection survient en lieu et place de la précédente :

et maintenant [de nouveau ils sont s/] [Pacs a dû mettre cent euros de sa poche] là
(blaJV11:287)

Les constructions inachevées sont identiques aux constructions reformulées, mais ne sont pas remplacées par une autre unité de rection. Cet aspect pourra poser des problèmes d'interprétation, en particulier lorsque l'élément recteur n'a pas encore été exprimé ; en cas d'incertitude, l'hypothèse de l'unité de rection est privilégiée :

[ja/] mais [c'est sérieux] [la semaine dernière] (blaJD11:82)

2.1.5. RELATIONS MICROSYNTAXIQUES ENTRE URS

Toutes les formes de constituants régis sont réductibles à des formes pronominales. Une unité de rection elle-même peut ainsi se trouver être régie par une autre et, de ce fait, peut être

³⁸ Le GARS représente ces phénomènes de reformulations et de répétitions par une « analyse en grille » (Blanche-Benveniste 1990b). Cette représentation permet de trancher entre la prolongation d'une même unité de rection ou la formulation d'une nouvelle construction. Dans l'exemple suivant, l'analyse en grille montre que c'est bien un dispositif clivé qui est répété, plutôt qu'une nouvelle unité de rection à dispositif de relative :

ben [c'est toi qui m'as /
qui m'as /
qui m'as /
qui m'as montré comment on devait faire] (rires) (blaAT11:95)

substituée dans son ensemble³⁹. Il serait donc possible d'annoter la relation entre les deux unités de rection de la façon suivante :

[je suis pas sûre moi [que ça se passe comme ça]₂ vraiment]₁ [tu vois]₃ (blaATil:51)

[j'en suis pas sûre] moi tu vois

Cette situation n'est pas un problème dans le cadre de l'approche du GARS, qui va, en microsyntaxe, décrire ce phénomène comme un cas d'*enchâssement* de domaines régis, et en macrosyntaxe comme une suite *noyau + suffixe* régulière. Pour une identification des périphéries, en revanche, les cas d'enchâssement sont une problématique réelle et demande d'appliquer une autre démarche d'annotation. Dans l'exemple ci-dessus, trois éléments ne sont pas régis : '*moi*' est associé à l'UR₁ ; '*vraiment*' est associé à l'UR₂ et à droite de celle-ci ; enfin, '*tu vois*', est un cas de rection faible (cf. chapitre 1) portant sur l'énoncé dans son ensemble et ne se rapportant ni à l'UR₁, ni à l'UR₂. Trois problèmes sont identifiables :

- (1) '*moi*' devrait, s'il est suivi d'un élément régi, être annoté comme un associé en incise séparant en deux l'UR₁, alors que la pronominalisation de l'UR₂ montre qu'il se comporte comme un élément périphérique ;
- (2) si l'UR₂ est régie par l'UR₁, alors '*vraiment*', nonobstant son fonctionnement périphérique dans l'UR₂, doit être lui aussi annoté comme un élément régi ;
- (3) '*tu vois*', en étant annoté comme une unité de rection, est isolé de l'énoncé auquel il se rapporte, alors qu'il constitue un cas de rection faible caractérisé par son utilisation périphérique.

Le problème d'une annotation stricte des unités de rection enchâssées est de produire un « silence » pour ces trois types de phénomènes. Les éléments en périphérie droite ne sont pas les seuls concernés : c'est également le cas des associés, marqueurs et recteurs faibles en périphérie gauche.

³⁹ L'extrait suivant ne reflète pas les choix d'annotation.

Afin de conserver, dans le cadre de ce mémoire, une notation réduite, l'annotation représente donc chaque unité de rection comme un segment isolé d'un continuum, n'articulant pas d'autres unités de rection⁴⁰. En conséquence, le début d'une unité de rection marque nécessairement la limite de la précédente ; la relation d'enchâssement entre deux unités de rection n'est donc pas explicitement annotée. La conjonction '*que*' et ses locutions dérivées ne sont pas considérés comme des éléments régis mais comme un connecteur « inter-syntaxique » reliant les deux unités de rection.

[je suis pas sûre] moi que [ça se passe comme ça] vraiment [tu vois]₃ (blaAT1l:51)

[c'est dingue] que [ce soit euh aussi petit d'ailleurs] (blaAT1l:76)

Cette notation s'applique également aux relations entre unités de rection pour lesquelles une UR fonctionne comme un associé – substituable à une proforme « semi-lexicale » :

[c'était clair] que [Bulle pouvait y a/ n'est pas restée] enfin (famES1r:25)

[c'était clair] **ça** enfin

Les influences possibles de l'enchâssement de domaines régis sur l'occurrence et la nature des segments périphériques droits mériteront d'être explorées dans les données.

2.2. LE GROUPE INTONATIF : ANNOTATION DES DÉPENDANCES MACROSYNTAXIQUES

Outre les relations de rection décrites ci-dessus, le GARS identifie d'autres liens grammaticaux de dépendance et d'interdépendance, qui sont qualifiés en termes de relations macrosyntaxiques. Contrairement à la *Grammaire de la période*, qui définit une unité maximale de la macrosyntaxe, le GARS ne définit pas où s'arrêtent les relations

⁴⁰ Ce choix d'annotation a également l'avantage de limiter les biais liés à une analyse des liens de subordination, dont de nombreuses études (notamment Benzitoun 2005 pour les constructions en '*quand*'; Blanche-Benveniste 2002 pour la problématique générale) ont montré que les « subordonnées » correspondent à différents liens pragmatiques et syntaxiques très différents. Leur distinction requiert une observation fine recourant une batterie de tests syntaxiques, ne se fiant pas à l'occurrence de tel ou tel mot subordonnant.

macrosyntaxiques ; annoter des « unités macrosyntaxiques » poserait donc des difficultés majeures dans une application consistante des théories du GARS.

Dans l'objectif d'étudier les périphéries droites, je développe un code d'annotation supplémentaire, supérieur aux unités de rection que j'ai définies précédemment. L'objectif de ce code d'annotation est de capter le mieux possible les éléments *postfixes* à partir des UR, en effectuant des tests permettant de simuler à la fois les relations macrosyntaxiques et la congruence de celles-ci avec des *regroupements intonatifs*⁴¹ prototypiques.

2.2.1. LE GROUPE INTONATIF « PAR DÉFAUT »

Pour bon nombre d'auteurs, la prosodie découpe la chaîne parlée en tronçons, dotés de frontières d'une certaine force, mettant ainsi en place une organisation hiérarchique. Les périphéries droites existent ainsi toujours en fonction d'un ensemble supérieur qui les rattache à la phrase à partir d'un lien de dépendance. Si le GARS n'envisage pas d'unité maximale en macrosyntaxe, j'explique ici qu'à partir de la modélisation d'une « intonation par défaut », une identification des « dépendances » peut néanmoins capter des relations qui sous-tendent un large ensemble de structures macrosyntaxiques. Cette approche factuelle, basée sur une estimation de faits de substances, limite le biais d'annotation en évitant une annotation *directe* de la macrosyntaxe. Je propose, à la suite de Mertens, d'extrapoler la notion de dépendance et de macrosyntaxe au domaine de la prosodie ; la cible recherchée est dès lors celle des « groupes intonatifs ».

⁴¹ J'emprunte et revisite le terme proposé par Delais-Roussarie & Post (2008) qui décrivent deux groupes « intonatifs ». Le *groupe intonatif syntaxique* (GI_{synt}) est « un domaine prosodique dont l'extension correspond à la phrase racine [...] – c'est-à-dire à la séquence *groupe nominal* et *groupe verbal* (SN SV) » ; caractérisé par sa frontière droite, le groupe intonatif est, dans une structure prototypique non marquée, congruent avec la syntaxe. Il correspond dans son principe aux URs que nous avons proposées dans la partie précédente. Le GI_{synt} se distingue du *groupe intonatif informationnel* (GI_{info}), lequel sera l'objet de notre annotation : ce dernier, qui imbrique les GI_{synt}, porte un « contour terminal » ; assimilé à la *phrase intonative*, il serait borné, à droite, par une pause ou un allongement très important, ainsi que par une frontière intonative majeure. Il se caractérise par une complétude informationnelle et constitue une unité discursive. Dans l'annotation, GI_{synt} sont les correspondants approxiatifs par les *unités de rection* (URs) ; analogiquement, les GI_{info} sont opérationnalisés par les *groupes intonatifs* (GIs).

Les notions de *dépendance*, de *construction* et de *macrosyntaxe*, empruntées au GARS, constituent le volet déductif de l'« approche prédictive de l'intonation » proposée par Mertens (2009). Le postulat formulé ici est celui d'une concordance *a priori* entre ces structures et une certaine intonation qui serait *non marquée*. Mertens (2009) utilise ainsi les représentations syntaxiques pour prédire une « structure prosodique par défaut » : à partir des données syntaxiques, il serait possible de dériver une intonation unique et standard. À la manière des transformations opérées sur la structure profonde en grammaire générative, l'écart entre cette intonation non marquée et l'intonation réellement utilisée par le locuteur sera dès lors productrice de sens (Mertens 2009: 2) ; les formes marquées sont significatives par leur présence, mais également par leur absence.

Jun (2003: 144) modélise elle aussi une « intonation par défaut » (*'default phrasing'*), en cherchant à établir la réalité psycholinguistique de cette notion. Cette structure serait également influencée, outre la syntaxe, par la longueur des mots, des groupes de mots et de l'ensemble de l'unité prosodique ; même en tant que réalité implicite, même en tant que « structure profonde » de la grammaire de l'intonation, Jun fait ainsi remarquer que cette intonation archétypale serait tout autant structurée par la métrique que par la syntaxe ; elle relèverait donc simultanément de la première et de la seconde articulation du langage.

Les cibles de cette seconde passe d'annotation sont les *segments périphériques droits*, caractérisés par leur forme d'*appendice*, à partir de l'unité supérieure dont ils dépendent et à laquelle ils se réfèrent. Cette unité intonative supérieure, qualifiée de « phrase intonative » dans la mesure où il s'agit de la plus grande structure autonome.

2.2.2. VARIÉTÉ DES RELATIONS À UNE UR

Pour une annotation complète, les possibilités de description des relations entre unités de rection telles que je les ai décrites sont nombreuses. À partir de la définition des unités de rection (UR) telle que je la propose, une annotation riche de la macrosyntaxe distinguerait en effet six niveaux d'étiquetage – dont les deux derniers, de nature discursive, jouent dans

l'annotation un rôle excluant. Ainsi, pour un élément x – syntagme isolé ou UR – et une unité de rection y , il est possible de proposer les relations suivantes^{42, 43} :

- (1) *L'interdépendance* : x implique y et y implique x , à l'instar des « noyaux complexes » de type '*plus ... plus ...*' ou '*autant ... autant*' ;
- (2) *La rection* : x est substituable dans y par une proforme pronominale ou/et répond positivement au test d'extraction dans y .
- (3) *L'association* : x est sémantiquement et discursivement relié à y ; il est sous-entendu par y ou représenté dans celui-ci par une proforme pronominale ;
- (4) *La modalité* : x modalise y et fonctionne envers celui-ci en tant que marqueur de discours ou marqueur pragmatique⁴⁴ ; x perd son sens sans y . Cette relation n'est pas discontinue avec la relation d'association ;
- (5) *La cohésion* : un connecteur discursif permet de rattacher x à y ; ce connecteur ou « introducteur » correspond à un procédé d'« empilement » de plusieurs unités de rection (Benzitoun, et al. 2010:2084)
- (6) *La cohérence* : le seul lien entre x et y est un voisinage discursif et une continuité sémantique.

À partir de la notation de ces six relations, de la position de l'élément considéré (à gauche ou à droite de l'UR) et de sa nature (unité de rection ou non), de nombreux phénomènes macrosyntaxiques pourraient ainsi être observés et décrits. Le tableau ci-après illustre le lien

⁴² Précisons que la notation de cette relation, des types 1 à 6, ne présage nullement de la fonction discursive de l'élément considéré. Ainsi, un élément disloqué à droite pourra tout à fait jouer un rôle de cohérence en rappelant un topic ; c'est néanmoins son lien d'association avec l'UR qui intéresse l'annotation à ce stade.

⁴³ Trois types de relations proprement *macrosyntaxiques* sont possibles en dehors de la rection : (1) l'interdépendance, (2) l'association et (3) la modalité – discursive ou pragmatique. Le GARS ne considère explicitement que les deux premières, tandis que la troisième forme de relation, telle que je la suggère est empruntée à la *Grammaire de la période* (Berendonner, 2012).

⁴⁴ Cette catégorie, non couverte explicitement par le GARS qui les considère généralement comme un sous-ensemble des associés, a été largement étudiée par la *Grammaire de la période* qui la décrit à travers plusieurs « routines macrosyntaxiques ». Elle concerne les cas de marqueurs discursifs et pragmatiques que j'ai rangés dans la catégorie des « corrélats » de périphéries droites (Blanche-Benvenise, 1990: 140-143).

entre ces relations, le concept de périphérie tel que je l'ai défini dans le premier chapitre, et les préfixes-suffixes-postfixes décrits par le GARS. Le « rattachement » de l'unité n'est pas une notion théorique, mais le critère opératoire d'identification des regroupements intonatifs dans l'annotation, que je détaille ci-après.

RELATIONS POSSIBLES ENVERS UNE U.R.

NATURES POSSIBLES ENVERS UNE UR						
Nature de la relation	1. Corrélation isomorphe ⁴⁵	2. Rection	3. Association	4. Modalité	5. Cohésion	6. Cohérence
Nature de l'élément						
UR	Noyau complexe	Rect. enchâssée	Segment périphérique			
Syntagme (non-UR)		Compl. régi				
Position de l'élément						
Gauche de l'UR		Préfixes				
Droite de l'UR		Suffixes	Postfixes			
Rattachement	INTERDÉPENDANCE	DÉPENDANCE			INDÉPENDANCE	
Annotation	1 unité de rection	1 groupe intonatif			2 groupes intonatifs	

Cette proposition d'annotation répond à la définition proposée dans la partie théorique de ce mémoire : les éléments que je qualifie de « segments périphériques » sont en dehors des unités de rection tout en étant compris dans les dépendants et, en conséquence, macrosyntaxiquement reliés à cette même UR. Ces segments correspondent aux postfixes et entretiennent envers les URs des relations qui vont de l'association à la modalité – les segments dépendants par lien de rection étant considérés comme des suffixes. Elles sont caractérisées par un lien de « rattachement » et sont « dépendantes » de l'UR.

Dans le cadre de ce mémoire, je propose donc d'identifier, dans un premier temps, le *rattachement* à l'UR – dépendance ou indépendance – des unités de rection (1) entre elles et (2) avec les éléments non régis. Le rattachement n'est *pas* à comprendre comme un concept théorique : c'est un critère d'identification correspondant à une méthode de jugement, visant à projeter sur la notion d'*unités de rection* celle de dépendances macrosyntaxiques que l'approche pronominal (cf. Delofeu 1991) applique, pour sa part, à partir des noyaux.

⁴⁵ Cette catégorie de relation est empruntée à Roig (2014) qui revisite les « noyaux complexes » (p. ex. '*plus ... plus ...*' ou '*autant ... autant ...*') décrits par le GARS. Par une série d'arguments, elle propose de considérer ces « noyaux complexes » comme un cas particulier de relations hypotactiques entre deux propositions.

2.2.3. ANNOTATION DES GROUPES INTONATIFS

L'identification des groupes intonatifs se résume à une annotation de leurs frontières ; celles-ci sont révélées au travers de quatre étapes successives que je décris ci-après à partir de Delofeu (1991), et de Mertens (2009)⁴⁶. Ces tests ne couvrant pas tous les cas de figure possibles, les indices de prosodie laissés par la transcription sont analysés à partir des thèses de Delais-Roussarie & Post (2008). Ces tests et indices de la transcription, toutefois, ne seront pas suffisants dans tous les cas ; il sera alors nécessaire de recourir au jugement intuitif de l'annotateur pour identifier certaines démarcations prosodiques ; la méthodologie que je propose alors, dans un souci de reproductibilité, extrapole certaines thèses entourant le concept de *prosodie implicite* ('*implicit prosody*', cf. notamment Fodor 2002 et Jun 2003). En conséquence, le produit de cette étape n'est pas une annotation entièrement objectivable : subjective et indicative, elle permet de conserver, dans le cadre du repérage des périphéries droites, une trace de l'interprétation de l'annotateur.

(1) La première étape d'identification des groupes intonatifs implique un repérage des dépendances macrosyntaxiques. Les éléments non-régis comme les unités de rection préalablement identifiées comme UR sont traités comme autant d'« îlots syntaxiques ». S'ensuit un test de suppression : chaque îlot est supprimé un à un ; ceux qui perdent leur valeur informationnelle, illocutoire, ou ne fonctionnent plus comme marqueurs d'une relation discursive en raison de cette absence d'éléments sont considérés comme non-autonomes et dépendants de l'unité qui a été supprimée. La dépendance envers une unité ne présage pas de l'autonomie de cette dernière : l'unité peut, à son tour, être de la sorte dépendante d'une autre et récursivement. Dès que cette récursion de dépendance s'arrête, l'annotateur se trouve face à l'unité racine, qui présentera une autonomie discursive et/ou illocutoire par laquelle elle sera alors assimilée à un noyau macrosyntaxique. Supprimée, cette unité-noyau provoque l'agrammaticalité de toute une séquence, laquelle sera considérée comme l'ensemble de ses dépendants et la première hypothèse de *groupe intonatif*. Lors de cette passe d'identification, il

⁴⁶ La notion de dépendance, de construction et de macrosyntaxe est le volet déductif de l'approche prédictive de l'intonation proposée par Mertens (2009).

est possible que de nouveaux noyaux se révèlent, témoignant d'une autonomie macrosyntaxique alors qu'ils ne mobilisaient pas de rection. Ceux-ci seront rétroactivement annotés comme unités de rection, étendant la liste de travail des « formules nucléaires ».

(2) En second lieu, l'annotateur identifie les marques explicites de prosodie laissées dans la transcription. Les normes de transcription de la banque de données Valibel donnent ici peu de latitude interprétative. L'annotation, cantonnée aux données de surface laissées par la transcription orthographique, est réduite à des choix binaires. Les *pauses longues*, représentées par une barre oblique double ('/') sont seules à être systématiquement considérées comme des frontières prosodiques majeures si aucun dépendant ne s'est manifesté lors de l'étape précédente. La présence de certains indices mineurs mais concomitants peut également aboutir à l'annotation d'une démarcation à partir de calculs que je propose d'opérationnaliser comme suit : la présence de pause simple ('/') vaut ½ démarcation ; l'interruption avec reprise vaut ½ démarcation ; la présence d'une pause pleine ('*eah*') vaut ½ démarcation.

(3) Après cette étape d'identification d'indices explicites, le jugement implicite de l'annotateur est convoqué dans la troisième étape de l'identification pour établir des liens ou des séparations entre unités. Pour rendre ce jugement reproductible, la thèse de la *prosodie implicite* (Fodor 2002) est utilisée. Toutes choses étant égales par ailleurs, le contour prosodique projeté lors de la lecture silencieuse correspond ainsi au « contour prosodique le plus naturel pour la construction »⁴⁷, et est, sur un texte ponctué, similaire à celle qui est effectuée lors de la lecture à voix haute. La thèse de la prosodie implicite a pu être attestée expérimentalement à de nombreuses reprises (notamment Frazier, Carlson & Clifton:2006) ; il a pu être démontré qu'elle permettait aux lecteurs de résoudre les ambiguïtés syntaxiques (Jun & Bishop 2015: 2). Cette thèse a deux implications pour l'annotation. La première est l'application du contour intonatif « par défaut » pour une construction donnée ; pour ce faire,

⁴⁷ La thèse de la prosodie implicite est exprimée par Janet Dean Fodor en ces termes : “The *Implicit Prosody Hypothesis*: In silent reading, a default prosodic contour is projected onto the stimulus, and it may influence syntactic ambiguity resolution. Other things being equal, the parser favors the syntactic analysis associated with the most natural (default) prosodic contour for the construction.” (Fodor, 2002, reprise dans Jun, 2003:221)

l'annotation se réfère aux conventions proposées par Mertens (2009, pp. 7-12). La prosodie implicite, outre le fait de calquer un contour intonatif « standard », obéit de façon annexe⁴⁸ à un ensemble de règles métriques liées aux capacités de stockage mémoriel⁴⁹ ; elle conditionne la longueur maximale des segments susceptibles de survenir sans frontière prosodique marquée. Au-delà des groupes de rection, le nombre maximal de syllabes d'un ensemble de dépendants intonatifs contigus, hors-UR autonome, est standardisé à huit, en correspondance à une loi rythmique (Morel 2007: 44). De la même façon, un groupe prosodique sera constitué, conventionnellement, de huit éléments – UR ou périphéries – au maximum. L'hypothèse est qu'au-delà de huit éléments, les contraintes métriques conduiront à la formation de deux unités intonatives plutôt qu'une seule.

(4) Dans la dernière étape, l'annotateur applique le postulat de congruence : à des frontières syntaxiques majeures correspondent des frontières prosodiques majeures. Autrement dit, en l'absence de dépendants, d'indices explicites ou implicites, les frontières syntaxiques majeures – les frontières des URs et, *a fortiori*, des tours de paroles – sont considérées comme correspondantes aux unités prosodiques dont les limites n'ont pu être déterminées lors des étapes précédentes.

2.3. LES PÉRIPHÉRIES DROITES : DÉLIMITATION ET SÉQUENÇAGE

Dans une dernière passe de l'annotation, les frontières des segments périphériques droits sont également fixées. Une fois les unités de rection et les groupes intonatifs annotés, leur repérage est aisé : elles constituent tous les segments situés à droite d'une UR et dans le même groupe intonatif que celle-ci. Seules les pauses pleines de type 'euh' [ø], sont, à ce stade, systématiquement exclues des segments périphériques.

⁴⁸ En ce sens qu'une contrainte rythmique sera « plus faible » qu'une contrainte syntaxique (Jun, 2003, p. 245).

⁴⁹ Jun (2003) évoque également des contraintes liées au débit de lecture, qui ne seront pas considérées ici, mais constituent un biais possible d'annotation. De façon analogue à la vitesse d'élocution pour une oralisation explicite, la vitesse de lecture influence la granularité du découpage intonatif. En général, des unités prosodiques lues en silence tendent à être de la même longueur, qui est fonction croissante par rapport au débit de lecture.

S'ensuit un séquençage de la périphérie droite en « éléments », rendant compte de la structure interne du segment périphérique droit. Quels que soient les segments repérés, ceux-ci sont considérés comme des ensembles dès lors qu'ils correspondent à l'une des quatre catégories de segments en périphérie droite que j'ai décrites dans le premier chapitre. La typologie correspond à celle que j'ai tirée de Blanche-Benveniste (1990, pp. 138-144) : *marqueurs discursifs et pragmatiques* ; *associés lexicaux ou pronominaux* ; *recteurs faibles* ; *verbes « défocalisés »*. Les « éléments » ainsi annotés sont toujours considérés selon leur forme la plus réduite. Le cas des marqueurs composés n'est ainsi analysé, à ce stade, que comme un phénomène de cooccurrence : l'hypothèse appliquée *a priori* lors de l'identification est celle de la conservation d'un sémantisme individuel et indépendant pour chaque terme cooccurent, sans préjuger du fait que certaines compositions de marqueurs fussent lexicalisées.

Certains groupes de rection sont réévalués comme des rections faibles ou comme marqueurs : les rections présentes en périphérie droite sont, lors du séquençage, évaluées en fonction de leur possibilité d'accueillir des modalités ; en cas de réponse négative, elles sont considérées comme des rections faibles. Les formes hautement récurrentes et présentant des indices manifestes de figement (Apothéloz 2003, pp. 243-249) sont alléguées à la catégorie des marqueurs – et parfois ré-identifiées rétroactivement après le constat de leur récurrence. Enfin, dans les structures de rection focalisées, l'annotation retient l'hypothèse d'une prononciation postfixe si le terme non focalisé du dispositif est inférieur à huit syllabes (cf. Morel 2007: 44).

3. TYPES ET FONCTIONS DES PÉRIPHÉRIES DROITES

La seconde étape de l’annotation des périphéries droites dans le corpus Valibel repose sur un étiquetage des segments préalablement identifiés. Un étiquetage est assigné à chaque élément périphérique observé, ce qui signifie qu’une périphérie droite complexe – constituée d’une série d’éléments – se verra attribuer une série de tags. Pour davantage de précision sur ces tags, l’application de la méthodologie et des exemples de formes annotées, je renvoie à la grille d’annotation en annexe (*Annexe II*).

3.1. TYPOLOGIE MICROSYNTAXIQUE

Le chapitre précédent a annoté quatre « corrélats microsyntaxiques » des périphéries droites (*section I.3.1*). Deux d’entre eux correspondent ainsi à des associés de la rection : (1) les marqueurs discursifs et pragmatiques et (2) les associés lexicaux et diverses formes de dislocation. Deux autres types de périphérie sont quant à eux constitués d’éléments recteurs : (3) les structures de rection faible et (4) les constructions verbales que je qualifie de « défocalisées ». Une couche d’annotation (*‘type’*) du corpus Valibel reprend ces quatre catégories de corrélats microsyntaxiques et les assigne à chaque élément identifié comme en périphérie droite. Méthodologiquement, l’identification de ces catégories repose *a priori* sur les catégories traditionnelles de la morphosyntaxe.

Un pronom tonique ou un syntagme nominal identifiés comme périphériques se verront systématiquement attribué un tag *‘asso’* les étiquetant comme des associés de l’unité de rection. Cette catégorie concerne les différents types de dislocation que j’ai mentionnés dans la partie précédente (*section II.2*), lesquels ne se limitent pas aux compléments obligatoires de la valence du verbe, mais s’étendent également à des compléments optionnels de la rection⁵⁰ ; les éléments étiquetés *‘asso’* comprennent ainsi, notamment, des éléments dont la valeur

⁵⁰ Pour rappel, des formes traditionnellement rangées dans la vaste catégorie des adverbes (*‘alors’*, *‘là-bas’*), sont considérés par le GARS comme des éléments à valeur pronominale sur le principe qu’ils peuvent occuper la même place de rection que les syntagmes nominaux.

sémantique est locative '*près de chez nous*', '*là-dedans*' ou temporelle '*cette année*' pour lesquels l'hypothèse d'une intonation d'appendice est à privilégier en raison de leur position ou des contraintes de focalisation de la structure de l'UR précédente.

Un élément de nature verbale ou correspondant à une unité de rection annotée est considéré soit comme une construction défocalisée, soit comme un recteur faible ; la frontière entre les deux catégories est décidée à partir d'une batterie de tests syntaxiques destinés à en mesurer la grammaticalisation : les recteurs faibles peuvent être employés comme des parenthétiques au milieu d'une unité de rection ; leur complétive n'est plus substituable à une proforme clitique ; ils restreignent leurs propriétés flexionnelles et la gamme des clitiques qu'ils régissent (voir Apothéloz 2003:243-249).

Dans l'acception comprise par cette grille d'annotation, la catégorie des « marqueurs et ponctuels » englobe des segments de différentes natures (adverbes, prépositions, interjections mais aussi des noms, verbes, etc.) que la littérature qualifie de diverses manières : marqueurs de discours, marqueurs pragmatiques, ponctuels,... Leur principale caractéristique commune est d'être « grammaticalisés » ou « pragmatiques » pour intervenir au niveau du discours ou de l'interaction. L'annotation de cette catégorie est établie à partir d'un inventaire de formes récurrentes⁵¹, enrichi lors de l'annotation par les nouveaux éléments rencontrés et possédant des caractéristiques similaires de grammaticalisation et de pragmatique⁵².

3.2. POSITIONS DANS LE TOUR DE PAROLE

L'objectif de cette annotation est de comprendre comment l'interaction structure la périphérie droite, et comment celle-ci structure à son tour l'interaction. Pour l'identification de ces

⁵¹ Cet inventaire de formes de marqueurs, , a été affiné suivant les passes d'annotation successives ; il est présenté en annexe de ce mémoire.

⁵² La forme '*tu vois*', en raison de son haut degré d'occurrence (190 tokens), est ainsi devenue la seule forme verbale considérée comme un marqueur : elle témoigne d'une grammaticalisation complète et affiche un degré d'occurrence similaire aux marqueurs les plus fréquemment rencontrés ('*quoi*' : 509 tokens ; '*hein*' : 439 tokens ; '*tu vois*' : 190 tokens ; '*mais*' : 151 tokens ; '*enfin*' : 142 tokens)

positions, je me réfère à Degand (2014) qui identifie trois positions possibles, à droite, pour les marqueurs de discours : « clause-final » (dans le domaine de la rection), « utterance-final » (structurant discursif) et « turn-final » (outil interactionnel). Seuls les deux dernières positions sont extérieures aux UR et par conséquent pertinents pour l'étude des périphéries.

Les segments en finale et tour et en finale d'énoncé ont donc été observés grâce à une couche d'annotation spécifique recensant les positions occupées, élément par élément, dans chaque tour. Un élément est considéré comme en « fin de tour » s'il est réalisé par le locuteur juste avant que celui-ci cède la parole. Dans d'autres cas, un élément comme en « fin d'énoncé », dans la mesure où son caractère final se traduit par le fait de borner une UR tout en faisant partie du même groupe intonatif⁵³. Dans de nombreux cas, un élément périphérique sera à la fois en finale d'énoncé et en finale de tour, ce qui conduira à une annotation reprenant les deux traits positionnels.

Certains éléments en périphérie droite ressortissent d'une autre catégorie de position : celle des éléments précédant une *afterthought* ou épexégèse, et celle des éléments situés eux-mêmes dans une *afterthought* – ce dernier cas, rarissime, comprend les cas dans lesquels une périphérie est formulée en deux temps, avec une démarcation prosodique des reformulations après-coup. Ces deux catégories, non présentes dans les premières versions de l'annotation, constituent des cas limites d'identification. La pertinence de leur rattachement au phénomène général des périphéries droites sera questionnée à travers les résultats de l'annotation.

3.3. ANNOTATION FONCTIONNELLE

L'analyse fonctionnelle est la dernière étape de l'annotation. Au niveau de l'annotation manuelle, il s'agit d'une étape difficile à opérationnaliser et largement conditionnée au jugement de l'annotateur. L'intérêt scientifique d'une telle annotation n'en est pas moins

⁵³ Cette position correspond *de facto* à une fin d'énoncé, dans la terminologie du GARS qui considère un énoncé comme l'ensemble d'un noyau et de ses affixes.

crucial pour autant : à partir d'un corpus de données annotées suffisamment grand, le développement de modèles prédictifs automatisés est en effet théoriquement envisageable.

Les théories du discours qui se sont intéressées au rôle joué par les périphéries en viennent bien souvent à opposer les fonctions que je qualifierai de *textuelles* et les fonctions *illocutoires*, en présentant généralement ces catégories comme mutuellement excluables. Danon-Boileau et al. (1991) différencient ainsi deux fonctions des « segments postrhématiques » l'*intégration discursive* de l'*intégration syntaxique*, lesquels s'excluent mutuellement. Detges & Walereit soutiennent la même opposition, mais la voient davantage comme un continuum entre un rôle de cohérence et un rôle modal. Une de leurs observations décisives est que ce dernier rôle « modal » serait pour eux plus fréquent en raison de la tendance des périphéries droites, liée à leur position dans la chaîne temporelle, à être les lieux d'une *renégociation* des actes de langage.

Les annotations de fonctions *textuelles* et *illocutoires* sont susceptibles de s'appliquer concomitamment au même élément, ce qui va permettre une prise en compte de la polyfonctionnalité de certaines formes – telles que certains marqueurs ayant à la fois une fonction de marqueurs de discours – textuelle – et une fonction de marqueurs pragmatique – illocutoire.

III. ANALYSE DES DONNÉES

Ce chapitre propose une analyse des données collectées lors de l’annotation. La démarche de description des données est avant tout quantitative dans un souci de produire des conclusions générales, utilisables dans des modèles destinés à l’implémentation.

Le chapitre est divisé en trois parties. Dans la première, je décris les résultats de la première étape de l’annotation : l’identification des périphéries droites (ici, RP). La deuxième partie s’intéresse aux relations entre l’interaction et les segments périphériques droits : sont considérées les données relatives aux positions occupées par les RP dans le tour de parole, fréquences relatives, complexité des éléments et enfin types formels de périphéries droites. Enfin, la troisième partie, la plus importante, investigate en profondeur la structure interne des périphéries droites : les longueurs relatives des RP ; les comportements positionnels des marqueurs et ceux des autres éléments ; les relations entre catégories formelles et fonctionnelles d’éléments en RP ; et enfin l’analyse des séquentialités et des possibles liens syntagmatiques entre catégories formelles et fonctionnelles.

1. DONNÉES DE L’IDENTIFICATION

La première étape de l’annotation manuelle consistait à identifier les segments périphériques à partir de deux entités opérationnalisées, les « unités de rection » (UR) et les « groupes intonatifs » (GI). Un segment périphérique droit (RP) est considéré comme l’ensemble des éléments à droite d’une unité de rection appartenant au même groupe intonatif.

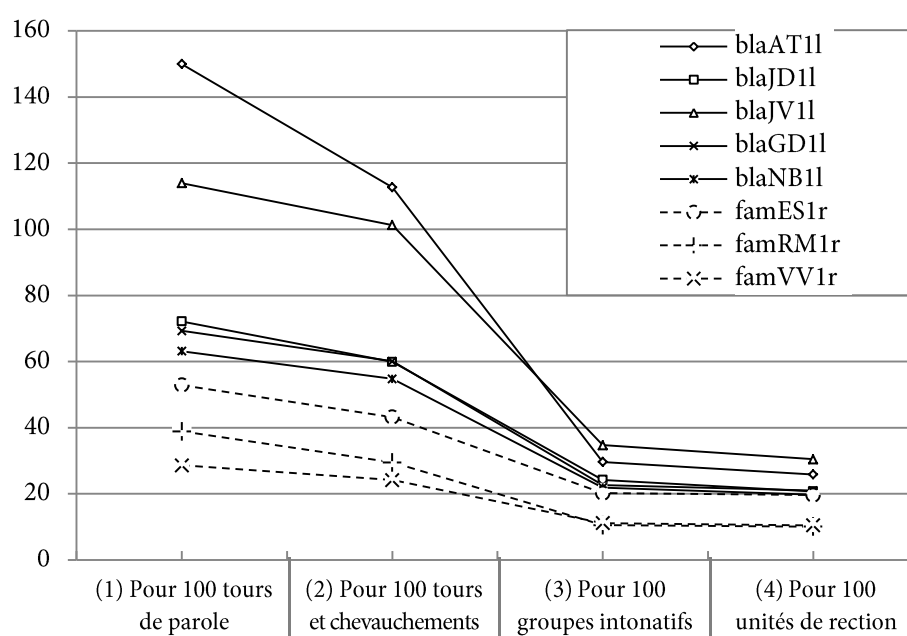
Cette démarche d’identification fournit de nombreuses données sur le plan qualitatif et quantitatif, qui permettent de mieux comprendre la nature des périphéries droites et leur relation aux autres entités de segmentation des discours. Dans cette partie, je m’intéresse en particulier à la distribution des périphéries droites à travers les différents corpus et transcriptions, pour montrer que la fréquence du phénomène varie largement d’une transcription à l’autre ; plusieurs facteurs explicatifs sont considérés.

Dans un deuxième temps, cette partie questionne l'heuristique d'identification des RP en analysant la notation des *incertitudes d'identification* dans le corpus. L'orientation est une analyse quantitative, repérant les cas les plus problématiques de l'annotation afin de circonscrire les généralisations ultérieures dans un cadre méthodologiquement adéquat.

1.4. FRÉQUENCE DES PÉRIPHÉRIES DROITES

L'annotation des corpus de conversations informelles *fam* et *bla* (ci-après « les deux corpus ») a identifié 2252 segments en périphérie droite. La fréquence d'occurrence de ces segments affiche des disparités notables d'une transcription à l'autre, et *a fortiori* d'un corpus à l'autre. Certaines transcriptions contiennent ainsi une proportion de RP largement supérieure à ce qu'en laisse présager leur taille, tandis que d'autres sont étonnamment pauvres en segments de ce type. Le contraste s'observe en particulier entre les deux corpus. Le corpus *fam*, reprenant des conversations dans le cadre familial, contient ainsi moitié moins de segments en RP que le corpus *bla*, moins spécifique dans ses contextes d'interaction.

Figure III.1 – Fréquence des segments RP dans les transcriptions



Le graphique ci-dessous montre les différences de fréquence des RP dans chacune des transcriptions des corpus *fam* et *bla*. Ces fréquences sont lissées en fonction de la proportion de RP pour 100 unités de segmentation, à partir de quatre mesures de segmentation

différentes – les droites joignant les points permettent de suivre chaque transcription d’une mesure à l’autre. Les tours de parole, avec et sans prise en compte des chevauchements, sont des données déjà présentes dans les transcriptions : pour la mesure (1), un tour de parole correspond à une ligne de transcription ; pour la mesure (2), la prise en compte des chevauchements de tours de parole donne un meilleur aperçu de l’« interactivité » de la conversation transcrite. Les mesures (3) et (4) ne sont pas issues de la transcription de base, mais évaluent la fréquence des RP à partir des groupes intonatifs⁵⁴ et les unités de rection⁵⁵, produits lors de l’annotation. La mesure (3) représente la proportion de RP pour 100 groupes intonatifs, tandis que la mesure (4), quant à elle, indique la proportion de RP pour 100 unités de rection.

De ce graphique se dégage un radical contraste, quelle que soit la donnée mesurée, entre les deux corpus – tous deux pourtant constitués de conversations informelles. Si l’on reprend la moyenne des transcriptions, cette proportion passe du simple au double. Plusieurs explications peuvent être avancées à cette disparité d’occurrence des RP. Si deux corpus relevant pourtant du même type de productions varient à ce point, la question de l’influence du contexte et des conditions de l’interaction sur l’occurrence des segments en périphérie droite se pose en premier lieu : les périphéries droites pourraient, par exemple, être très sensibles dans leurs occurrences au degré de familiarité ou de proximité entre les locuteurs impliqués. Un autre facteur explicatif serait celui de la divergence de certaines transcriptions par rapport à au reste du corpus auquel elles se rattachent. Ainsi, les transcriptions *blaAT11* et *blaJV11* contiennent beaucoup plus de RP par tours de parole que les trois autres

⁵⁴ Les groupes intonatifs opérationnalisés pour l’annotation sont décrits dans le chapitre précédent (II.2.2) et des exemples sont donnés dans la grille d’annotation (*Annexe II*). Dans le corpus *bla*, ceux-ci surviennent à une fréquence de 3,5 pour 1 tour de parole ; cette fréquence descend à 3 dans le corpus *fam*.

⁵⁵ Les unités de rection sont définies dans le chapitre précédent (II.2.1) et des exemples sont donnés dans la grille d’annotation (*Annexe II*). Le corpus *bla* contient une proportion de 3,9 unités de rection par tour, tandis que le corpus *fam* n’en contient que 3,1 par tour. Concrètement, cela signifie qu’un groupe intonatif sur 10 contiendra deux unités de rection dans le corpus *bla*, tandis que ce rapport est de 1 pour 20 dans le corpus *fam*.

transcriptions du corpus *bla*. Il s'agit ici d'un artefact lié aux choix du transcripteur d'y noter moins fréquemment les passations de tours de parole : ces transcriptions sont en effet caractérisées par de longs tours ininterrompus, ponctués de chevauchements d'une longueur conséquente. Dès lors, mesurer la fréquence des RP à partir de données motivées telles que les GI ou les UR, issues de l'annotation d'identification, prend tout son sens.

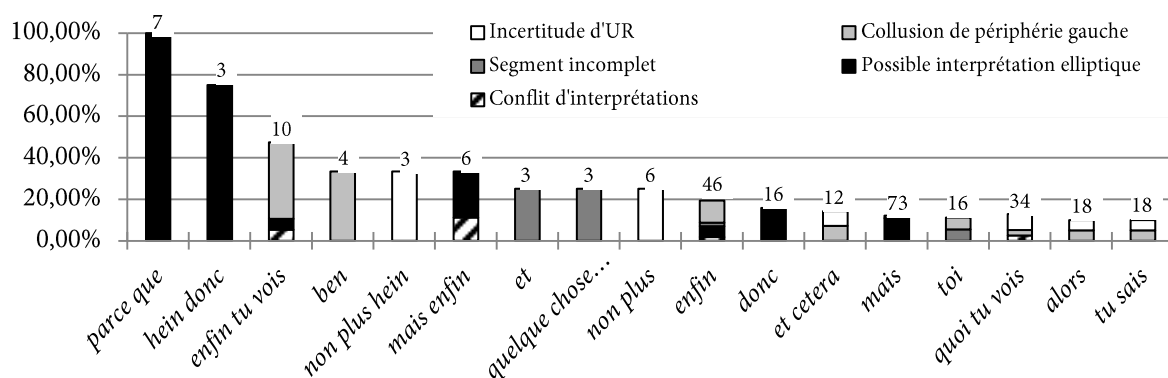
1.5. INCERTITUDES ET BIAIS D'IDENTIFICATION

La phase d'identification des périphéries droites, qui n'a été réalisée que par un seul annotateur et sans données de prosodie, a conduit à annoter les incertitudes d'identification. Plutôt que d'évaluer l'accord inter-annotateurs *a posteriori* à partir des codes attribués, c'est la difficulté *a priori* de prendre une décision d'annotation qui est ici mesurée.

Les segments annotés RP qui se sont vu attribuer un tag d'incertitude sont au nombre de 187 (17% des segments identifiés). Les cas recensés correspondent à des catégories différentes : les *incertitudes d'UR* [63 annotations] : les segments pour lesquels les unités de rections situées dans l'environnement du segment identifié ne peuvent pas être délimitées avec certitude ; la RP est annotée sous réserve que tout ou une partie de ses éléments ne sont pas dépendants d'une UR. D'autres incertitudes fréquentes sont des *collusions de périphérie gauche* [55 annotations] : dans de tels cas, les segments sont bel et bien indépendants de l'UR, mais l'indétermination porte sur le groupe intonatif (GI) parent : le segment observé est potentiellement en RP de l'UR précédente ou en LP de l'UR suivante. Les *interprétations elliptiques* sont présentes en nombre plus restreint [34 annotations] mais concernent de façon récurrente un inventaire restreint de formes. Enfin, quelques cas de *segments incomplets* [16 annotations] et de *conflits d'interprétations* [17 annotations] sont également recensés dans l'annotation.

Le graphe en colonnes ci-après reprend les dix-sept segments qui, de façon récurrente⁵⁶, ont suscité une incertitude d'identification. L'objectif, ici, n'est pas d'expliquer la raison de chaque incertitude, mais d'attirer l'attention sur des cas fréquents. J'attire donc l'attention du lecteur sur le fait que ce graphique ne propose pas un aperçu de l'ensemble des RP incertaines ; il recense les RP pour lesquels un contenu produit *de façon répétée* des incertitudes d'annotation. Dans ce graphique, chaque colonne représente, en pourcentages, la proportion de cas dans lesquels le segment aura été problématique pour l'annotation. Les différentes catégories d'incertitudes sont représentées par empilement dans chacune des colonnes ; les nombres au-dessus des colonnes représentent la fréquence absolue de tags annotés, à partir de laquelle est calculée la proportion en pourcentage.

Figure III.2 – Segments récurrents pour les incertitudes d'identification



La catégorie des possibles *interprétations elliptiques* [34 notations] se distingue par ses formes emblématiques : ‘*parce que*’, ‘*donc*’, ‘*hein donc*’, ‘*mais enfin*’, ‘*mais*’, qui se trouvent être tous les quatre composés de conjonctions dont la suite a été – potentiellement – omise⁵⁷. Ainsi, ‘*parce que*’, pour les quelque 7 occurrences dans lesquelles il se présente en périphérie droite d’une UR, suscite une possible interprétation elliptique dans tous les cas (100%). Deux formes sont à la fois régulièrement génératrices d’incertitudes et fréquentes dans les deux corpus étudiés :

⁵⁶ Je considère un problème d’identification comme récurrent lorsqu’il se présente dans au moins 10% des occurrences de la RP, pour un minimum de 3 incertitudes annotées.

⁵⁷ Ces segments ont pour point commun de tous reposer sur une conjonction qui, en périphérie droite, peut être employée de façon autonome comme marqueur pour ouvrir vers un sens partagé par les interlocuteurs et donc actif au niveau interpersonnel. L’interprétation elliptique, ici, serait en revanche de considérer que le locuteur conserve l’emploi initial de conjonction, et que la suite, si elle est *non verbalisée*, n’en est pas moins *verbalisable* et ne fait tout simplement pas l’objet d’une réalisation.

‘*donc*’ [16 notations d’incertitude] et ‘*mais*’ [73 notations ‘*iid*’]. L’examen de leur contexte d’apparition montre que ces cas d’ellipse surviennent surtout avant une pause pleine ‘*eah*’.

Les *collusions de périphérie gauche* et les *incertitudes d’UR*, les plus fréquentes [55 et 63 notations], sont moins rattachées à des segments typiques. Il existe des éléments récurrents : a suite ‘*enfin tu vois*’ de même que ‘*ben*’ posent régulièrement des difficultés lors de l’annotation en introduisant une confusion sur leur rattachement entre deux UR et, de là, une collusion de périphérie gauche⁵⁸. Les *incertitudes d’UR* connaissent également quelques formes emblématiques ; ici, ‘*non plus*’ et ‘*non plus hein*’ sont les plus remarquables, malgré une fréquence faible. L’examen détaillé des données révèle que les *collusions de périphérie* et les *incertitudes d’UR* sont associées en particulier à des hapax du corpus (par segments annotés, la médiane de leurs fréquences est à 1). La forte présence des *collusions de périphérie gauche* et des *incertitudes d’UR*, associées à la non-prévisibilité des formes problématiques, souligne l’importance des données d’intonation pour trancher ce type de cas limites.

Les *segments incomplets* et les *conflits d’interprétation* [16 et 17 notations] sont les moins représentés dans le corpus ; présentant peu de régularité au niveau des formes, ces deux types d’incertitudes sont à considérer au cas par cas.

Des données relatives aux incertitudes d’identification, on peut donc conclure que plusieurs segments sont susceptibles, de façon régulière, de biaiser le repérage des segments en périphérie droite. Les biais intervenant lors de l’identification sont importants, car ils impactent le reste de l’annotation et des résultats. Ainsi, les marqueurs de *hedging* prenant la forme de conjonctions (‘*parce que*’, ‘*mais*’, ‘*mais enfin*’, ‘*donc*’) ou les formes typiques de collusion de LP (‘*enfin tu vois*’, ‘*ben*’) devront être exclues de certaines généralisations de la description.

2. LES PÉRIPHÉRIES DROITES DANS L’INTERACTION

⁵⁸ Ces cas correspondent à des phénomènes de ce type :

[ouais] // [voilà] / ben / à mon avis [je crois] qu’[on peut] // [*arrêter là] (blaGD11:205)
Dans ces exemples, il est difficile de rattacher ‘*ben*’ autant à la deuxième UR du tour ‘*voilà*’ qu’à la troisième ‘*je crois*’ ; l’hypothèse RP est privilégiée, mais un tag d’incertitude d’identification ‘*iid*’ est ajouté.

Le fonctionnement dynamique des périphéries droites a été approché, lors de l'annotation, par un enregistrement des positions occupées par les périphéries dans le tour de parole ; cette notation est décrite dans le chapitre précédent (II.3.2). Ces positions liées à la structure de l'interaction sont de cinq groupes différents : les éléments en finale de tour de parole ('*ft*') ; les éléments en finale d'énoncé ('*fe*') ; les éléments simultanément en finale de tour et en finale d'énoncé ('*ff*') ; les éléments précédant ou amorçant une *afterthought* ou épexégèse ('*ba*') et enfin les éléments intégrés dans une '*afterthought*', tout en présentant les caractéristiques d'une périphérie droite ('*ia*'). Dans cette partie sont observées les fréquences de ces différents tags et leurs relations avec d'autres caractéristiques annotées des périphéries droites. La question posée par ces notations est en effet d'investiguer l'influence de l'interaction sur certaines caractéristiques de leur structure : les RP complexes, les types microsyntaxiques corrélés aux RP et enfin les fonctions jouées par celles-ci dans le discours et au niveau pragmatique.

2.1. FRÉQUENCES DE RP RELATIVES AUX POSITIONS DANS L'INTERACTION

Intéressons-nous tout d'abord aux fréquences de chacun de ces tags dans les deux corpus considérés. Les éléments en fin d'énoncé ('*fe*'), qui ne marquent pas la fin d'un tour de parole mais bien celle d'une unité macrosyntaxique, sont de loin les plus fréquents : les deux corpus en comptent 1267, parmi lesquels 984 se trouvent dans des RP simples (contenant un seul élément). En comparaison, les éléments en périphérie droite qui marquent la fin d'un tour de parole ('*ft*' et '*ff*') sont au nombre de 712 dans les deux corpus réunis ; ces éléments finaux sont significativement moins présents ($p = 0,014$) dans le corpus *fam* que dans le corpus *bla* ; cette différence est à relier à la plus grande densité de RP dans ce dernier corpus.

2.2. COMPLEXITÉ DES RP RELATIVE AUX POSITIONS DANS L'INTERACTION

Les données de la table ci-dessous montrent la répartition des différentes complexités observées de RP – de 1 à quatre éléments constitutifs – en fonction des différentes positions de l'interaction dans lesquelles elles peuvent survenir.

Table III.a – Position dans le tour et complexité des RP

Position de l'élément	Complexité de la RP				Total
	1 el.	2 els.	3 els.	4 els.	
Finale d'énoncé ('fe')	984	247	30	6	1267
Finale de tour					
en RP simple ('ff')	594	-	-	-	601
Finale de tour					
en RP complexe ('ft')	118	137	14	5	274
Précession d' <i>afterthought</i> ('ba')	74	10	0	0	84
Intégration en <i>afterthought</i> ('ia')	13	3	0	0	16

Les données de cette table suggèrent une distribution relativement lisse et proportionnée. Les seules anomalies de distribution concernent les RP internes aux tours de parole ('fe'), qui montrent une faible tendance à accueillir des RP plus courtes, alors que les finales de tour en RP simple ('ff') seront à l'inverse tendanciuellement plus complexes. Statistiquement, toutefois, la position des éléments n'exerce pas une influence significative sur le nombre d'éléments contenus dans les segments en RP : de fait, un test d'indépendance de la relation entre position finale de tour et nombre d'éléments en RP ne rejette pas l'hypothèse nulle ($p = 0,23$). Les données indiquent donc que les périphéries droites ne seront ni plus longues, ni plus courtes en fonction de leur position dans le tour de parole.

2.3. TYPES ET FONCTIONS RELATIVES AUX POSITIONS DANS L'INTERACTION

Les figures III.3 et III.4 ci-après montrent, pour chaque position annotée d'un élément dans le tour de parole, (1) la répartition des types morphosyntaxiques préalablement identifiés comme corrélats de RP et (2) la distribution des rôles fonctionnels joués par les RP aux niveaux discursif et métadiscursif.

Figure III.3 – Position et type de RP

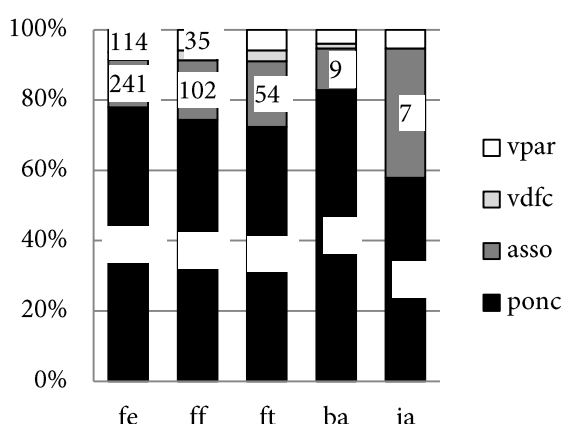
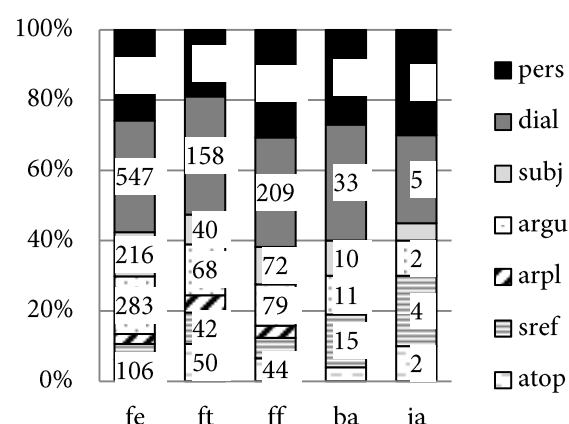


Figure III.4 – Fonction et type de RP



Le type microsyntaxique des éléments périphériques est en partie conditionné par leur position dans le tour de parole. La proportion de marqueurs de discours et de marqueurs pragmatiques (*ponc*) est ainsi plus importante à l'intérieur d'un tour de parole qu'à la fin de celui-ci ; les associés (*asso*), en comparaison, affichent justement la tendance inverse : ils sont plus fréquents en fin de tour de parole. Cette différence de répartition entre les marqueurs et les associés en fonction de leur position n'est pas aléatoire : un test de dépendance permet de rejeter l'hypothèse nulle et de montrer que la position dans le tour de parole est hautement significative ($p = 0,006$) pour déterminer la catégorie qui y sera la plus probable.

Les fonctions véhiculées par chaque élément périphérique sont également influencées par leur position dans le tour de parole. Les éléments en fin de tour, spécialement lorsqu'ils sont également en finale d'une RP complexe (*ff*), sont ainsi caractérisés par une plus grande fréquence des fonctions d'« ajout sémantique » : les anti-topics et les spécifications référentielles. Cette tendance est encore plus manifeste dans les périphéries droites précédant ou intégrant une *afterthought* (*ba* et *ia*), qui contiennent jusqu'à 10% d'anti-topics et 20% de spécifications référentielles. Les fonctions interpersonnelles (*pers*) sont également intéressantes à observer en fin de tour : dans des éléments isolés (*ff*), elles sont plus fréquentes qu'attendu, tandis qu'elles sont justement plus rares en finale de RP complexes ; cette différence est très significative dans les données observées ($p > 0,001$ pour la probabilité d'obtenir un tag *pers* et non un autre dans ces deux positions).

3. STRUCTURE INTERNE DES PÉRIPHÉRIES DROITES

3.1. LONGUEUR RELATIVE DES SEGMENTS EN RP

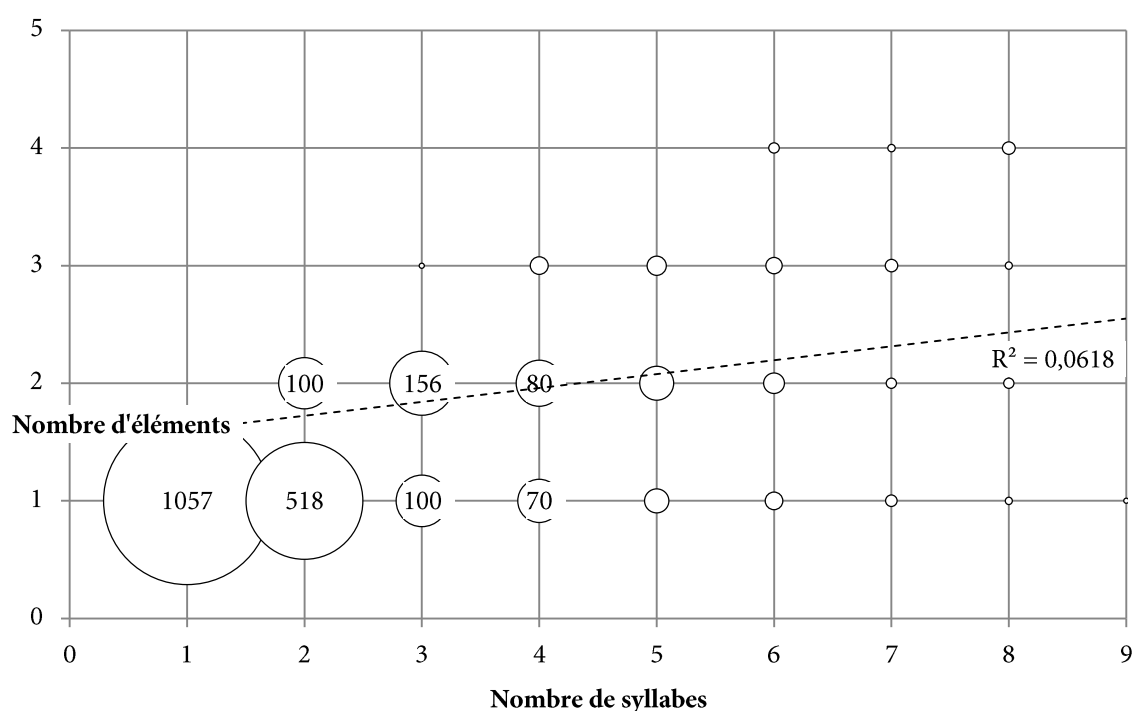
Mesurer la longueur des segments en périphérie droite peut se faire à partir de deux mesures de longueur pertinentes en regard du phénomène concerné. D'une part, l'annotation produit un découpage des segments en RP en « éléments » qui forment des unités de sens dans les RP et peuvent être considérées comme des formes lemmatiques⁵⁹. Une autre mesure de la longueur des périphéries droites est celle de leur longueur en syllabes, pertinente sur le plan métrique pour prédire les régularités de regroupement intonatif⁶⁰. D'après les propositions de Morel (2007), un choix méthodologique pour l'annotation a consisté à limiter les segments à huit syllabes en cas d'hésitation sur les autres critères. L'hypothèse est qu'au-delà, les règles métriques du français contraignent à la formation de deux unités intonatives distinctes.

La figure ci-dessous montre la relation entre, en abscisses, le nombre de syllabes et, en ordonnée, le nombre d'unités qui constituent le segment en périphérie droite. Des cercles symbolisent la fréquence d'individus de RP affichant des valeurs identiques en x et en y .

⁵⁹ Certaines seront considérées comme des *multiword units* ('MWU') par des lemmatiseurs automatiques. L'annotateur DisMo, par exemple, analyse 'tu vois' comme une seule unité.

⁶⁰ La longueur en syllabes mesurée ici ne reflète pas la réalité des formes effectivement prononcées. Dans le corpus Valibel, en effet, les transcriptions ne cherchent pas à reproduire fidèlement les réalisations phonologiques. Bien au contraire : le choix de la banque de données Valibel est celui d'une unification des graphies indépendamment des variations de prononciation, pour approcher une correspondance 1:1 entre la graphie et les morphèmes. À un type morphologique correspond donc une forme (ortho)graphique. Cela implique notamment que la réduction phonologique, pourtant intéressante dans l'étude des périphéries discursives, ne peut être observée directement depuis la transcription. Ainsi marqueur 'enfin' ne sera pas différencié en fonction de ses réalisations courantes /ãfẽ/, /fẽ/, /fɛ/ ; mimétiquement à sa graphie standardisée reproduite dans les transcriptions, il sera toujours considéré comme bi-syllabique.

Figure III.5 – Nombre d'éléments en RP en fonction du nombre de syllabes



La relation entre le nombre de syllabes et le nombre d'éléments contenus dans une périphérie droite affiche une corrélation modérément forte ($r_{\text{Pearson}} = 0,645$), mais toutefois hautement significative ($p = 0,004$) à partir d'une analyse des variances. La droite de régression indique qu'on peut s'attendre à ce qu'une RP comprenne deux éléments plutôt qu'un seul à partir du moment où elle est constituée de 4 syllabes ou plus. La limite estimée à huit syllabes se vérifie, mais présente deux exceptions : un individu est constitué de 9 syllabes et de 1 élément, tandis qu'un autre, non représenté dans le graphique, est constitué de 11 syllabes et 2 éléments.

Ce graphique montre également la tendance forte des périphéries droites à être composée de segments très courts : avec jusqu'à deux syllabes et un seul élément, 1575 individus sont couverts, soit 70% des segments RP annotés dans les deux corpus⁶¹. À ce titre, on peut noter les distributions respectives du nombre d'éléments et du nombre de syllabes : une RP contient en moyenne 1,24 élément, pour une médiane à 1, avec un écart-type de 0,5 élément et un coefficient d'asymétrie γ de 2,31. En comparaison, elle contiendra en moyenne 2,05 syllabes

⁶¹ Notons encore que cette caractéristique est liée à la non prise en compte des marqueurs composés.

pour une médiane à 2 syllabes, avec un écart-type de 1,4 syllabe et un coefficient d'asymétrie de 1,76. La distribution du nombre de syllabes est donc non seulement plus étalée, mais sa nature fortement asymétrique conduira à observer des RP très courtes. Cette plus forte asymétrie de la distribution syllabique, associée à une régression relativement importante des données, suggère qu'une tendance à un faible nombre de syllabes en RP ne contraindrait pas la taille des éléments, tandis que l'inverse serait vrai. En d'autres termes, la RP serait un environnement propice aux phénomènes de réduction syllabique ; cette indication, cependant, demande à être vérifiée à partir des données phonologiques liées aux transcriptions.

3.2. FRÉQUENCES ET POSITION DES ÉLÉMENTS EN RP

Cette partie recense, dans les périphéries droites, les éléments les plus fréquemment annotés en tant qu'unités autonomes – les compositions d'éléments, à ce stade, ne sont pas encore prises en compte. La colonne des occurrences totales affiche la fréquence de chaque élément dans le corpus et la proportion de cet élément par rapport à l'ensemble des formes observées. Quatre colonnes distribuent ces fréquences absolues relativement aux différentes positions occupées par l'élément dans les segments périphériques ; les pourcentages liés représentent la proportion d'occurrences dans cette position par rapport aux occurrences totales de la forme.

Les positions sont de quatre types : la position *isolée* concerne les éléments qui forment à eux seuls un segment périphérique et couvre, au total, deux tiers des formes [66,7% des els.] ; la position *initiale* porte sur les éléments observés en début de RP, juste après la fin de l'unité de rection [1,9% des els.] ; la position *médiale*, la plus rare, ne survient que dans les RP comprenant au moins trois éléments [14,8% des els.] ; la position *finale* recense les éléments qui se trouve à la fin d'un segment RP et, en conséquence, en fin d'énoncé [14,8% des els.].

Table III.b – Usage positionnel d'éléments fréquents en RP

Élément RP	Usage positionnel observé				Occurrences totales
	isolé	initial	médial	final	
<i>quoi</i>	360 (70,5%)	120 (23,5%)	11 (2,2%)	20 (3,9%)	509 (22,4%)
<i>hein</i>	319 (72,5%)	49 (11,1%)	6 (1,4%)	66 (15%)	439 (19,3%)
<i>tu vois</i>	93 (48,7%)	6 (3,1%)	8 (4,2%)	84 (44%)	190 (8,4%)
<i>mais</i>	83 (55%)	21 (13,9%)	5 (3,3%)	42 (27,8%)	151 (6,7%)
<i>enfin</i>	57 (40,1%)	35 (24,6%)	10 (7%)	40 (28,2%)	142 (6,3%)
<i>là</i>	68 (69,4%)	18 (18,4%)	1 (1%)	11 (11,2%)	98 (4,3%)
<i>moi</i>	45 (67,2%)	8 (11,9%)	5 (7,5%)	9 (13,4%)	67 (3%)
<i>ça</i>	49 (77,8%)	8 (12,7%)	1 (1,6%)	5 (7,9%)	63 (2,8%)
<i>aussi</i>	33 (55,9%)	21 (35,6%)	1 (1,7%)	4 (6,8%)	59 (2,6%)
<i>quand même</i>	23 (46,9%)	15 (30,6%)	2 (4,1%)	9 (18,4%)	49 (2,2%)
<i>et tout ça</i>	26 (60,5%)	17 (39,5%)	0	0	43 (1,9%)
<i>je crois</i>	28 (66,7%)	13 (31%)	0	1 (2,4%)	42 (1,9%)
<i>donc</i>	19 (46,3%)	1 (2,4%)	0	21 (51,2%)	41 (1,8%)
<i>comme ça</i>	26 (81,3%)	4 (12,5%)	0	2 (6,3%)	32 (1,4%)
<i>tu sais</i>	20 (64,5%)	4 (12,9%)	2 (6,5%)	5 (16,1%)	31 (1,4%)
<i>alors</i>	20 (74,1%)	6 (22,2%)	0	1 (3,7%)	27 (1,2%)
<i>ou quoi</i>	21 (87,5%)	3 (12,5%)	0	0	24 (1,1%)
<i>en fait</i>	17 (85%)	2 (10%)	0	1 (5%)	20 (0,9%)
<i>dis</i>	10 (55,6%)	0	0	8 (44,4%)	18 (0,8%)
<i>etcetera</i>	15 (88,2%)	2 (11,8%)	0	0	17 (0,7%)
<i>et tout</i>	13 (86,7%)	2 (13,3%)	0	0	15 (0,7%)
<i>Tot. els. courants</i>	1451 (66,7%)	369 (16,8%)	53 (1,9%)	397 (14,6%)	2270
<i>Tot. corpus</i>	1783 (64,1%)	469 (16,9%)	61 (2,2%)	469 (16,9%)	2782

En termes de fréquences absolues des lemmes, cette table montre qu'une classe excède de loin toutes les autres : celle des marqueurs discursifs et pragmatiques. Ainsi, les cinq lemmes '*quoi*', '*hein*', '*tu vois*', '*mais*', et '*enfin*', exclusivement utilisés comme marqueurs en RP, représentent, ensemble, 63% de tous les éléments présents en RP. Les deux premiers associés, '*moi*' et '*ça*' sont loin derrière, constituant seulement 5,8% de l'ensemble des formes. La première réaction faible qui connaisse une récurrence relativement importante est la forme '*je crois*', forte de 42 occurrences ; précisons ici que ce fait est lié au choix de considérer '*tu vois*' comme un marqueur plutôt que comme une entité de réaction.

Plusieurs observations notables peuvent être faites concernant les positions de ces formes les plus fréquentes en périphérie droite ; je me contenterai ici de 6 phénomènes portant sur les marqueurs, particulièrement révélateurs de certaines régularités liées aux formes observées.

(1) Le marqueur '*quoi*', tout d'abord, a pour caractéristique d'être omniprésent au début des RP complexes. Il a une légère tendance à survenir au début de ce type de RP (23,6% de ses occurrences, soit +0,56 écarts-types [σ] au-dessus de la moyenne des autres formes). Ce fait peut s'expliquer par les nombreuses cooccurrences de marqueurs dont il est le premier terme : '*quoi tu vois*' (50 coocc.), '*quoi mais*' (30 coocc.), '*quoi enfin*' (14 coocc.), '*quoi donc*' (10 coocc.). Le même marqueur '*quoi*' est en revanche rare en finale de RP complexe (3,7% de ses occurrences, -0,68 σ).

(2) Le marqueur '*tu vois*', contrairement à '*quoi*', n'est presque jamais en initiale de RP complexes (3,2% de ses occurrences, -1,25 σ) et survient très fréquemment en finale de RP (44,2% de ses occurrences, 1,75sig au-dessus de μ). Parmi les marqueurs les plus fréquents, c'est celui qui affiche, dans les corpus étudiés, le plus grand nombre d'emplois isolés (48% de ses occurrences, +0,39 σ).

(3) La forme '*enfin*' se démarque curieusement par un usage très fréquent en position médiane, toutes proportions gardées (10 cas recensés parmi 45 RP à trois éléments ; 7% de ses occurrences, 2,35sig au-dessus de μ).

(4) Les marqueurs de type '*general extenders*' ('*et tout ça*', '*etcetera*', '*et tout*') montrent eux aussi une tendance à se trouver en début de périphérie. Cette tendance se comprend facilement par leur fonction d'extension, les amenant à suivre directement la forme à laquelle ils se rapportent, laquelle sera bien souvent située juste avant la RP.

(5) Malgré une différence de fréquence, '*mais*' et '*donc*' affichent un comportement similaire : leurs emplois majoritaires sont soit isolés (55% et 46% de leurs occurrences respectives), soit en finale de RP complexes (28% et 51% de leurs occurrences respectives). Notons que dans presque tous les cas annotés, ils sont identifiables en RP à une forme particulière de '*hedging*' qui invite l'allocutaire, par l'incomplétude qu'ils introduisent, à inférer un présupposé. Tant en emploi isolé qu'en finale de RP complexe, ils sont d'ailleurs souvent suivis par une pause

pleine : *'mais euh'* représente ainsi près de 18% des *'mais'* finaux ou isolés (22 occ.); *'donc euh'* représente 37% des *'donc'* finaux ou isolés (23 occ.).

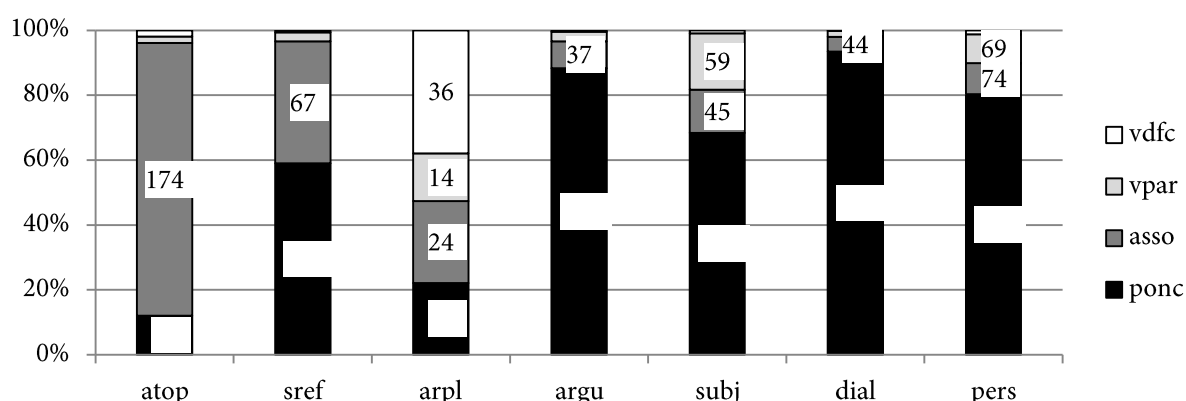
(6) Le marqueur *'hein'* se démarque par une grande fréquence d'occurrence qui, pourtant, ne semble indiquer aucune position spécifique. Certes, une préférence se dégage pour un emploi isolé, à associer à un usage relativement moindre en initiale de RP complexe ; mais c'est surtout le comportement du marqueur, réparti presque uniformément entre les positions, qui mérite d'être souligné. Tout indique que ce marqueur participerait d'un emploi indépendant, qui agencerait une articulation purement paratactique.

Au-delà de ces phénomènes observés pour des cas de marqueurs périphériques, la prise en compte des positions au sein des segments en périphérie droite montre que les éléments présents n'occupent pas des positions aléatoires ($p < 0,001$). Les marqueurs, qui sont de loin la catégorie la plus fréquente d'éléments en RP, ne présentent pas une tendance générale à certaines positions ($p < 0,001$), mais affichent, individuellement, des distributions qui suggèrent l'influence de contraintes syntagmatiques.

3.3. TYPE ET FONCTION DES ÉLÉMENTS

Une question cruciale de l'annotation réalisée pour ce mémoire est de comprendre comment la périphérie droite agence les fonctions textuelles, discursives et interactionnelles, et comment ces fonctions, à leur tour, sont traduites par des éléments formels identifiables. Le graphique ci-dessous montre, pour chaque fonction annotée en périphérie droite, les différentes catégories formelles de RP qui les matérialisent dans le discours.

Figure III.6 – Distribution des catégories formelles de RP d'après les fonctions annotées



Comme le montre ce graphique, les catégories formelles de périphéries droites ne sont pas uniformément distribuées à travers les fonctions discursives. Face à la quasi-omniprésence des marqueurs discursifs et pragmatiques en périphérie droite (*'ponc'*), deux catégories de fonctions se dégagent : celles qui les convoquent très largement des marqueurs discursifs et celles qui convoquent également d'autres catégories. Dans le premier groupe, trois classes de fonctions reposent majoritairement sur des marqueurs : les éléments intervenant dans la structure argumentative du discours (*'argu'*), les éléments valeur illocutoire subjective (*'subj'*) et les éléments à valeur illocutoire intersubjective, de type dialogique (*'dial'*) ou interpersonnel (*'pers'*). Dans le second groupe : les *anti-topics* (*'atop'*), les *spécifications référentielles* (*'sref'*) et les *marquages d'arrière-plan* (*'arpl'*) ; trois fonctions qui jouent un rôle aux niveaux textuel et référentiel.

Lorsqu'on s'intéresse aux fréquences absolues, une autre information importante est illustrée par ce graphique : plus une fonction est représentée dans les conversations, et plus elle comprendra une proportion élevée de marqueurs. Ainsi, la fonction « dialogique » (*'dial'*) est la première fonction annotée avec non moins de 960 notations sur 2773 éléments (soit 34,6% des éléments en RP) ; elle est réalisée dans plus de 90% des cas par des marqueurs de discours ou des marqueurs pragmatiques (*'ponc'*). Les autres catégories de fonctions suivent cette tendance : plus une fonction est fréquente, et plus forte est la présence des marqueurs. Seule la catégorie des *anti-topics* tranche résolument avec cette tendance : sur 207 formes observées, 84% (174) sont des associés lexicaux ou pronominaux.

Au niveau des catégories formelles, ce graphique montre que les *associés* jouent principalement un rôle d'*anti-topic* ('*atop*') [37,4% des formes] : ils introduisent, après l'énoncé, le topic informationnel qui permet de situer l'énoncé dans le discours et de construire la cohérence. L'intérêt traditionnellement porté, dans la littérature, au phénomène de dislocations à droite, trouve donc une justification dans les données présentées ici. Il ne doit cependant pas occulter d'autres fonctions jouées par les associés, notamment au niveau subjectif ou intersubjectif ('*subj*', '*dial*' et '*pers*'), qui concernent 35% des formes.

L'une des catégories formelles de RP est relativement peu courante : celle des *verbes défocalisés* ('*vpar*') Hormis de rares cas, cette catégorie a toujours été annotée comme un marquage d'arrière-plan : cette tendance se laisse comprendre à partir de l'outillage même qui a permis de les identifier : l'étude des processus de grammaticalisation. Les verbes défocalisés, faiblement grammaticalisés, peuvent difficilement jouer un rôle informationnel ou discursif au-delà de l'énoncé auquel ils se rapportent et sont donc cantonnés au rôle « par défaut » de la périphérie droite : celui d'amener une information secondaire, conformément à une routine routine [*action* ▷ *confirmation*].

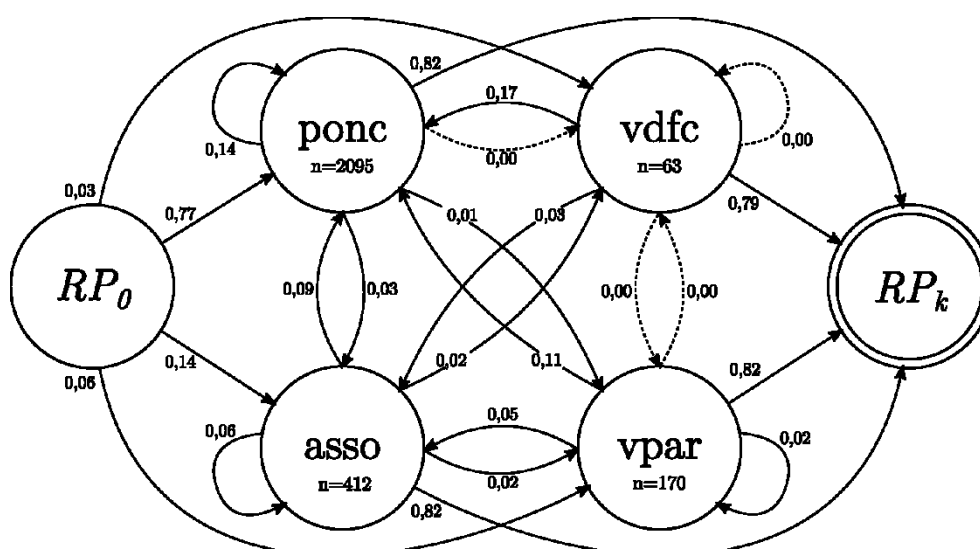
3.4. SÉQUENCES DE CATÉGORIES FORMELLES

Pour comprendre plus profondément la structure interne des périphéries droites, intéressons-nous aux suites d'éléments observées dans le corpus. Cette étude des séquences se fera d'abord au niveau des catégories formelles qui ont été annotées, et ensuite au niveau des catégories fonctionnelles. Il est possible de représenter les suites de catégories en consignant, à partir des données observées, la fréquence de transition d'une catégorie vers une autre.

Le graphe de transitions ci-après, inspiré du modèle stochastique des chaînes de Markov, représente et généralise la fréquence avec laquelle, dans les données annotées, chaque élément en suivra un autre, sous la forme de proportions des fréquences observées F_{obs} telles que $0 < F_{obs} < 1$. Le point de départ du schéma est donc la balise $< rp >$, qui initie dans le corpus une périphérie droite annotée et est représentée ici par le symbole RP_0 . Une fois qu'on obtient une catégorie donnée pour le premier élément, la question est de savoir avec quelle fréquence cet

élément s – le dernier – le graphe conduit alors immédiatement vers la balise fermante $\langle rp \rangle$ – représentée par le symbole RP_k ou bien, alternativement, avec quelle fréquence un élément successif se présentera – le graphe mène alors vers une autre catégorie, pour laquelle se pose à nouveau la question des fréquences⁶². Ci-après, chaque flèche représente la transition d’une catégorie vers une autre avec, associée à celle-ci, la probabilité de cette transition à partir des données observées ; chaque nœud représente une catégorie formelle d’élément figurant dans l’annotation. Les flèches en pointillés indiquent une transition entre catégories non attestée dans les données.

Figure III.7 – Fréquence des transitions entre catégories formelles d’éléments dans une RP



Une particularité des graphes de transitions du type des chaînes de Markov est de concevoir des inférences probabilistes sur les chaînes d’éléments : de la fréquence observée d’une transition est extrapolée une probabilité de transition d’un élément vers un autre. De là, une certaine chaîne de marqueurs acquiert une probabilité d’occurrence égale au produit des probabilités de transitions respectives d’élément en élément. À partir de ce diagramme de transitions, on peut ainsi prédire, d’après les données observées, qu’une RP de 4 éléments aura

⁶² Cette visualisation des données donne une meilleure compréhension du fonctionnement des séquences d’éléments, mais comporte également ses limites : elle ne contient pas l’historique des itérations précédentes, ce qui pose un problème d’approximation dès la seconde transition du graphe, qui reste ignorante de la transition précédente. Les relations entre deux éléments ou plus sont de ce fait difficilement expliquées.

0,1% de chances de se produire et que, dans ces cas de figure, une suite de marqueurs est la plus probable. Similairement, une suite $vpar \succ asso \succ ponc$ se produira donc avec une probabilité induite P telle que $P(el_{ponc}|el_{asso}|el_{vpar}) = 0,00033$, soit dans un cas sur 30030. Les chances d'observer un tel phénomène dans le corpus considéré auraient donc été de 9,26% si on s'en réfère à cette modélisation des données ; une étude des marqueurs complexes désireuse de produire des résultats quantitativement significatifs pour les périphéries droites aura donc besoin de recourir à des corpus substantiellement plus importants que les deux corpus étudiés dans le cadre de ce mémoire.

Ces graphes de transitions est également un bon indicateur des comportements séquentiels de chaque catégorie. Un fait tangible est ainsi l'importance des marqueurs de discours '*ponc*' dans la distribution des transitions. Au début d'une RP, un marqueur de discours est présent dans plus de 3 cas sur 4 ; la catégorie a ensuite 82% de chances d'être terminale et 14% de chances de se répéter, tandis qu'elle n'a que 4% de chances d'être suivie d'une autre catégorie. Ce constat permet de faire le lien entre plusieurs observations réalisées plus haut : les marqueurs sont à la fois très nombreux et fortement présents en initiale de RP ; ils sont souvent isolés. À ces confirmations d'observations, on peut ajouter qu'il est rare que des marqueurs introduisent d'autres catégories d'éléments : malgré leur fréquence, ils ne sont jamais, dans le corpus, suivis par un verbe défocalisés ('*vdfc*'), et très rarement par un recteur faible ('*vpar*').

Au niveau des successions de catégories, certaines tendances peuvent être constatées ; les successions des différents types formels de RP varient toutefois largement en fonction de leur distribution générale. Si les marqueurs tendent ainsi à succéder aux recteurs périphériques ('*vpar*' et '*vdfc*'), cela n'influence nullement la probabilité qu'ils soient terminaux ; les transitions correspondent aux fréquences respectives de chaque type d'élément. Le début d'une RP est lui aussi tout à fait arbitraire : la distribution des transitions établissant les éléments initiaux correspond aux fréquences globales. Elle montre cependant une préférence envers les associés et une contre-préférence envers les marqueurs. D'autres données

Table III.c – Fréquence des transitions entre catégories fonctionnelles d'éléments dans une RP

	<i>atop</i>	<i>sref</i>	<i>arpl</i>	<i>argu</i>	<i>subj</i>	<i>dial</i>	<i>pers</i>	<i>RP_k</i>
<i>RP₀</i>	0,068 (+0,001)	0,06 (+0,002)	0,031 (+0,002)	0,156 (+0,007)	0,117 (+0,004)	0,335 (+0,021)	0,233 (-0,038)	-
<i>atop</i>	0,022 (+0,01)	0,049 (+0,038)	0,022 (+0,017)	0,036 (+0,007)	0,022 (+0,001)	0,063 (+0,003)	0,04 (-0,011)	0,746 (-0,061)
<i>sref</i>	0,022 (+0,008)	0,026 (+0,015)	0,016 (+0,01)	0,031 (+0,003)	0,016 (-0,006)	0,084 (+0,024)	0,052 (+0,001)	0,754 (-0,053)
<i>arpl</i>	0 (-0,013)	0,01 (-0,001)	0 (-0,006)	0,01 (-0,018)	0,021 (0)	0,135 (+0,076)	0,01 (-0,041)	0,813 (+0,005)
<i>argu</i>	0,01 (-0,002)	0,01 (-0,001)	0,004 (-0,001)	0,044 (+0,016)	0,027 (+0,006)	0,1 (+0,04)	0,104 (+0,053)	0,7 (-0,107)
<i>subj</i>	0,011 (-0,002)	0,008 (-0,003)	0 (-0,006)	0,027 (-0,001)	0,038 (+0,017)	0,049 (-0,01)	0,162 (+0,11)	0,704 (-0,103)
<i>dial</i>	0,017 (+0,004)	0,008 (-0,003)	0,005 (-0,001)	0,017 (-0,012)	0,026 (+0,005)	0,036 (-0,023)	0,109 (+0,058)	0,782 (-0,025)
<i>pers</i>	0,012 (-0,001)	0,006 (-0,005)	0,003 (-0,002)	0,033 (+0,004)	0,008 (-0,013)	0,037 (-0,022)	0,062 (+0,011)	0,839 (+0,032)

Un constat générale qui ressort de ce tableau est que les fréquences de transition observées varient relativement peu par rapport aux fréquences attendues dans une hypothèse de distribution aléatoire. Sur un ensemble de 63 transitions possibles, 48 présente un écart inférieur à 1% [$F_{obs} - F_{th} > 0,01$], ce qui suggère, en général un agencement aléatoire des transitions entre fonctions telles que l'annotation en a rendu compte.

Quelques exceptions notables sont cependant à signaler. Alors que peu de restrictions contextuelles de transitions sont observables, deux sont particulièrement fortes : les éléments à fonction *argumentative* ou *subjective* conduisent beaucoup moins souvent vers une finale de RP [RP_{fin}] que les autres éléments, et que ce que leur distribution laisserait présager [resp. $F_{obs} - F_{th} = -0,107$; $F_{obs} - F_{th} = -0,103$].

Les autres observations majeures concernent des distributions qui se présentent comme contextuellement favorisées. Ainsi, une tendance observée est celle d'une transition favorisée des anti-topics vers les spécifications référentielles [$F_{obs} - F_{th} = +0,038$], qui montre encore une fois que ces fonctions, proches, sont bien souvent complémentaires.

Les éléments marquant une fonction *interpersonnelle* sont également intéressants dans leur comportement individuel. D'une part, ils ont fortement tendance à être les éléments finaux d'une périphérie droite [$F_{obs} - F_{th} = +0,032$], alors qu'ils sont plus rarements initiaux [$F_{obs} - F_{th} = -0,038$] ; d'autre part, ces mêmes éléments interpersonnels sont en grande majorité précédés par des éléments à fonction *argumentative* [$F_{obs} - F_{th} = +0,053$] ou *dialogique* [$F_{obs} - F_{th} = +0,058$]. Ces observations indiquent que les fonctions interpersonnelles auraient un rôle de clôture, lié au fait qu'elles interviendraient souvent en combinaison avec des éléments régulateurs de l'interaction. Ce constat est étonnant dans la mesure où la *section 2.3* de ce chapitre a montré que les fonctions interpersonnelles sont moins fréquentes en fin de tour ('*ft*') que les fonctions dialogiques.

Les fonctions de type dialogique ('*dial*') ou interpersonnel ('*pers*') sont intéressantes dans leurs interrelations. Ainsi, une tendance forte observée à porter sur des éléments initiaux de RP [$F(el_{dial} | RP_0) = 0,335$; $F(el_{pers} | RP_0) = 0,233$] ; les fonctions interpersonnelles sont largement moins présentes qu'attendu, alors que le contraire est observé pour les fonctions dialogiques. Les séquences faisant intervenir les deux fonctions sont également intéressantes : dans une RP complexe, les éléments à fonction *dialogique* précèdent très souvent des éléments possédant une fonction interpersonnelle, alors que l'inverse n'est pas vrai. Formulé autrement, cela signifie qu'un élément régulateur de l'interaction aura tendance à donner lieu, dans l'élément suivant, à une adresse personnelle à l'interlocuteur.

De façon plus forte que ce qui a pu être constaté au niveau des catégories formelles, la structure fonctionnelle des RP, montre des régularités statistiques qui témoignent d'une forte structure sous-jacente, et cela spécialement au niveau des deux fonctions de type « intersubjectif » : les éléments à fonction dialogique et les éléments à fonction interpersonnelle. Cette donnée est importante pour comprendre, d'une manière générale, le rôle fonctionnel propre des périphéries droites.

IV. PERSPECTIVES D'IMPLÉMENTATION

Dans ce chapitre sont proposés deux modèles de traitement susceptibles de garantir une identification efficace des segments en périphérie droite à partir de données transcrites. Le problème spécifique de traitement fait l'objet d'une brève description. Ensuite, deux approches sont proposées pour l'implémentation. D'un côté, l'analyse syntaxique en dépendances est orientée vers un traitement riche, et proche de la démarche d'annotation ; de l'autre, l'analyse statistique à partir de champs conditionnels aléatoires (*'conditional random fields'*, CRF) est une approche intéressante pour parvenir à un modèle robuste et capable de considérer un input multifactoriel.

1.6. DESCRIPTION DU PROBLÈME

Si la finalité de la détection automatique des périphéries droites est une analyse linguistique, leur détection automatique peut se comprendre, en termes de traitement et d'output, comme deux tâches différentes et en pratique complémentaires : l'extraction d'information et la catégorisation d'entités.

1.6.1. EXTRACTION D'INFORMATIONS

Dans le cadre d'une extraction d'information, chaque périphérie droite repérée dans une transcription, consiste en une séquence qu'il s'agit de restituer comme un ensemble structuré de données :

- le contenu textuel du segment identifié, en texte plain ou enrichi d'une analyse (p. ex. de type *'parts-of-speech'*) ;
- le contexte d'occurrence du segment : l'identifiant unique du tour de parole et des sous-unités parentes du segment RP ; un concordancier contenant les n mots du cotexte droit et gauche ; le texte des segmentations pertinentes auxquels la RP se rattache (unité de rection, groupe macrosyntaxique, tour de parole) ;
- la subdivision de la RP en éléments constitutifs ;

- les informations relatives à ces éléments constitutifs : unité mono- ou polylexicale, nombre de syllabes, données positionnelles relatives à l'interaction, catégorie formelle (p. ex. 'ponc', 'asso', 'vpar', 'vdfc'), catégorie fonctionnelle (p. ex. 'atop', 'sref', 'arpl', 'argu', 'subj', 'dial', 'pers').

Si l'identification des RP est comprise comme une tâche d'extraction, elle a donc pour objectif de constituer une base de données de périphéries droites – texte et informations – avec pour finalité d'en permettre une analyse linguistique fine.

1.6.2. CATÉGORISATION

Outre une extraction d'information, la conservation du matériel initial et de l'analyse automatique sont deux compléments souhaitables. Formellement, le problème de l'identification peut être présenté comme celui d'une catégorisation. D'une façon similaire à l'étiquetage en *parts-of-speech* (POS) on cherche, étant donné une forme et sa position, à calculer une étiquette qui correspondra, dans ce cas-ci, à l'étiquetage d'un élément comme *RP* ou *non-RP*. Cet étiquetage peut être raffiné par une ou plusieurs étiquettes catégorisant l'élément du point de vue morphologique ou fonctionnel.

Dans cette perspective, l'identification des périphéries droites est un problème de *classification* qui consiste à discriminer les segments RP dans les chaînes textuelles. Cette classification connaît deux sous-problèmes :

- le repérage du début d'une périphérie droite – détecter quand s'achève la rection et la syntaxe des constituants et quand commence un segment en périphérie droite ;
- le repérage de la fin d'une périphérie droite – discriminer une RP d'une nouvelle unité de rection, d'une afterthought, d'une LP.

Il s'agit d'une tâche de catégorisation de formes, et non de segmentation : les données annotées ont en effet montré, dans le chapitre précédent, que les RP sont bien souvent constituées d'un seul élément et que certaines formes y sont particulièrement récurrentes.

1.6.3. TRAITEMENT DES TÂCHES ENVISAGÉES

Plusieurs stratégies d'implémentation sont possibles pour réaliser ces deux tâches d'extraction et de catégorisation.

Le travail de catégorisation peut être envisagé de façon similaire à l'annotation manuelle, en concevant une démarche d'identification automatique des unités de rection ('UR') et les groupes intonatifs ('GI'), voire d'autres unités d'analyse pertinentes et opérationnelles pour un reconnaiseur automatique. Une première approche serait dès lors basée sur des règles et permettrait une classification indépendante, basée sur la position des éléments; les informations sont produites à partir d'un module distinct.

Toutefois, comme l'a montré le chapitre précédent, les phénomènes liés aux périphéries droites sont souvent interdépendants; la probabilité de faisceaux de traits rend donc possible une seconde approche des données, par inférences statistiques.

1.7. APPROCHE PAR RÈGLES : L'ANALYSE SYNTAXIQUE PAR DÉPENDANCES

L'analyse syntaxique par dépendances est une formalisation des relations entre lemmes d'une phrase qui prend la forme d'un graphe acyclique, c'est-à-dire d'une forme particulière d'arbre dont les nœuds sont les différents lemmes, et donc les arcs, étiquetés ou non, correspondent aux différents liens de dépendance.

La dépendance, dans une phrase, peut se définir comme suit (Candito, Crabbé, Denis, & Guérin, 2009):

la dépendance syntaxique représente le fait que la présence d'un mot ou unité est légitimée par un autre mot, son gouverneur. De plus il s'agit de dépendances "surfaciques", c'est-à-dire des relations entre formes fléchies, où toute forme fléchie, même sémantiquement vide, est représentée, et a un et un seul gouverneur, sauf pour la forme tête de la phrase (Candito, Crabbé, Denis, & Guérin, 2009, p. 3)

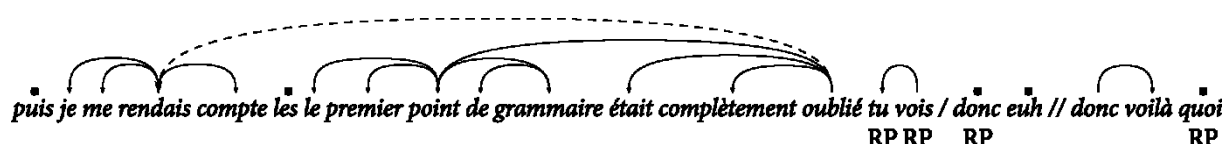
L'analyse en dépendances présente l'avantage de refléter directement la démarche d'identification manuelle; dans ce cadre, une approche sur base de règles est envisageable.

Ainsi, un point fort d'une analyse en dépendances est de repérer directement les unités de rection ; les éléments verbaux formeraient des têtes de dépendance, et les dépendants un domaine régi. Dans son principe, un parsing en dépendance reproduit avec fidélité les analyses microsyntactiques de l'approche pronominale.

1. APPLICATION DU FORMALISME À DES DONNÉES ORALES TRANSCRITES

La figure ci-dessous propose une annotation manuelle en dépendances – non-étiquetées – qui répondrait aux spécificités de l'analyse des transcriptions orales en français. Idéalement, un parseur doté d'une précision maximale serait en mesure de restituer cette structure à l'identique :

Figure x – Exemple d'annotation des dépendances pour des données orales transcrites



(blaAT11:34)

Si elle répond à certaines spécifications l'adaptant à l'oral conversationnel, l'analyse des dépendances est, même dans le cadre d'une annotation manuelle, un outil intéressant pour identifier les éléments recteurs et les limites de la rection. Dans la figure ci-dessus, la flèche en pointillés représente une dépendance distante ; ici, elle formalise ainsi un lien entre deux unités de rections, assimilable à une relation macrosyntaxique de type *préfixe + noyau*. L'heuristique d'identification des éléments recteurs, même pour des liens complexes entre UR, est relativement directe et reproduit celle de l'annotation manuelle : une tête de dépendance qui n'est pas dépendante d'une autre est notée comme un élément recteur ; s'il s'agit d'un verbe morphologiquement tensé et que celui-ci dépend d'un autre élément, on se trouve alors avec deux unités de rection dont l'une dépend de l'autre⁶⁵.

⁶⁵ Cf. section II.2.1.5 pour les liens de dépendance entre UR.

Une analyse en dépendances maximales réalisée sur des données orales devrait être en mesure de considérer certaines unités comme non-rattachées. De fait, plusieurs mots graphiques de l'exemple ci-dessus n'entreront pas dans une structure de dépendance : ici, le segment interrompu initié par '*les*', la pause pleine '*eah*' et les marqueurs '*quoi*' et '*donc*'. Le marqueur '*tu vois*' est également isolé s'il est lemmatisé comme une seule unité.

1.7.1. HEURISTIQUE D'IDENTIFICATION DES RP

Un segment en périphérie droite peut être déduit de cette structure de dépendances et correspondre à deux types d'informations récupérées par le parseur. Dans le premier cas de figure, une RP est identifiée à une structure de dépendance saturée qui répond à une série de critères : contenir un nombre réduits de dépendants, appartenir à un lexique de formes tel que celui des recteurs faibles ou des marqueurs composés. Dans le second cas de figure, une périphérie droite est identifiée à un élément indépendant ; certaines formes graphiques sont exclues ; un processus de décision doit être implémenté discriminer les périphéries gauches des périphéries droites, et exclure les formes potentiellement noyau, les signaux paraverbaux symbolisés par une graphie (rires '*haha*', exclamations '*pf*', etc.) et les pauses pleines ('*eah*').

Certaines différences d'approche demandent toutefois un traitement particulier. Les parseurs réalisant une analyse en dépendances sont conçus avec, en ligne de mire, l'objectif de décrire tous les éléments d'une « phrase » dans cette structure de dépendance ; la nature de la dépendance, ou l'éventualité que certains éléments soient tout simplement structurellement indépendants, n'est pas considérée. Le GARS, en comparaison, voit la rection comme un phénomène dont la portée est limitée ; au-delà de la rection, il existe certes des liens de dépendances, mais ceux-ci sont décrits dans le cadre de la macrosyntaxe.

1.7.2. ALGORITHMES DE PARSING

Un parseur en dépendances fonde son analyse sur des *mots* et des *phrases* ; il demande une étape de pré-traitement dans laquelle une lemmatisation et une découpe en phrases sont

effectuées⁶⁶ (Kübler, McDonald, & Nivre, 2009). Par « phrase » est ici considérée une séquence de mots w_i telle qu'une phrase S de longueur n contiendra la séquence $w_0 \succ w_1 \succ \dots \succ w_n$, en considérant que $\forall w_i, w_i \in V_S$. Un graphe de dépendance peut être non-labellisé ; l'objectif est alors de fournir à partir de S un ensemble de relations $(w_i, w_j) \in R$. Chaque relation de dépendance peut également être étiquetée, auquel cas l'information restituée par le parseur sera un ensemble de relations $(w_i, r, w_j) \in R$, dans lesquelles r appartient à un ensemble d'étiquettes $R = \{r_1, r_2, \dots, r_k\}$. Dans l'acception classique d'un graphe de dépendances étiquetées, deux mots w_i et w_j ne peuvent alors avoir qu'une et une seule relation r , sans quoi on se trouve dans des représentations « multi-stratales » (Kübler, McDonald, & Nivre, 2009, p. 12) représentant simultanément plusieurs analyses distinctes. Dans tous les cas, les données d'output auront la taille $V_S \times R \times V_S$.

Les analyses en structures de dépendances peuvent être effectuées à partir de plusieurs algorithmes : le *shift-reduce parsing*, comparable aux PCFG en analyse des constituants ; les algorithmes utilisant des graphes de transition ; les algorithmes de « parsing déterministe » tels que celui utilisé par le MaltParseur (Nivre, 2004).

Les PCFG (*probabilistic context-free grammars*) sont une première catégorie d'algorithmes pour l'analyse en dépendances. Elles sont la transposition directe des PCFG utilisés pour la syntaxe en constituants. Un analyseur grammatical de type PCFG, dans sa version la plus simple, travaille uniquement à partir de règles définies manuellement ou apprises sur un corpus pré-annoté. Par rapport aux CFG, les PCFG partent du principe qu'en langue naturelle, les structures ne connaissent pas des fréquences uniformes d'occurrence. Il s'agit donc de calculer le parsing le plus probable pour une phrase donnée, en évaluant les différentes alternatives. Concrètement, ce travail repose sur l'application *top-down* de règles de réécriture prédéfinies. Le parseur démarre ainsi avec une dépendance racine (par exemple, S ou C) et en recherche la réécriture la plus probable. À chaque règle est assignée une probabilité d'occurrence, de telle sorte que certaines dépendances entre mots seront plus

⁶⁶ Par « phrase » est ici considérée une séquence de mots w_i telle qu'une phrase S de longueur n contiendra la séquence $w_0 \succ w_1 \succ \dots \succ w_n$.

probables que d'autres. Un nœud de dépendance syntaxique identifié peut donc comporter plusieurs possibilités de dépendants, pondérés en fonction des données observées.

La première réécriture sera souvent $S \rightarrow V$ définissant une tête verbale mais des phrases à tête nominale, adverbiale, etc. sont également représentables : leurs probabilités d'apparition seront tout simplement moindres. L'interprétation d'un syntagme nominal se fera par exemple une alternative de type

$$R \ni \left\{ \begin{array}{l} N \xrightarrow{0.8} Det\ N \\ N \xrightarrow{0.2} Det\ N\ CP \end{array} \right.$$

dans laquelle un nom peut être réécrit tantôt par un lien de dépendance entre lui-même et un déterminant, tantôt par un double lien de dépendance entre lui-même, un déterminant et une phrase relative. Cette application des règles se poursuit jusqu'à n'obtenir plus que des nœuds terminaux. Un problème majeur des analyses en dépendances basées sur les PCFG est qu'une mauvaise décision prise par le parseur à un haut niveau de l'arbre peut se propager sur l'ensemble de la phrase. Si les PCFG sont, pour un parsing en dépendances non-étiquetées, relativement performants en termes de calcul⁶⁷, ils perdent ainsi beaucoup en précision ; pour cette raison, la plupart des parseurs PCFG sont non-déterministes : ils conservent lors de leur traitement plusieurs analyses en mémoire. Dans certains cas, ces analyses alternatives vont jusqu'à être produites en sortie.

Les deux autres algorithmes que je présente brièvement ici sont caractérisés par une séparation du modèle de parsing et des données linguistiques attribuées. Le premier algorithme est souvent qualifié de *graph-based* car il résume la tâche du parsing en dépendances à celle de représenter la phrase sous la forme d'un graphe dans lequel il s'agit d'estimer l'« arbre d'expansion maximal » (*maximum spanning tree*, MST) (voir p. ex. McDonald & Pereira, 2006). Ce MST applique la contrainte des arbres *projectifs* : si une

⁶⁷ Chaque mot w_i demandant V_S comparaisons itératives de mot à mot, chaque fois en vertu d'un ensemble de relations $R_i \subseteq R$, la complexité calculatoire des PCFG est d'ordre $V(V \times R \times V) \Rightarrow \theta(n^3)$.

dépendance existe entre i et j , alors aucun élément situé entre i et j ne peut être dépendant d'un élément situé au-delà de ces derniers. Pour le calculer, McDonald et Pereira propose de considérer chaque mot de la phrase comme un vertex – i.e. un nœud au sein d'un graphe –, dans un graphe orienté dont les arcs sont chacun dotés d'une probabilité a priori de dépendance entre mots.

Le dernier algorithme présenté ici est celui du parseur MALT. Le MALT appartient à la famille des algorithmes de *shift-reduce parsing*. Ceux-ci consistent à évaluer les dépendances à partir de règles selon une progression *bottom-up*, à l'opposé des PCFG ; le MALT s'en distingue par le fait qu'il réalise une analyse « pseudo-déterministe » et qu'il considère d'autres types de transitions (Nivre, 2008, p. 522). Le principe est le suivant : selon un fonctionnement, *bottom-up*, les dépendances proches sont produites en premier lieu, par évaluation de la relation entre un lemme w_i et des lemmes suivants $w_{j+1}, \dots, w_{j+n}$, itérativement. Chacun des lemmes, sa catégorie POS – qui doit avoir été préalablement produite avec la lemmatisation – et de possibles autres traits généralement binaires sont utilisés pour la comparaison. Suite à ce travail de comparaison, quatre transitions existent : (1) la transition (*'shift'*), qui remet la formulation d'une relation r à une passe ultérieure ; la décision d'une dépendance peut être alternativement (2) *gauche* ou (3) *droite* – seule la tête de dépendance est alors conservée pour la passe suivante (*'reduce'*) ; et enfin (4) ; la décision de faire de w_i une nouvelle tête si rien n'est trouvé. L'analyse est bouclée quand la dernière dépendance assignée est la plus distante, et joint la phrase en un seul arbre de dépendances. Le parseur MALT s'est avéré être une implémentation particulièrement performante⁶⁸ pour l'analyse en dépendances. En outre, même s'il est en général moins précis que d'autres algorithmes plus complexes tels que le MST, il se montre efficace pour les dépendances étiquetées (Candito, Nivre, Denis, & Anguiano, 2010).

⁶⁸ Nivre (2008) indique que la complexité temporelle du parseur MALT pour le traitement des dépendances de chaque lemme est de $O(n^2)$, ce qui amène à une complexité d'ordre $O(n^3)$ pour l'analyse d'une phrase complète.

L'analyse par dépendances, qui permet une extraction et une catégorisation fine des RP, présente toutefois des limites. Si l'analyse est effectuée dans le détail et contient un grand nombre d'informations utiles à l'analyste, elle demande un temps de traitement important. En outre, elle présente le même défaut que la plupart des approches d'identification par règles : des règles mal paramétrées peuvent mener à des interdépendances erronément calculées, ce qui produira des biais importants, d'autant plus susceptibles de se propager et de générer des données bruitées.

Les modélisations statistiques, en particulier sous forme de graphes (réseaux bayésiens, modèles de Markov, modèles d'Ising) sont une solution à ce type de problème : une distribution complexe est ramenée à un produit de plusieurs facteurs locaux, lesquels s'applique à des sous-ensembles de variables plus réduits (Sutton & McCallum, 2006). Une partie de l'information globale est perdue, mais l'information locale est traitée de façon plus adéquate pourvu que les dépendances conditionnelles observées soient pourvues d'une récurrence statistique suffisante. Le chapitre précédent a montré que de telles récurrences sont observables et significatives au niveau des transitions entre catégories d'éléments, ce qui conduit à postuler d'une pertinence de ces modèles pour catégoriser les segments RP.

1. PERTINENCE DES MODÈLES CRF

Un modèle, en particulier, est en mesure d'analyser les périphéries droites selon un modèle statistiquement adéquat par rapport au type de données : les champs conditionnels aléatoires (*'conditional random fields'*, CRF). Les CRF linéaires, en particulier, reposent sur des graphes de transitions acycliques ; ils sont adaptés pour le traitement des données linguistiques. En traitement automatique du langage, ils ont été appliqués à de nombreuses tâches : la reconnaissance d'entités nommées et de traits descriptifs relatifs à ces entités nommées, le *'shallow parsing'*, ou encore l'identification de rôles sémantiques dans les séquences de mots (Sutton & McCallum, 2006). Une recherche liminaire en la matière est celle de Sha & Pereira

(2003), dans laquelle ils appliquent les CRF à l'étiquetage des séquences de mots ; ce faisant, ils parviennent à une précision comparable, voire meilleure que les reconnaisseurs '*ad hoc*' travaillant à partir de règles écrites.

Pour le problème spécifique de l'identification-catégorisation des périphéries droites, les CRF possèdent deux propriétés intéressantes : d'une part, la capacité de calculer des indépendances de façon conditionnelle, et d'autre par le fait de permettre une factorisation de variables influant sur les décisions de transition (Sutton & McCallum, 2006).

La propriété d'indépendance conditionnelle des algorithmes CRF facilite l'élaboration de modèles spécialisés pour des phénomènes locaux⁶⁹. Les périphéries droites entrent dans cet ordre de phénomènes. De fait, elles représentent un problème de catégorisation de formes plus que de segmentation à proprement parler : si elles constituent des suites de formes, elles sont en grande majorité constituées d'une suite d'éléments ; un grand nombre d'entre elles sont des formes récurrentes ; leur structure produit un certain nombre de fréquences positionnelles.

De l'autre côté, la propriété de factorisation, pondérant les inférences du modèle, est une façon efficace de concevoir des algorithmes inférentiels. L'avantage des CRF est ainsi d'être adaptables à des environnements dans lesquels les traits à prendre en compte sont nombreux. Pour les périphéries droites, il sera ainsi possible d'introduire certaines règles basées sur le traitement des formes, par des recours à des lexiques de taille réduite : lexique de verbes

⁶⁹ La *reconnaissance d'entités nommées* est par exemple une tâche pour laquelle les CRF se sont montrés particulièrement efficaces (Sutton & McCallum, 2006, pp. 12-13). Dans la *reconnaissance d'entités nommées*, l'objectif est d'identifier les entités faisant partie d'une certaine classe : les noms propres. Ensuite, de leur assigner une sous-classes : noms de personnes, noms de lieux, noms d'organisations. Pour reconnaître les entités nommées, il est en pratique irréalisable de se baser sur un lexique de formes. De nombreuses EN sont en effet trop rares pour apparaître dans un corpus, même très important. En outre, les entités nommées sont souvent constituées de plusieurs mots graphiques, ce qui rend inefficace tout algorithme analysant en dur les formes individuelles (p. ex. les modèles de Markov naïfs).

cognitifs ou expressifs permettant d'établir les recteurs faibles, ou encore lexiques de formes fréquemment marqueurs.

1.8.1. LIMITES DES APPROCHES BAYÉSIENNES

Les approches bayésiennes représentent, dans leur version naïve, l'approche la plus simple d'un reconnaisseur à base algorithmique statistique. Le principe est le suivant : en prenant un ensemble de formes que l'on considère comme observées, représentées dans l'ensemble $X = \{x_1, x_2, \dots, x_k\}$, et un ensemble $Y = \{y_1, y_2, \dots, y_k\}$, de tags y que l'on souhaite prédire, on peut calculer, en les stockant dans une matrice V , chaque probabilité conjointe $p(x, y)$, *i.e.* la fréquence à laquelle on obtiendra chaque x pour chaque y ; cette modélisation entre dans le cadre des formalisations bayésiennes classiques ; on postule de l'indépendance de X et de Y telle que

$$p(X, Y) = p(X)p(Y)$$

Une difficulté évidente des approches bayésiennes est leur *complexité calculatoire* : pour un ensemble de X formes observées pour Y tags, X^Y possibilités doivent pouvoir être calculées. Notons que cette complexité peut être réduite à partir d'une application algorithmique de l'entropie maximale de Shannon.

Une implémentation bayésienne absolument naïve est en outre problématique du fait que la *confiance du modèle* est établie sur des catégories beaucoup trop fréquentes. En ce qui concerne les périphéries droites, nous avons ainsi constaté l'asymétrie de fréquence des catégories : les marqueurs, par exemple, sont largement majoritaires dans les données. Cet aspect posera des problèmes au niveau de l'apprentissage dans la mesure où l'identification sera sur-adaptée à la reconnaissance des marqueurs en initiale de RP et sous-estimera, par exemple, les périphéries droites commençant par un recteur faible '*vpar*'.

Ensuite, on postule de l'*indépendance* de chaque tag vis-à-vis d'un tag adjacent. Dans une représentation des données qui postulerait que chaque probabilité d'avoir y étant donné x est indépendante de ses éléments adjacents, on aurait une distribution de probabilités qui serait

tout simplement le produit des probabilités de chaque $p(y|X)$. Cette distribution correspond donc à une distribution de type

$$p(y|X) = \prod_k (x_k|y_k)$$

Ce postulat est non seulement une perte d'information compte tenu de la récurrence de certaines suites de tags, mais omet en outre la nature nécessairement séquentielle des chaînes textuelles par laquelle certaines formes x prendront un certain tag compte tenu de leurs prédécesseur x_{-1} ou successeur x_{+1} , et par laquelle certaines formes, inversement, marqueront des frontières plus ou moins importantes et seront alors indépendantes soit de x_{-1} , soit de x_{+1} . À ce sujet, l'annotation a montré que les éléments en périphérie droite peuvent localement montrer que des séquentialités – *i.e.* des dépendances de transition – sont observables. Les *marqueurs*, par exemple, ne sont jamais suivis par un *verbe défocalisé*, et très rarement par un *recteur faible*; au niveau des fonctions – dont beaucoup sont corrélées à certaines lexies –, il a également été montré que les éléments à fonctions *interpersonnelle* sont précédés par des éléments à fonction *dialogique* ou *subjective*. Une approche d'entropie maximale n'est pas en mesure de tenir compte de ces dépendances conditionnelles entre éléments.

Pour une tâche de reconnaissance, la production d'un modèle génératif, en mesure de prédire lui-même des suites de segments, n'est pas nécessaire. Le calcul de la probabilité conjointe $p(x, y)$ n'est pas l'objectif du modèle; il suffit de calculer directement la probabilité conditionnelle $p(x|y)$ pour assigner un tag y à partir d'une forme x donnée.

1.8.2. MODÈLE CRF : DÉFINITION ET PRINCIPE

Le problème principal d'un HMM est qu'on ne puisse pas recourir à d'autres traits de surface (majuscules, environnement syntagmatique, préfixes, suffixes, etc.). En comparaison, les CRF ont pour avantage d'être adaptables à des environnements dans lesquels les traits à prendre en compte sont nombreux. Ainsi, pour les périphéries droites, il faudra prendre en compte des lexiques de formes : lexique de verbes cognitifs ou expressifs permettant d'établir les recteurs

faibles ; lexiques de formes fréquemment marqueurs, prenant en compte des probabilités basées sur des fréquences.

Le problème de la reconnaissance des périphéries droites est celui de la prédiction d'un $Y = \{y_1, y_2, \dots, y_k\}$ de variables aléatoires à partir d'un vecteur $X = \{x_1, x_2, \dots, x_k\}$. Dans un modèle qui prend en compte les dépendances séquentielles entre les éléments, on ajoute à cette probabilité la probabilité de ce tag y_k d'être suivi, dans un modèle donné, par un élément y_{k+1} . Cette combinatoire est enregistrée dans une matrice V dont les dimensions sont égales au nombre de catégories $y \in V$ et $y_{+1} \in V$. C'est cette même matrice V qui sera toujours utilisée pour les transitions entre les éléments⁷⁰.

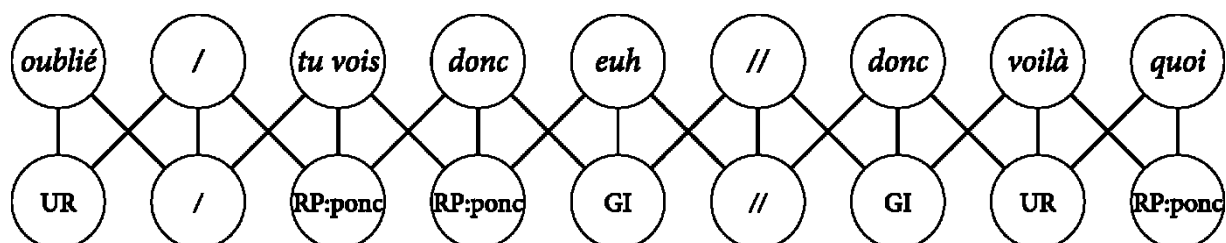
Dans un CRF linéaire, les calculs seront ainsi toujours effectués sur deux types de vecteurs : les vecteurs d'état et les vecteurs de transitions – ces derniers étant formellement comparables à des représentations du type des modèles de Markov cachés. Ces deux groupes de vecteurs contiennent des traits relatifs à un x donné à une position donnée, mais peuvent également dépendre de propriétés plus larges des éléments en entrée (Sha & Pereira, 2003, p. 2). À partir de ce calcul vectoriel, les tâches de classification peuvent dès lors se résumer à attribuer une seule variable de classe ; toutefois, l'intérêt principal des modélisations par CRF – et plus généralement des modélisation par graphes – est de permettre la modélisation de plusieurs variables interdépendantes.

1.8.3. APPLICATION AUX SEGMENTS RP

La reconnaissance des RP, à partir des CRF, peut être envisagée comme le calcul de l'assignation d'un tag '*RP*' / '*non-RP*' à l'ensemble des lemmes d'une transcription. Conjointement, les CRF peuvent également faire intervenir des calculs permettant d'induire les catégories formelles et fonctionnelles des RP, voire même catégoriser des éléments comme

⁷⁰ Le chapitre précédent, en présentant les fréquences de transitions entre éléments, a fourni un exemple de ce type de matrices (III.3.3 et III.3.4) ; ces observations n'avaient alors qu'une valeur exploratoire et se cantonnaient à l'observation des probabilités séquentielles de transition entre les catégories.

finaux ou initiaux d'UR ou de GI. Le graphe ci-après est un exemple d'analyse par CRF. Chaque input x (nœuds supérieurs) est susceptible d'être modifié par les catégories environnantes.



À partir des CRF, la reconnaissance de chaque élément de la RP peut être opérée moyennant un faible cout, tant en termes de performances que de précision du modèle. Un certain nombre de traits de la RP peuvent contribuer aux inférences du modèle statistique : la taille des mots, la position dans le tour de parole, la fréquence de certains inputs, etc.

Un aspect intéressant des CRF est qu'ils sont en mesure d'effectuer plusieurs tâches de traitement simultanément. Ainsi, Sutton et al. (2007, cités par Sutton & McCallum, 2006) ont implémenté un CRF à deux niveaux dynamiques effectuant simultanément une tâche de *chunking* et une tâche d'identification des NP et une tâche de POS-tagging. Les CRF rendent également possible l'assignation de catégories multiples aux mêmes éléments, une propriété utile en particulier pour effectuer une annotation fonctionnelle des RP : de fait, celle-ci doit idéalement pouvoir rendre compte de nombreux cas de polyfonctionnalité.

CONCLUSION

À travers ce mémoire auront pu être proposées deux définitions des périphéries droites, l'une théorique et l'autre opératoire, ainsi que deux méthodes d'identification de celles-ci, l'une manuelle et l'autre automatisée.

Dans la première partie, les périphéries droites sont définies à partir de différentes théories grammaticales portant sur la segmentation du français en usage. Après avoir posé la question de la notion de « phrase » qui questionne le lien entre syntaxe et discours, ce chapitre parcourt plusieurs théories grammaticales qui en proposent une description alternative. Dans ce cadre se distinguent des approches « orientées syntaxe » et des approches « orientées discours ». Ces dernières, dans l'objectif d'une automatisation de l'identification, sont écartées au profit des approches "orientées syntaxe". L'Approche pronominale du groupe aixois de recherche en syntaxe ('GARS'), en particulier, est retenue comme socle définitionnel pour les périphéries droites ; le modèle de la Grammaire de la période est également intégré pour certains aspects.

La périphérie droite est ainsi définie, à la suite du premier chapitre, comme une entité en essence macrosyntaxique, correspondant aux postfixes. Intonativement, cette forme entre en correspondance avec l'intonation d'« appendice » qui se caractérise par une réalisation mélodique en plage basse, non modulée - bien que quelques exceptions sont possibles. La macrosyntaxe, toutefois, n'est pas observable directement ; c'est pourquoi la définition retenue pour les périphéries droites prend en compte des faits de surfaces corrélés aux éléments postfixes. Quatre groupes de périphérie droites sont ainsi identifiables par leur forme : les associés à la rection ; les marqueurs discursifs et pragmatiques ; les recteurs faibles ; et enfin les "verbes défocalisés". Ces quatre catégories ne sont pas seulement attestées par l'*Approche pronominale* : elles ont été observées à de nombreuses reprises dans la littérature sur les périphéries.

Dans la deuxième partie a été développée une méthode d'annotation manuelle des périphéries droites. Les corpus considérés pour la tâche d'annotation manuelle proviennent de

conversations informelles issues de la banque de données Valibel. Les normes de transcription de ces corpus conditionnent l'annotation des éléments périphériques : les conventions graphiques orientent les observations vers les faits morphosyntaxiques et produisent notamment un silence sur certaines différences de réalisations morphologiques. Les notations de la prosodie, quant à elles, sont réduites, tandis que les notations des tours de parole et la mise en scène de l'interaction sont insuffisamment normalisés pour rendre compte de certaines informations contextuelles, de nature pragmatique ou interactive, liées à l'occurrence des périphéries.

Dans la deuxième partie de ce deuxième chapitre, j'ai proposé une méthode opératoire d'identification des périphéries droites qui s'abstrait des données d'intonation. L'identification des périphéries droites suppose d'établir, en amont, les limites de deux unités : les *unités de rection* et les *groupes intonatifs* ; un segment périphérique droit (RP) est considéré comme l'ensemble des éléments à droite d'une *unité de rection* appartenant au même *groupe intonatif*. L'*unité de rection*, tout d'abord, opérationnalise les regroupements de rection théorisés par l'approche pronominale et se veut plus inclusive, notamment par l'intégration de « recteurs averbaux » et de « formes nucléaires » ; elle est identifiable, manuellement, à partir de deux tests proposés par le GARS : la substitution pronominale et l'extraction en clivée. Le *groupe intonatif*, de son côté, se fonde sur des indices textuels et syntaxiques pour établir des liens de dépendance macrosyntaxiques, avec une attention particulière portée sur les éléments postfixes ; pour être opérationnalisé, celui-ci demande d'objectiver les impressions de l'annotateur concernant l'annotation d'une « structure prosodique par défaut », et de catégoriser les liens de dépendance au-delà du domaine de la rection pour capter un ensemble de dépendants incluant les éléments périphériques, et uniquement ceux-ci.

À la suite de ces dispositions relatives à l'identification des segments, l'annotation des types et fonctions des éléments périphériques proposée pour ce mémoire est également décrite dans ses aspects méthodologiques. Plusieurs catégories formalisables par la microsyntaxe sont opérationnalisées : marqueurs discursifs et pragmatiques ; associés lexicaux ou pronominaux ; recteurs faibles ; verbes « défocalisés ». La position des éléments dans le tour de parole est

également annotée : fins de tours, fins d'énoncés et positions liées aux afterthoughts. Enfin, une annotation fonctionnelle des éléments est proposée, qui annote d'une part les fonctions textuelles jouées par les éléments périphériques droits - cohérence discursive, *reference tracking*, structurant argumentatif - ainsi que leurs fonctions illocutoires – de type subjectif ou intersubjectif.

Le troisième chapitre de ce mémoire explore les données collectées lors de l'annotation. Y sont décrites dans un premier temps les données relatives à l'étape d'identification des segments, à partir des unités de rection et des groupes intonatifs. Il y est constaté une grande disparité du phénomène des périphéries droites : celles-ci varient grandement en fréquence d'une transcription à l'autre. Plusieurs biais importants de l'identification sont épinglés.

Les positions occupées par les périphéries dans le tour de parole sont également investiguées. Les données indiquent que les périphéries droites ne seront ni plus longues, ni plus courtes en fonction de leur position dans le tour de parole. Plusieurs données indiquent un comportement interactionnel des périphéries droites. En matière de microsyntaxe, une présence significative des associés est constatée en fin de tour de parole, ce qui laisse présager d'un rôle interactionnel attribué aux éléments disloqués dans l'interaction ; les éléments périphériques portant une fonction interpersonnelle sont quant à eux caractéristiques des périphéries droites en finale de tour.

En ce qui concerne la structure interne des périphéries, l'analyse dégage plusieurs données cruciales. Les périphéries droites composées d'un seul élément monosyllabique sont de loin les plus fréquentes ; en termes de métrique, la périphérie droite serait un environnement propice aux phénomènes de réduction syllabique. Certaines formes tiennent le haut du pavé en termes d'occurrences : '*quoi*', '*hein*' ou encore '*tu vois*' forment ainsi, de loin, les périphérie droites les plus typiques. Les marqueurs affichent certaines préférences de positions, certaines cooccurrences, certains agencements de catégories ; mais d'autres témoignent également d'articulations radicalement paratactiques.

L'analyse des catégories étiquetées de l'annotation montre également que les catégories formelles de périphéries droites ne sont pas uniformément distribuées à travers les fonctions discursives. Une observation générale est que plus une fonction est fréquente, et plus forte est la présence des marqueurs. Les données montrent, au niveau des séquences observées, que certaines séquences de catégories sont plus fréquentes que d'autres et que la périphérie droite, en plus d'être un agencement relativement ordonné de formes, est également ordonnée, *parfois* dans ses catégories microsyntaxiques et *assez fréquemment* au niveau des catégories fonctionnelles d'éléments. La structure fonctionnelle des périphéries droites montre enfin des régularités statistiques qui témoignent d'une forte structure sous-jacente, spécialement au niveau des fonctions intersubjectives.

Dans la dernière partie de ce mémoire, plus brève, ont été proposées plusieurs perspectives d'implémentation pour la détection automatique des périphéries droites. Le problème en termes de traitement, qui est décrit comme une tâche d'extraction et de catégorisation, peut se résoudre à partir de l'implémentation de règles écrites ou bien à partir de modèles statistiques. Il s'agit d'une tâche de catégorisation de formes, et non de segmentation : les données annotées ont en effet montré, dans le chapitre précédent, que les RP sont bien souvent constituées d'un seul élément et que certaines formes y sont particulièrement récurrentes.

Dans son principe, un parsing en dépendances reproduit avec fidélité les analyses microsyntaxiques de l'approche pronominale et apparaît donc comme la démarche la plus adéquate d'implémentation sur base de règles. Pour le problème spécifique de l'identification-catégorisation des périphéries droites, les CRF possèdent deux autres propriétés intéressantes qui rendent un modèle statistique pertinent : d'une part, la capacité de calculer des indépendances de façon conditionnelle, et d'autre par le fait de permettre une factorisation de variables influant sur les décisions de transition.

Un intérêt d'une recherche ultérieure serait d'effectuer le même travail d'annotation en tenant compte de l'intonation celle-ci peut avoir une valeur ajoutée, voire changer certaines conclusions auxquelles ce mémoire est arrivé. Les fonctions d'« arrière-plan » et les

constructions défocalisées, deux phénomènes intrinsèquement liés, n'ont pu qu'être reconstitués au travers de ce mémoire, dans les limites des indices qu'ils laissaient dans une transcription orthographique. L'utilisation de l'intonation n'a pas seulement une pertinence exploratoire : au niveau de l'implémentation, un modèle statistique entraîné sur une proportion suffisante de données pourrait être en mesure de prédire certaines récurrences des structures intonatives à partir de la transcription seule.

Ce mémoire a été l'occasion de proposer une heuristique *sémasiologique* pour les segments en périphérie : la périphérie droite devient ainsi un paradigme observable au sein duquel de nouvelles formes peuvent être découvertes. Au-delà de ce que ce mémoire a pu montrer, une implémentation de l'approche proposée permettra de mieux comprendre les rôles fonctionnels de la périphérie et, par elle, les modalités selon lesquelles la temporalité structure la langue et les interactions.