Annotation et analyse syntaxique de corpus hétérogènes : le cas du français médiéval

Mathilde Regnault

Thèse de doctorat
de
l'Université Sorbonne Nouvelle - Paris 3

Lattice, Inria

École doctorale N°622, Langage et Langues
Spécialité du doctorat : Sciences du Langage

Annotation et analyse syntaxiques de corpus hétérogènes : le cas du français médiéval

---

présentée et soutenue publiquement par

Mathilde Regnault

le 16 juin 2022

sous la direction de Sophie Prévost, d'Isabelle Tellier† et d'Eric Villemonte de la Clergerie

Composition du jury
Sylvain Kahane, Professeur (Université Paris Nanterre), rapporteur,
Laura Kallmeyer, Professeur (Heinrich Heine Universität Düsseldorf), rapportrice,
Sophie Prévost, Directrice de recherche CNRS (Lattice), directrice de thèse,
Eric Villemonte de la Clergerie, Chargé de recherche (Inria), co-directeur de thèse,
Béatrice Daille, Professeur (Université de Nantes), examinatrice,
Annie Forêt, Maître de conférence HDR (Université de Rennes 1), examinatrice,
Achim Stein, Professeur (Universität Stuttgart), examinateur.

Annotation et analyses syntaxiques de corpus hétérogènes : le cas du français médiéval

Le français médiéval couvre les états de langue d'ancien français (9e-13e s.) et de moyen français (14e-15e s.). Nous disposons de données annotées pour ces états de langue, dont un corpus arboré d'ancien français (Stein et Prévost 2013). Il est cependant difficile d'obtenir plus de données annotées syntaxiquement, parce que les spécialistes sont peu nombreux et parce qu'il n'existe pas encore d'outil dédié pour l'ensemble de la période. Développer ce genre d'outil permet d'obtenir des annotations plus facilement et d'en contrôler la qualité. Cependant, ce n'est pas une tâche simple parce que les différents états de langue sont soumis à la variation, due à plusieurs facteurs, notamment l'absence de norme graphique, la variation dialectale, la souplesse de l'ordre des mots, l'évolution de la morphologie et de la syntaxe (sur sept siècles), qui fait passer le français d'une langue SOV à une langue SVO. La nature des écrits se diversifie aussi à mesure que la littérature évolue et que le latin est délaissé au bénéfice du français comme langue administrative et juridique. Les données à analyser sont donc hétérogènes, ce qui rend difficile le traitement automatique.

Pour obtenir un parseur du français médiéval, nous proposons d'adapter la métagrammaire du français contemporain FRMG (Villemonte de la Clergerie 2005). Bien que les différents états de langue présentent des différences manifestes, les points communs sont suffisants pour rendre possible la modification d'un système existant pour obtenir un outil dédié. Les changements concernent essentiellement l'ordre des mots (constituants majeurs, modifieurs du nom, position des pronoms conjoints). Pour utiliser cet outil sur corpus, il est nécessaire d'enrichir le lexique d'ancien français (Sagot 2019), d'une part pour obtenir une couverture lexicale satisfaisante sur les textes, et, d'autre part, pour y intégrer des informations syntaxiques et sémantiques nécessaires à l'analyse syntaxique.

Mots clés : annotation syntaxique, parsing, grammaire d'arbres adjoints, métagrammaire, français médiéval, ancien français, corpus hétérogène

Syntactic Analysis and Parsing of Heterogeneous Corpora : The Case of Medieval French

Medieval French is an umbrella term for Old French (9th-13h c.) and Middle French (14th-15th c.). We have annotated data for these stages, including a dependency treebank of Old French (Stein et Prévost 2013). However, obtaining more treebanks is difficult, because there are few experts of Medieval French and we do not have yet a dedicated parser for the whole period of Medieval French. A dedicated tool would make it easier to annotate new corpora and it would enable to control the quality of the annotation. Nevertheless, it is not a trivial task, because the states of language are subjected to variation. It comes from several sources, including the absence of standard spelling, dialects, flexible word order, evolution of morphology and syntax over seven centuries, with seminal phenomena like the transition from a SOV language to a SVO language. Text genres do also evolve as the number of literature writings rises and Latin is replaced by French for official texts such as treaties, contracts and chronicles. The data available for Medieval French are therefore heterogeneous, which makes it difficult to annotate them automatically.

We chose to adapt the *French Metagrammar* (FRMG, Villemonte de la Clergerie (2005)) in order to develop a parser for Medieval French. Even if the differences between Medieval French and Contemporary French are striking, there are enough similarities to obtain a satisfactory parser. The main changes ensure the word order is properly analysed (ex. major constituents, noun modifiers, position of clitics). In order to annotate a new corpus, adapting the lexicon OFrLex (Sagot 2019) is mandatory : new entries as well as new syntactic and semantic information were added.

Keywords : syntactic annotation, parsing, Tree-Adjoining Grammar, metagrammar, Medieval French, Old French, heterogeneous corpus

Remerciements

Mes tout premiers remerciements vont à mes directeurs de thèse, Sophie Prévost, Isabelle Tellier et Eric de la Clergerie. De la préparation du sujet, intégré au projet ANR Profiterole, jusqu'à la soutenance, en passant par de nombreuses étapes et quelques perturbations, ils m'ont offert le soutien dont j'avais besoin. Merci pour votre gentillesse et pour tout le temps que vous m'avez consacré, il m'a été très précieux. Je remercie infiniment Sophie Prévost pour tout ce qu'elle m'a appris, sur le français médiéval en particulier, et aussi plus globalement sur la méthodologie de la recherche en linguistique. Je remercie aussi Eric de la Clergerie d'avoir accepté de prendre plus de place dans mon encadrement, et de m'avoir guidée dans l'apprentissage des métagrammaires de type *FRMG*. Mille mercis pour les réponses à mes mille questions et toutes les discussions sur les méthodes symboliques. Enfin, je souhaite rendre hommage à Isabelle Tellier, dont je garderai le souvenir d'une pédagogue et d'une chercheuse hors pair.

Je remercie les membres du jury d'avoir accepté d'évaluer ce travail. Leurs questions, leurs remarques et leurs conseils m'ont permis d'améliorer des éléments importants de ce manuscrit et d'envisager autrement les prochains travaux.

Les comités de suivi de thèse ont été des moments salutaires pour ce projet. Je remercie les membres de ce comité de m'avoir aidée au fil des années, notamment dans l'orientation de mon travail.

Je remercie les membres du Lattice d'avoir fait de ce laboratoire un lieu où j'avais hâte de me rendre tous les matins, pour son ambiance de travail stimulante et chaleureuse. Merci aussi aux résidents de Thésardland^{TM} pour la camaraderie, le soutien sans faille et les éclats de rire. Je remercie également Frédérique de m'avoir accueillie dans son bureau.

Je remercie l'équipe ALMAnaCH de m'avoir accueillie toutes les semaines dans ses locaux. Merci pour les moments d'échanges et pour tout ce que vous m'avez appris.

Ich möchte mich jetzt bei den Mitgliedern vom ILR bedanken. Ihr habt mich von Anfang an herzlich willkommen geheißen und jeder hat mir Zeit gegeben, um mir meinen Amtsantritt und meinen Einzug in Stuttgart zu erleichtern. Vielen Dank an Achim für die Zeit, die du mir gegeben hast, um meine Dissertation zu beenden. Vielen Dank an Tom für die Diskussionen über die Kurse und wie man die letzten Monate der Promotion überlebt. Die Diskussionen, die wir im Forschungskolloquium oder während den Pausen gehabt haben, haben mir sehr geholfen. Herzlichen Dank !

Cette thèse a vu le jour grâce au projet ANR Profiterole. J'en remercie les membres pour leur disponibilité, leurs conseils et leurs enseignements.

Je remercie toutes les personnes qui ont rejoint le groupe de travail Déméter et avec qui j'ai pu échanger au sujet des grammaires formelles. Je remercie aussi l'équipe *Computer Science* de l'université Heinrich Heine de Düsseldorf pour leur invitation et les discussions sur les métagrammaires.

Je remercie les membres du club des CINQ, qui sont au nombre de six : Marie-Amélie, Marine, Auphélie, Loïc, Tian et Yoann. Selon l'expression consacrée, vous êtes le houmous du houmous, et je vous dois beaucoup de la joie que j'avais à faire cette thèse. Merci également à Clémentine, Alix, Tanti, Murielle,

Marie et à toutes les personnes que j'ai rencontrées au fil des années et que j'ai eu plaisir à côtoyer. Enfin, merci à Pedro de m'avoir enseigné l'Art du café, mais surtout pour son soutien et pour sa présence.

Je remercie ma grande famille pour tout son soutien à travers les années, et tout particulièrement mes grands-parents, mes parents, Paul et Anne. Merci infiniment.

Acronymes

AF	ancien français (9e–13e s.)
AND	Anglo-Norman Dictionary
BFM	Base de français médiéval
FC	français contemporain
FMed	français médiéval (9e–15e s.)
HTR	Handwritten Text Recognition (reconnaissance de l'écriture manuscrite)
MCVF	Modéliser le changement : les voies du français
MF	moyen français (14e–15e s.)
NCA	Nouveau Corpus d'Amsterdam
OCR	Optical Character Recognition (reconnaissance optique de caractères imprimés)
PROFITEROLE	PRocessing Old French Instrumented TExts for the Representation Of Language Evolution
PROIEL	Pragmatic Resources in Old Indo-European Languages
SRCMF	Syntactic Reference Corpus of Medieval French
SRCMF-UD	SRCMF au format Universal Dependencies
TAF	très ancien français (9e–11e s.)
UD	Universal Dependencies

Table des matières

Introduction

Nous disposons actuellement de peu de données annotées syntaxiquement pour le français médiéval, composé de l'ancien français (9e-13e siècles) et du moyen français (14e-15e siècles). Pour la syntaxe en dépendances, on ne compte que le corpus arboré *SRCMF* (Stein et Prévost 2013), une ressource de 250 000 mots qui couvre l'ancien français. Le but du projet PROFITEROLE (ANR-16-CE38-0010, 2017–2022, dirigé par Sophie Prévost), dans le cadre duquel cette thèse est financée, est de fournir d'une part une extension du *SRCMF* pour en faire une ressource de français médiéval d'un million de mots, et d'autre part de développer des outils de traitement pour ces états de langue. Les différents travaux de cette thèse sont menés en vue du développement d'outils, en particulier pour l'annotation syntaxique.

Le traitement de ressources de langue ancienne constitue un défi pour les méthodes d'annotation automatique, qui reposent sur l'hypothèse que des phénomènes réguliers peuvent être déduits des données. Or, ces données sont souvent très hétérogènes. Le français médiéval est ainsi fortement soumis à la variation. De l'émergence du français (9e s.) au développement des divers types d'écrits (littéraires, juridiques, didactiques...), la langue a considérablement évolué d'un point de vue phonétique, morphologique et syntaxique. De plus, ce qu'on appelle le "français" est en fait le regroupement des langues d'Oïl, des parlers régionaux qui diffèrent notamment par leurs systèmes phonologiques (et donc morphologiques). Cela entraîne des variations importantes entre les variétés dialectales et, du fait que la graphie n'est pas encore strictement fixée, au sein même des textes. La syntaxe est, elle aussi, soumise à la variation, en diachronie, mais aussi en synchronie. Nous présentons les particularités de notre objet d'études et celles des langues anciennes en général dans le chapitre 1.

Pour traiter ce type de données hétérogènes, plusieurs stratégies semblent efficaces. Certaines concernent globalement les problématiques d'adaptation au domaine, et d'autres concernent plus particulièrement les langues anciennes, comme la lemmatisation. Nous en faisons un tour d'horizon en nous concentrant sur le français médiéval (cf. chapitre 2). Cette partie se clôt sur des expériences préliminaires d'annotation morpho-syntaxique, dirigées par Isabelle Tellier.

Le projet principal de cette thèse est l'adaptation d'un parseur symbolique du français contemporain au français médiéval, en nous appuyant sur les connaissances linguistiques rassemblées pour contrôler l'annotation. Le parseur choisi est *FRMG*, basé sur la métagrammaire du même nom (Villemonte de la Clergerie 2005), que nous présentons dans le chapitre 3 avec les formalismes qui la sous-tendent. Les modifications apportées à cette ressource sont décrites dans le chapitre 4. Ce travail s'accompagne de l'adaptation du lexique *OFrLex* (Sagot 2019) et du segmenteur *SxPipe* (Sagot et Boullier 2008) pour les besoins de l'analyse syntaxique (cf. chapitre 5). Les résultats de ce travail sont présentés et comparés à ceux d'autres parseurs dans le chapitre 6, et nous proposons quelques réflexions sur notre démarche dans le chapitre 7.

A travers ce travail sur le français médiéval, nous souhaitons explorer des méthodes de traitement pour les données hétérogènes. Le corpus *Profiterole* regroupe en effet plusieurs sources d'hétérogénéité : le français se constitue peu à peu comme langue littéraire, entrainant une évolution dans le type d'écrits

et les sujets abordés, avec une croissance du vocabulaire. Le traitement de notre corpus rejoint donc les problématiques d'adaptation au domaine. A cela s'ajoutent l'évolution syntaxique de la langue et les variations d'ordre des mots et de graphie, sur un volume de données limité, faisant de ces états de langue un terrain dense et intéressant pour étudier le traitement de données hétérogènes.

1 Le français médiéval : particularités, corpus et environnement linguistique

Nous disposons actuellement de peu de données annotées pour les langues anciennes. L'édition numérique de textes anciens est elle-même difficile, car il faut rassembler l'ensemble des manuscrits d'un texte, le cas échéant, et faire des choix éditoriaux entre les variantes (Parussa 2010). L'annotation de ces textes, et a fortiori l'annotation syntaxique, est très coûteuse et encore peu développée en comparaison de ce qui existe pour les états de langue modernes. Cela s'explique, entre autres, par le nombre limité d'experts, quelle que soit la langue considérée, et l'absence de locuteurs natifs.

Cependant, l'étude de tels textes est nécessaire à différents égards. Ce sont avant tout des ressources littéraires, historiques, juridiques ou didactiques. Les travaux en humanités numériques ont permis de développer des méthodologies pour les exploiter sous forme de vastes corpus. Ces textes constituent aussi un objet linguistique en eux-mêmes, et servent de support à l'étude de l'évolution de la langue, puisque c'est à travers eux seuls que nous avons accès aux états de langue anciens. Du point de vue de l'analyse syntaxique automatique (ou *parsing*), ils représentent un objet intéressant, car il est difficile de constituer et d'évaluer des modèles pour des états de langue peu dotés, ou dont les données disponibles sont hétérogènes.

Le travail présenté dans cette thèse porte sur le français médiéval, qui est une succession d'états de langue du 9e au 15e siècle soumis à une forte variation. Le français médiéval (désormais FMed) est considérablement différent du français contemporain (désormais FC), ce qui rend sa compréhension difficile pour tout locuteur actuel qui ne l'a pas appris (Prévost 2005). Ce chapitre en présente les particularités les plus remarquables. Dans un premier temps, nous proposons une description générale du FMed. Puis nous présentons les divers corpus numérisés disponibles. Enfin, nous donnons un aperçu des ressources annotées d'autres langues anciennes, qu'elles soient disponibles ou en cours de construction.

1.1 Etendue chronologique, géographique et littéraire du français médiéval

1.1.1 Naissance et constitution du français médiéval

Le FMed s'étend sur une période de sept siècles, du 9e au 15e siècle. Cette période ne correspond pas au Moyen Age historique. Celui-ci commence à la chute de l'Empire romain d'Occident en 476, alors que les premiers textes considérés comme du français apparaissent bien plus tard, au 9e siècle. La fin de cette période historique est habituellement définie par la prise de Constantinople par les Turcs, en 1453, mais le FMed dépasse ce cadre. Nous traitons en effet les écrits jusqu'à la fin du 15e siècle, qui est la frontière reconnue, même si certains la situent à la moitié du 16e siècle. En comparaison, le FC ne compte qu'un peu plus de cent ans d'existence, car on détermine traditionnellement son émergence à la fin du 19e siècle.

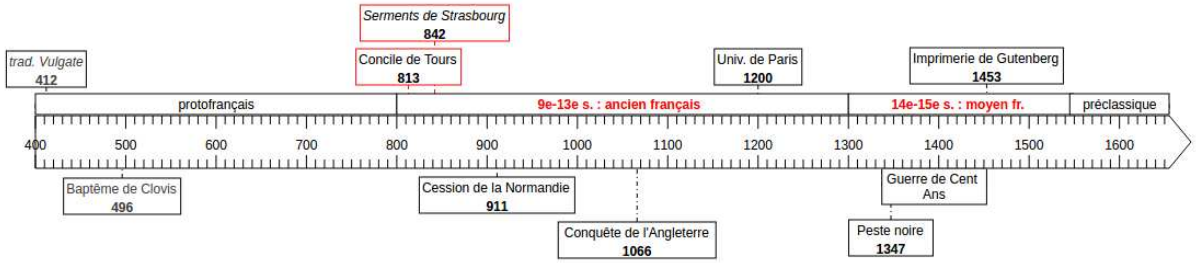


FIGURE 1.1 – La période du français médiéval
En haut : événements linguistiques
En bas : événements historiques

Le français est issu du latin tardif, dont nous avons peu de traces, et qui résulte lui-même de l'évolution du latin classique. Ce latin "tardif" est influencé par les langues d'origine des territoires (les substrats), ce qui explique sa diversification en plusieurs langues romanes. Les invasions germaniques au 5e siècle créent un clivage entre le nord et le sud, et on distingue alors langues d'Oïl (au nord), influencées par un superstrat germanique, et langues d'Oc (au sud). Ces noms viennent de l'ouvrage *De vulgari eloquentia* de Dante Alighieri, qui est le premier à avoir trouvé une origine commune à l'italien, au français et à l'occitan. Il désigne ces langues par leur mot "oui" (respectivement, "si", "oïl" et"oc"). Ce sont les langues d'Oïl qui constituent les premiers états du français.

La première trace de l'existence de cette langue est le témoignage du *Concile de Tours* en 813, où les évêques recommandent aux ecclésiastiques de prononcer leurs sermons en "linguam romanam rusticam" (c'est-à-dire en langue vulgaire), car le latin "classique" n'est plus compris par les populations. Le premier texte considéré comme contenant du "français" est l'accord signé en 842 entre Charles le Chauve et Louis le Germanique à la mort de Charlemagne, *Les Serments de Strasbourg*, consigné par Nithard.

Le FMed regroupe les états de langue depuis ces premiers témoignages jusqu'à la fin du 15e siècle, soit avant l'instauration des premières normes langagières strictes, au 16e siècle. Traditionnellement, on partage la période du FMed en ancien français (désormais AF), du 9e au 13e siècle, et en moyen français (MF), du 14e au 15e siècle (MARCHELLO-NIZIA 1999) ou jusqu'au milieu du 16e siècle, avant la période de français pré-classique. Dans cette thèse, nous considérons la fin du 15e siècle comme le terme final de la période étudiée.

1.1.2 Territoires

Les frontières du Royaume de France évoluent considérablement au fil des siècles. L'organisation du pouvoir est féodale, ce qui confère aux seigneurs une grande autonomie. Les alliances successives formées avec les seigneurs créent une certaine unité, mais le pouvoir n'est pas assez centralisé pour empêcher le morcellement dialectal sur l'ensemble du territoire.

Parmi les langues d'Oïl, on compte "le picard, le wallon, le normand, le champenois, le lorrain, le bourguignon, le gallo", mais aussi le poitevin et le saintongeais, qui les rejoignent plus tardivement (SIOUFFI 2020a). Le normand est un parler influencé par les Vikings, auxquels Charles le Simple cède la Normandie en 911. Ce dialecte se diffuse en Angleterre après la bataille de Hastings en 1066 et évolue vers ce qu'on appelle l'anglo-normand.

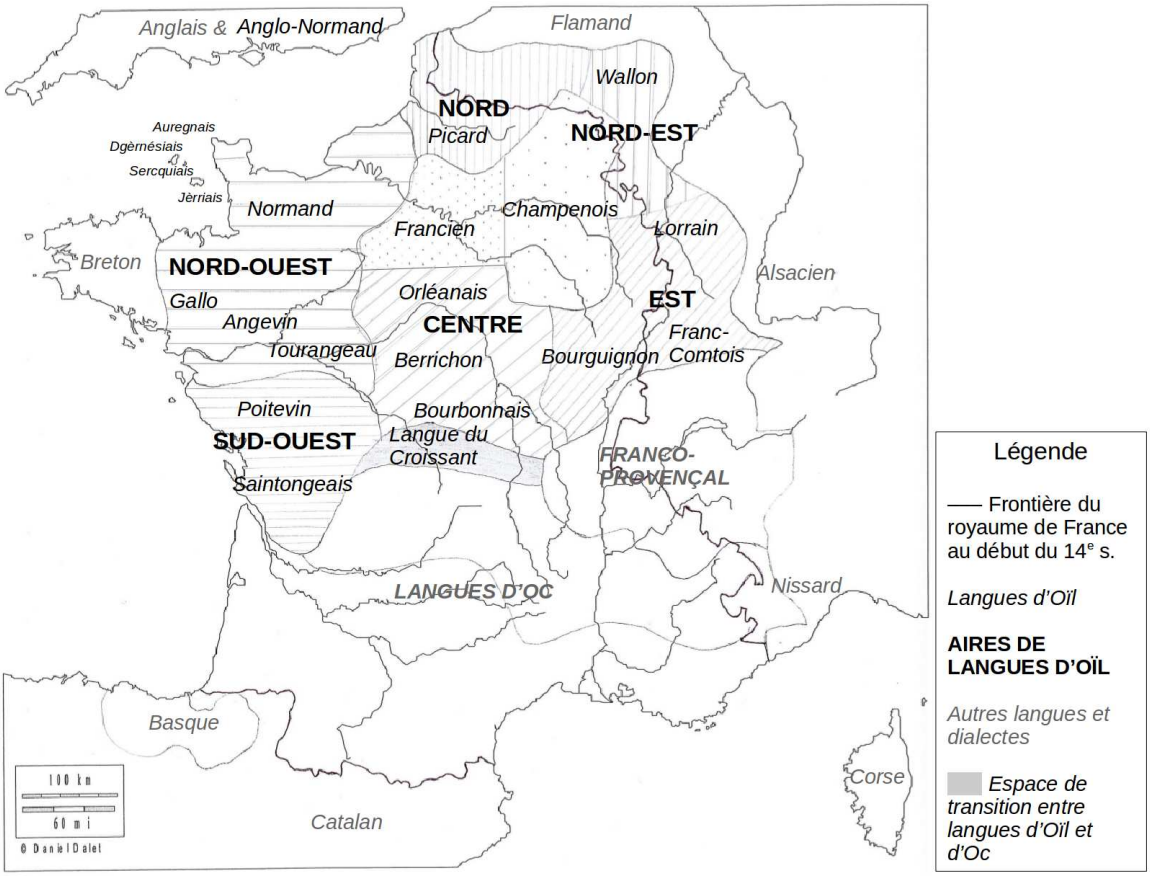


FIGURE 1.2 – Langues et aires dialectales au 14e siècle (BURIDANT 2000 ; DESTEMBERG 2017)

1.1.3 Production littéraire

La seule source de description du FMed dont nous disposons sont les manuscrits conservés de cette époque (PRÉVOST (2011), p. 53). Nous avons peu de données pour les états les plus anciens, car la production était plus faible et peu d'écrits sont parvenus jusqu'à nous. Nous disposons en revanche de plus de données à partir du 12e siècle. L'édition et la numérisation de ces textes étaient auparavant faites manuellement, mais les technologies de reconnaissance automatique de l'écriture manuscrite (HTR) accélèrent désormais ce processus. L'apport de nouveaux textes permet d'affiner les connaissances sur la langue médiévale.

A l'exception des *Serments de Strasbourg*, les premiers textes d'AF portent sur des thèmes religieux, comme *La Séquence de Sainte Eulalie* (881). Peu de textes de très ancien français (9e-11e siècle, désormais TAF) nous sont parvenus. À partir du 12e siècle, les genres des textes se diversifient. Les chansons de gestes, de tradition orale, sont consignées à l'écrit. Ce sont les premiers textes de littérature. Ils sont suivis par les romans, le théâtre et la poésie. Les écrits de didactique et d'histoire se développent surtout à partir du 13e siècle, ils étaient jusqu'alors en latin. Les premiers textes juridiques en langue vernaculaire dont nous disposons sont des chartes qui datent du 13e siècle. Les textes scientifiques, que nous classons dans le domaine didactique, n'apparaissent qu'au 14e siècle. Les premiers textes sont essentiellement en vers, mais la prose se développe fortement à partir du 13e siècle.

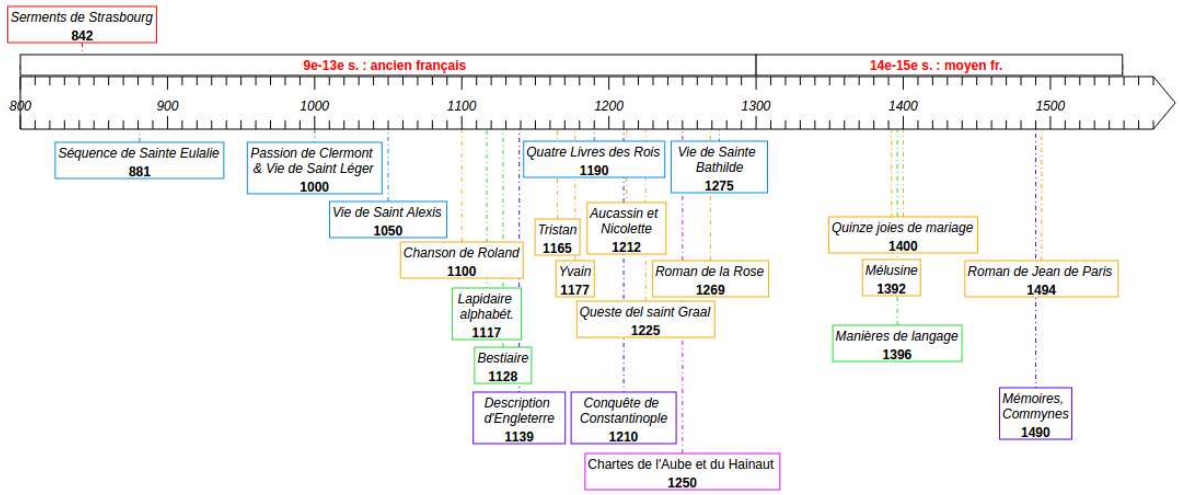


FIGURE 1.3 – Oeuvres majeures
Légende
Bleu : écrits religieux
Jaune : littérature
Vert : écrits didactiques (technique, sciences, langues...)
Violet : écrits historiques
Rose : écrits juridiques

Dans le cadre d'expériences de *parsing* menées avec Loïc Grobol, Pedro Ortiz Suarez et Benoit Crabbé, nous avons rassemblé toutes les données de FMed disponibles pour la recherche. Cela permet de donner une représentation de la quantité de données numérisées et disponibles pour chaque domaine.

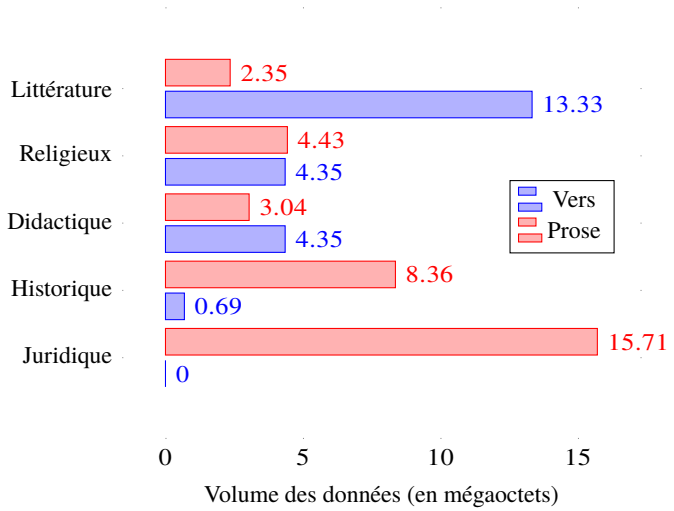


FIGURE 1.4 – Distribution des données actuellement disponibles (9e-15e s.) selon leur domaine et leur forme

Ce volume de données est remarquable pour une langue ancienne, mais il reste faible en comparaison des données disponibles pour les états de langues contemporaines. Nous avons présenté les types principaux d'écrits de FMed, et nous abordons à présent leurs particularités linguistiques.

1.2 Particularités linguistiques

Le FMed ne peut être compris par des locuteurs du français qui ne l'ont pas étudié (Prévost 2020). Cette situation est due en grande partie à la graphie variable des mots, inhabituelle pour un lecteur contemporain, et à des particularités morphologiques et syntaxiques majeures.

1.2.1 Ordre des mots et déclinaisons

En FMed, l'ordre des constituants majeurs est libre, même si SVO devient très tôt l'ordre dominant. Par ailleurs, le sujet peut être omis si le référent est identifiable et jugé suffisamment "accessible" (Marchello-Nizia et Prévost 2020). Par exemple, dans les premiers vers de *La Chanson de Roland* [1], le sujet "li reis Marsilie" est introduit une première fois, et il n'est pas repris dans les deux phrases suivantes (cf. exemple 1.1).

Li reis Marsilie *esteit en Sarraguce.*	**Le roi Marsile** se tient à Saragosse.
Alez en est en un verger suz l'umbre.	**[Il]** s'en est allé dans un verger, à l'ombre.
Sur un perrun de marbre bloi se culchet.	Sur un perron de marbre bleu **[il]** se couche...

Tableau 1.1 – *Incipit* de *La Chanson de Roland*

L'expression du sujet devient plus courante à partir du 13e siècle (Marchello-Nizia et Prévost 2020), avec un emploi plus fréquent des pronoms.

La souplesse de l'ordre des mots affecte aussi les compléments de nom. Ils sont plutôt placés à droite du nom qu'ils modifient, mais ils peuvent aussi être à gauche, avec ou sans préposition. Leur antéposition à gauche est cependant limitée à quelques noms comme *Dieu*.

(1) Seignors fait el por **Deu merci** Saintes reliques voi ici
NOMcom VERcjg PROper PRE NOMpro NOMcom ADJqua NOMcom VERcjg ADVgen
'Seigneurs, dit-elle, par **la grâce de Dieu**, je vois ici les saintes reliques.'
Tristan de Béroul, v. 4 197 (fin 12e s.)

L'ordre libre des constituants n'est pas compensé par une déclinaison nominale strictement respectée, comme c'était le cas en latin. L'AF hérite du latin une déclinaison à deux cas (contre six cas en latin) : le cas sujet pour les sujets, les attributs du sujet et les vocatifs, et le cas régime pour toutes les autres fonctions. Mais dès les premiers textes, et surtout à partir de la fin du 12e siècle, il est rare que la déclinaison soit strictement respectée, et c'est souvent le cas régime qui se substitue au cas sujet. Progressivement, à partir du 13e siècle, la déclinaison perd du terrain pour finalement disparaître, au profit du cas régime (Marchello-Nizia 1979). D'autres indices permettent d'identifier les fonctions, dont la valence verbale (Schøsler 1984).

Toutes les combinaisons des constituants majeurs n'apparaissent pas dans les textes de manière égale. Jusqu'au 11e siècle, l'ordre majoritaire est SOV, hérité du latin. Son usage recule progressivement pour laisser la place à SVO, qui est aujourd'hui encore l'ordre standard dans les phrases déclaratives. Les fréquences d'apparition des combinaisons de constituants majeurs varient selon les textes. L'évolution de la syntaxe du FMed n'est pas linéaire, et on observe aussi des phénomènes ponctuels. Par exemple,

1. L'extrait choisi de *La Chanson de Roland*, ainsi que sa traduction, sont disponibles sur *Wikisource* à cette adresse : https://fr.wikisource.org/wiki/La_Chanson_de_Roland/Joseph_B%C3%A9dier/La_Chanson_de_Roland/Bilingue/001-050.

l'ordre OSV, très rare au 13e siècle, connaît un pic d'utilisation aux 14e et 15e siècles (Combettes et Prévost 2011 ; Marchello-Nizia 2008).

1.2.2 Variation dialectale

Le FMed est marqué par la variation dialectale, même si un texte d'un dialecte donné ne présente pas toutes les particularités de son dialecte (Hasenohr et Raynaud de Lage 1993). Une langue littéraire commune est utilisée, et de nombreux textes ne sont pas, ou peu, marqués par des traits dialectaux. Cependant, ce qu'on désigne par "français médiéval" est en réalité un ensemble de parlers régionaux, les langues d'Oïl.

Ces *scriptas* régionales se distinguent avant tout par leur morphologie. Par exemple, certaines formes de pronoms personnels sont particulières à un dialecte. Les formes en *-i*, comme *mi* et *ti* (en FC "moi" et "toi") ne sont présentes que dans des textes en picard (Buridant 2000). On trouve des formes en *-us*, comme *nus* et *vus* en anglo-normand. L'utilisation syntaxique de ces pronoms peut être différente d'un dialecte à un autre. Ainsi, *lor* est normalement atone, et n'apparaît donc qu'à proximité du verbe conjugué. Cependant, il est souvent en position tonique, derrière une préposition, dans les dialectes de l'est et du centre-est.

Les pronoms ne sont pas la seule catégorie marquée par la variation dialectale. Les terminaisons des infinitifs, les conjugaisons et les articles sont aussi affectés (Buridant 2000). Même si la graphie est marquée par la prononciation du parler régional, les mots d'un même texte peuvent avoir plusieurs graphies, car il n'y a pas encore de conventions graphiques.

En comparaison, la syntaxe est peu soumise à la variation dialectale. Les grammaires traditionnelles identifient quelques "dialectalismes". Par exemple, les propositions hypothétiques sont normalement à l'indicatif, mais en anglo-normand, elles sont souvent au subjonctif (Buridant 2000). En picard, l'interrogation indirecte n'est pas introduite par *se*, et les conjonctions de subordination *com* et *que* sont souvent échangées.

Les traits dialectaux sont peu à peu effacés par les imprimeurs à partir du 15e siècle. Ils cherchent à donner de la régularité au français, créant ainsi de premières normes, qui ne sont pas encore l'orthographe, mais qui peuvent être appelées "orthotypie" ou "orthotypographie" (Siouffi 2020b).

1.2.3 Variation graphique

La variation graphique est causée en partie par la variation dialectale. Cependant, elle résulte aussi de la difficulté à traduire les sons du français à partir de l'alphabet latin, qui ne permet plus une adéquation entre lettres et phonèmes (Blanche-Benveniste et Chervel 1969). Le français, langue sans norme graphique fixe, devient alors une langue "à orthographe", que Cazal, Parussa et Llamas-Pombo (2020) (p. 512–513) définissent ainsi :

langue pour laquelle les graphèmes cumulent une fonction phonogrammatique (ils notent un phonème) et une fonction soit diacritique (ils aident à sélectionner une valeur phonématique parmi les valeurs possibles d'un graphème) soit idéographique (ils singularisent la forme écrite d'un mot pour faciliter l'identification de son signifié).

En l'absence d'une réelle orthographe, les scribes trouvent individuellement des solutions pour représenter les sons. On observe souvent l'alternance entre quelques lettres dans un même texte, notamment *i/y* et *k/q* (et parfois *c* selon le contexte). Par exemple, *Yseult* est parfois écrit avec un *i*. Le pronom conjoint

locatif ou datif *y* peut aussi être écrit *i*. La conjonction de coordination *que* se trouve sous cette forme, ou écrite *ke*, parfois *c'*.

1.2.4 Domaines et genres d'écrits

Suivant la classification adoptée pour la *BFM* (*Base de français médiéval*), nous répartissons les textes en cinq domaines majeurs, définis selon leur destination principale et le domaine d'activités auquel ils se rattachent : religieux (édifier), littéraire (divertir), juridique (réguler la vie sociale), historique (consigner/relater les événements du passé), didactique (enseigner, instruire). Ils regroupent différents genres, qui ont évolué jusqu'à aujourd'hui. Par exemple, le théâtre est non seulement présent dans la littérature (farce, comédie, tragédie), mais aussi dans les domaines religieux (drame liturgique, mystère) et didactique (moralités). Nous pouvons attribuer certains traits distinctifs à ces groupes.

La syntaxe évolue considérablement à travers la période, et elle présente des particularités dans les différents domaines. Elle est assez simple dans les chansons de geste, marquées par un recours important à la parataxe (Siouffi 2020b), comme dans cet exemple (*La Chanson de Roland*, v. 139-141) :

Li empereres en tint sun chef enclin.	L'empereur garde la tête baissée.
De sa parole ne fut mie hastifs,	Sa parole jamais ne fut hâtive.
Sa custume est qu'il parolet a leisir.	Telle est sa coutume, il ne parle qu'à son loisir.

A partir du 12e siècle, le nombre de subordonnées rejoint celui des principales. La syntaxe devient plus complexe encore aux 14e et 15e, où de nombreux textes antiques sont retranscrits en prose, amenant les scribes à adopter la période latine (Lecointe 1997).

(2) Oyant le roy d' Espaigne ce que le roy de France
VERppa DETdef NOMcom PRE NOMpro PROdem PROrel DETdef NOMcom PRE NOMpro
luy avoit dit, luy respondit...
PROper VERcjg VERppe PROper VERcjg
'Ayant entendu ce que le roi de France lui avait dit, le roi d'Espagne lui répondit...'
Roman de Jean de Paris, p. 92 (1461)

Les textes juridiques sont marqués par une syntaxe simple, contenant peu de subordonnées (Siouffi 2020b). Ils sont parfois redondants, pour échapper à l'ambiguïté, ce qui se traduit notamment par un emploi accru des pronoms sujets (Goux et Larrivée 2020), qui sont beaucoup moins utilisés dans les autres types d'écrits.

Les textes didactiques sont au contraire plutôt concis, ce qui provoque une certaine ambiguïté. Les auteurs de ces écrits équipent la langue en vocabulaire pour traiter de sujets techniques et scientifiques. L'influence du latin est conséquente, car on compte une part importante de traductions d'ouvrages antiques dans ce domaine. Elle est renforcée aux 14e et 15e siècles par la création de doublets synonymiques, donnant une origine savante à une partie du lexique.

On trouve des textes historiques dès le 12e siècle, notamment du fait de la conquête de l'Angleterre. Ce domaine se développe pendant les croisades. Cependant, son essor réel est tardif, corrélé à la baisse de l'utilisation du latin comme langue savante au 15e siècle.

1.2.5 Vers ou prose

La part de textes en vers est bien supérieure à celle du FC, mais elle va en diminuant au fil des siècles : d'abord faiblement au 12e siècle, et davantage par la suite. Nous proposons ici une brève comparaison de la syntaxe des vers et de la prose. Il a été énoncé précédemment que les premiers textes, en vers, sont dotés d'une syntaxe plutôt simple, même si ce n'est pas systématique. Ces écrits suivent le rythme de la déclamation, car leur transmission est traditionnellement faite à l'oral. Pour limiter la surcharge cognitive des locuteurs, c'est une structure informationnelle simple qui est adoptée (Combettes 2020) : les éléments de la proposition sont généralement brefs et ils s'enchaînent dans des constructions régulières. En effet, les propositions sont liées avec le contexte antérieur, ce qui place le thème en tête d'énoncé, et le deuxième élément est habituellement le verbe. La prose narrative en AF est très influencée par l'oral, sa syntaxe ressemble donc à celle des textes en vers.

Cependant, les textes en prose couvrent plus de genres, et ils présentent donc une grande variété syntaxique. La transcription de textes latins en prose vernaculaire permet aussi d'intégrer de nouvelles constructions dans la langue, la rendant plus complexe. Par exemple, le style périodique, qui se développe en MF, est inspiré de la période latine. L'émergence du type argumentatif rend plus diverse l'expression dans les textes en prose, notamment par le biais de la topicalisation. Les liaisons entre phrases peuvent aussi être implicites.

Le FMed est donc bien une succession d'états de langue fortement soumis à la variation. Les descriptions rapportées ici, extraites de travaux de grammairiens et de linguistes, n'ont été possibles que grâce à l'accès direct à des textes. Nous disposons désormais de plusieurs corpus, pour partie annotés, que nous présentons dans la section suivante.

1.3 Corpus de français médiéval

On compte actuellement plusieurs centaines de textes de FMed numérisés et librement accessibles. Le volume de données annotées en syntaxe reste cependant encore assez faible. Dans cette section, nous présentons les ressources qui peuvent être utilisées pour des travaux d'analyse syntaxique automatique.

1.3.1 Corpus et bases textuelles

Nous avons recensé huit corpus encore non annotés en syntaxe (cf. tableau 1.2). Leurs volumes sont très disparates. Les plus utilisés sont le *Nouveau Corpus d'Amsterdam* (*NCA*), la *Base de français médiéval* (*BFM*) et l'*Anglo-Norman Database* (*AND*).

Le *NCA* est la base accessible la plus ancienne. Elle regroupe un million de mots issus d'une précédente version du corpus. Celle-ci a été annotée manuellement en parties du discours et en catégories grammaticales par l'équipe d'Anthonij Dees. Le corpus actuel est lemmatisé et distribué sous format *XML*.

Le corpus *BFM 2019* contient près de 4,7 millions de mots annotés en parties du discours avec le jeu d'étiquettes *Cattex* (Guillot, Prévost et Lavrentiev 2013), développé pour ces états de langue, et adapté au format d'annotation du *SRCMF* (voir section suivante). Elle est intégrée à la plate-forme *TXM* (Heiden, Magué et Pincemin 2010), via un portail dédié également accessible en ligne[2]. Son annotation

2. L'adresse de la plate-forme est http://txm.ish-lyon.cnrs.fr/bfm/

Corpus	Annotation	Nb. txt	Période	Dialectes
NCA (Stein, Kunstmann et Glessgen 2011)	POS lemmes morpho.	299	1150-1350	variété
BFM (Guillot, Heiden et Lavrentiev 2018)	POS lemmes	170	9-15e s.	variété
AND (Rothwell et Trotter 2005)	non	78	1112-1440	anglo-nd
Doc. ling. galloromans (Glessgen 2003)	non	6 346	13-15e s.	variété
Tx. légaux de Normandie (*CRISCO*)	non	8	1150-1440	normand
Geste (Camps, Albarran et al. 2016)	POS lemmes morpho.	32	12-14e s.	variété
Chartes (van Reenen, Wattel et van Mulken 2006)	non	1	1270-1300	champenois
OpenMedFr (Wrisley 2018)	non	19	11-15e s.	variété

Tableau 1.2 – Corpus de français médiéval

est partiellement vérifiée. Son format d'annotation a été repris pour l'annotation du corpus *Geste* (Camps, Albarran et al. 2016).

La base de l'*AND*, quant à elle, est exploitable à partir du dictionnaire d'anglo-normand en ligne (Rothwell et Trotter 2005). Ces corpus sont privilégiés par les utilisateurs, car ils rassemblent de grands volumes de données et ils peuvent être exploités grâce à un moteur de recherche dédié. L'annotation de corpus de langue ancienne est très coûteuse et reste rare, notamment l'annotation syntaxique, car les spécialistes de ces états de langue sont peu nombreux et l'annotation de tels états de langue impose de constituer un guide d'annotation qui s'adapte au changement linguistique, ce qui demande un effort supplémentaire conséquent.

1.3.2 Corpus arborés

De nombreux textes de FMed sont accessibles en format numérique, mais peu d'entre eux bénéficient d'une annotation syntaxique manuelle. Il existe actuellement deux corpus arborés (ou *treebanks*) comprenant des textes de cette période.

MCVF

Le *MCVF* (*Modéliser le changement : les voies du français*, *MCVF*) est le premier et le plus grand corpus arboré de FMed, avec 361 283 mots. Ce projet a été mené par France Martineau et son équipe de 2005 à 2009. Le corpus est constitué de textes intégraux annotés manuellement en constituants, selon le modèle du *Penn treebank* (Taylor, Marcus et Santorini 2003). D'autres corpus arborés diachroniques ont été constitués en parallèle pour l'anglais et le portugais. Ce sont les premiers du genre, et un objectif des équipes était de permettre la comparaison entre ces états de langue anciens.

Les textes ont été choisis pour représenter l'évolution du français et sa variété dialectale. On compte vingt-neuf textes pour le FMed, qui couvrent six dialectes. Un corpus équilibré permet d'étudier l'évolution du français en dégageant des tendances dans les états de langue successifs. L'équipe du *MCVF* s'est par exemple intéressée au passage à l'ordre SVO, à la perte du sujet nul et à la montée du clitique objet et à la fixation de la position des adverbes négatifs.

Pour rendre compte des états de langue anciens, les *s* longs ont été conservés. Les agglutinations (ex. *lamour* pour "l'amour") apparaissent aussi, mais elles sont résolues et encodées pour garantir l'efficacité de la recherche par un moteur de recherche.

Les annotateurs ont procédé en deux étapes. Premièrement, ils ont effectué une annotation semi-automatique en parties du discours, et chaque annotation a été vérifiée deux fois. Contrairement au format *Cattex* (Guillot, Prévost et Lavrentiev 2013), les formes agglutinées (ex. *du* pour "de le") ne reçoivent pas une étiquette spécifique (ex. PRE.DETdef), mais deux étiquettes. Ces formes sont séparées pour préparer l'annotation syntaxique. Le protocole d'annotation a été conçu pour rester constant à travers le corpus. Deuxièmement, l'équipe a procédé à l'annotation syntaxique à partir des résultats de la première étape. Le guide d'annotation se base sur la théorie du gouvernement et du liage (Chomsky 1993) et attribue une place supposée aux sujets nuls. Ce corpus arboré a servi pour plusieurs études, notamment celles de Martineau (1990), Simonenko, Crabbé et Prévost (2020) et Stein (2018).

SRCMF

Le *Syntactic Reference Corpus of Medieval French* (*SRCMF*, Stein et Prévost (2013))[3] est un corpus arboré d'environ 250 000 mots, manuellement annoté en dépendances. Ses quinze textes proviennent de la *BFM* et du *NCA* (cf. tableau 1.3). L'équipe du *SRCMF* a aussi choisi de "panacher" les textes pour représenter les différents états de langue (Prévost 2015). Pour obtenir un ensemble équilibré, les textes supérieurs à 40 000 mots ne sont présents que sous la forme de trois extraits (début, milieu et fin du texte). Cinq dialectes sont couverts, ainsi qu'une variété de domaines. Le corpus ne représente pas tous les domaines et tous les dialectes à chaque siècle, car la production écrite n'est pas très diversifiée au Moyen Age, comme cela a été exposé dans la première section de ce chapitre.

Texte	Date	Domaine	Forme	Dialecte	Nb. mots
Serments de Strasbourg	842	juridique	prose	indéfini	115
Sequence de Sainte Eulalie	881	religieux	vers	indéfini	189
Vie de Saint Alexis	vers 1050	religieux	vers	normand	4 868
Passion de Clermont	ca. 1100	religieux	vers	franco-occ.	2 842
Vie de Saint Léger	ca. 1100	religieux	vers	franco-occ.	1 388
Chanson de Roland	ca. 1100	littérature	vers	normand	28 997
Lapidaire en prose	milieu 12e s.	didactique	prose	anglo-nd	4 765
Tristan de Beroul	1165-1200	littérature	vers	franco-pic.	27 052
Yvain, Chr. de Troyes	1177-1181	littérature	vers	champenois	41 702
Quatre Livres des Rois	fin 12e s.	religieux	prose	anglo-nd	13 061
Aucassin et Nicolete	fin 12e s. – déb. 13e s.	littéraire	mixte	picard	9 946
La Conqueste de Constantinople	après 1205	historique	prose	picard	33 969
Queste del Saint Graal	ca. 1220	littérature	prose	indéfini	40 636
Miracles, G. de Coinci (*NCA*)	1218-1227	religieux	vers	picard	4 963
Miracles, G. de Coinci	1218-1227	religieux	vers	picard	17 455
Roman de la Rose, J. de Meun	1269-1278	didactique	vers	indéfini	19 462

Tableau 1.3 – Textes du *SRCMF*

Le format d'annotation syntaxique[4] a été créé par l'équipe et ce sont les étiquettes morphologiques *Cattex* qui sont utilisées pour l'annotation morpho-syntaxique. Elles proposent une étiquette par token,

3. Le corpus arboré *SRCMF* est accessible en ligne à cette adresse : http://srcmf.org/.
4. Le guide d'annotation est disponible en ligne à cette adresse : http://srcmf.org/fiches/index.html.

même s'il s'agit d'un token double (ex. *sel* peut être étiqueté CONsub.PROper). Ce token est dupliqué dans le *SRCMF* pour permettre d'y assigner deux dépendances syntaxiques, en l'occurrence celles de relateur non-coordonnant (*RelNC*) et d'objet (*Obj*), comme dans cet exemple :

(3) Sel pois trover a port ne a passage
CONsub.PROper VERcjg VERinf PRE NOMcom CONcoo PRE NOMcom
'Si je le puis trouver aux défilés et aux passages'
Chanson de Roland, v. 657 (ca. 1100), trad. Léon Gautier

Le même procédé est aussi utilisé pour les pronoms relatifs, qui ne sont pas des formes complexes, mais ils sont à la fois *RelNC* et sujets ou objets de la proposition subordonnée.

Le schéma d'annotation syntaxique du *SRCMF* ne dépend pas d'une théorie syntaxique spécifique, contrairement au *MCVF* qui est largement influencé par la grammaire générative. Il n'y a donc pas de nœuds vides, ni de traces dans ce corpus. Ce format d'annotation permet de croiser les dépendances, par exemple en cas de constructions discontinues. La tête d'une phrase est un verbe conjugué. Il existe des séquences, appelées "non-phrases" qui sont des énoncés indépendants dont la tête n'est pas un verbe conjugué, comme les phrases nominales. On ne trouve pas de coordination entre verbes conjugués de principales ; chaque verbe conjugué est la tête d'une phrase. L'annotation du *SRCMF* ne se base pas sur des éditions critiques strictes, car la ponctuation a été retirée. De manière générale, les têtes lexicales sont préférées aux têtes fonctionnelles, qui peuvent ne pas être exprimées. Les relateurs, les déterminants et les prépositions ne sont en effet pas toujours exprimés. Par exemple, on trouve un complément du nom non introduit, "Dieu", dans *Miracles de Nostre Dame* de Gautier de Coinci (v. 1 277) :

(4) La mere Dieu merci cria
DETdef NOMcom NOMpro NOMcom VERcjg
'[Il] cria merci à la mère **[de]** Dieu'
Miracles de Nostre Dame de Gautier de Coinci, v. 1 277 (début 13e s.)

Toutefois, dans les formes verbales complexes, c'est l'auxiliaire qui est la tête.

La campagne d'annotation a été menée en double aveugle. Les annotateurs confrontaient ensuite leurs analyses et décidaient de l'annotation à conserver. Un forum a été mis en place pour permettre des échanges, ce qui permet de garder la trace des discussions (Fort 2012). En cas de désaccord, les superviseurs prenaient la décision. Le treebank est disponible au format *TigerXML* pour *TigerSearch* (Lezius 2002), dont le moteur est intégré à la plate-forme *TXM*.

Le *SRCMF* a été converti au format *Universal Dependencies* (*UD*, Nivre, De Marneffe et al. (2016)). Les frontières de phrase du corpus original ont été conservées, même si elles ne correspondent pas toutes à des séquences séparées par une marque de ponctuation forte. Le système de mises à jour *UD* permet de corriger progressivement les annotations, selon les besoins. Le *SRCMF* est l'une des premières ressources historiques disponibles dans ce format. Dans la prochaine section, nous donnerons un aperçu des avancées dans la création de corpus arborés pour les autres langues anciennes.

1.4 Autres ressources de langues anciennes

Le développement de ressources syntaxiques pour des états de langue anciens est rare, mais on constate une hausse de leur production. Cela est en partie dû aux progrès réalisés dans les techniques de numérisation de manuscrits (HTR) et d'imprimés (OCR). Avec le développement des humanités numériques,

de telles ressources sont de plus en plus demandées. Les besoins d'annotation diffèrent selon les projets. Nous présentons ici les corpus arborés disponibles pour des états de langue produits dans les mêmes conditions que le FMed, c'est-à-dire avant la diffusion de l'imprimerie. Nous ne décrirons pas les choix faits pour toutes ces ressources, mais les problèmes soulevés dans ces études rejoignent ceux exposés ici. La plupart de ces corpus sont annotés manuellement (Demske et al. (2004) pour l'allemand, Lee et Kong (2012) pour le chinois et Rögnvaldsson et al. (2012) pour l'islandais). Certains travaux mettent à profit des outils (Dukes et Buckwalter 2010) et utilisent des *treebanks* existants pour en annoter de nouveaux (Lee et Kong 2014).

1.4.1 Anglais médiéval

L'anglais est la seule langue à disposer de ressources syntaxiques comparables à celles du français, en ce qui concerne notre période. L'ensemble *English Parsed Corpora Series* recouvre trois *treebanks* d'anglais médiéval et deux autres d'anglais moderne. Il s'agit des ressources suivantes :

— *York-Helsinki parsed corpus of Old English poetry* (*YCOEP*, *The York-Helsinki parsed corpus of Old English poetry (YCOEP)* (2001))

— *Penn-Helsinki Parsed Corpus of Middle English II* (*PPCME2*, Kroch, Taylor et Santorini (2000))

— *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (*YCOE*, Traugott et Pintzuk (2008))

Corpus	Période	Dialectes	Nb. textes	Nb. mots
YCOEP	450-1100	-	14	71 490
PPCME2	1150-1500	Kentish, South, West/East Midlands, North	56	1,3 Mi
YCOE	9e-12e s.	Anglian Mercian, Kentish, West Saxon	100	1,5 Mi

Tableau 1.4 – Corpus arborés d'anglais médiéval

Ils reposent sur le même schéma d'annotation au format *Penn*, mais il est adapté en fonction des périodes. Ce travail d'annotation a été fait en parallèle de celui du *MCVF*. Pour analyser des états de langues anciens, des choix ont été faits, comme celui de ne pas représenter de syntagme verbal, car ses frontières sont trop floues. Certaines catégories ne sont pas homogènes : ce qu'elles annotent dépend de la période du texte.

Tout comme pour les données du FMed, les annotateurs se sont heurtés à l'annotation de tokens dont les frontières sont différentes de la langue contemporaine. Certaines formes ont été fusionnées, comme *never the less*, dont chaque token est annoté, puis les trois tokens ont été rassemblés pour former un ensemble adverbial, utilisé comme tel dans l'annotation syntaxique. À l'inverse, certaines formes sont des contractions, comme *bicause*, qui est analysé comme deux mots, *bi* et *cause*, formant un syntagme prépositionnel.

L'objectif du *treebank YCOE* n'est pas de proposer une analyse syntaxique totale des phrases, mais d'aider les utilisateurs à faire des requêtes dans le corpus. Les annotations sont donc moins complètes que dans les précédentes ressources. Le *PPCME2*, en revanche, a une annotation complète, ce qui a permis de procéder à des travaux de traitement automatique, comme l'entraînement d'un lemmatiseur (Trips et Percillier 2020). Le guide d'annotation a été adapté à l'islandais pour la constitution d'un *treebank* diachronique (Rögnvaldsson et al. 2012).

1.4.2 Portugais médiéval

Rocio et al. (2003) ont fait l'hypothèse qu'il était possible d'annoter un état d'une langue avec des outils et des ressources robustes conçus pour un autre état de cette même langue. Ils ont annoté un corpus de portugais médiéval (12-13e s.) avec une chaîne de traitement du portugais contemporain. Pour cela, le parseur et la grammaire *DCG* qu'il utilise ont été adaptés à la langue médiévale. Les auteurs ont également remplacé le lexique du portugais contemporain par un analyseur lexical créé pour ces travaux.

Ils ont ainsi obtenu une annotation partielle du corpus. Quelques obstacles techniques ont empêché l'adaptation totale des outils existants. D'une part, le portugais médiéval présente un ordre variable des constituants principaux. Prendre en compte toutes les possibilités aurait demandé des changements conséquents dans la grammaire. D'autre part, la couverture lexicale de l'analyseur était trop faible pour l'analyse syntaxique. Les informations morphologiques disponibles étaient parfois inexactes ou insuffisantes, et il aurait fallu des informations syntaxiques comme la valence pour guider l'analyse syntaxique.

1.4.3 *PROIEL*

Le projet *PROIEL* (*Pragmatic Resources in Old Indo-European Languages*, Haug et Jøhndal (2008)) a abouti à la création d'un *treebank* parallèle de grec ancien, de latin (antique), de gotique, d'arménien classique et de vieux-slave (*Old Church Slavonic*). Les textes choisis sont les traductions du *Nouveau Testament*, qui sont les premiers écrits en gotique, arménien et vieux-slave. Elles se basent sur la traduction grecque de la *Bible*. Les sous-corpus de grec ancien et de latin regroupent aussi des textes antiques. Ce corpus couvre ainsi treize siècles. Le *treebank TOROT* (Berdicevskis et H. Eckhoff 2020) est une expansion de *PROIEL* pour les langues slaves.

Langue	Période	Nb. tokens
grec ancien	2e au 1er s. av. JC	250 455
latin	1er s. av. JC – 5e s.	225 064
gotique	4e s.	23 513
arménien classique	5e s.	57 211
vieux-slave	11e s.	58 269

Tableau 1.5 – Distribution linguistique du *treebank PROIEL*

Le pré-traitement de ces textes anciens pose les mêmes problèmes que dans les autres corpus. Premièrement, la segmentation en phrases a été effectuée manuellement, reprenant parfois les choix des éditeurs, car la ponctuation seule ne permet pas de déterminer la frontière des phrases. Deuxièmement, ces textes comportent de nombreuses formes contractées, qui ont été dissociées pour les besoins de l'analyse syntaxique.

L'annotation syntaxique suit globalement le modèle du *Prague Dependency Treebank* (*PDT*, Hajic (1998)) car la syntaxe en dépendances se prête bien aux langues à ordre libre, qui constituent la majorité du corpus (H. Eckhoff et al. 2018). Ce modèle se base sur le cadre *LFG* (Haug 2012). On note cependant quelques changements par rapport à l'annotation du *PDT*. Par exemple, les objets directs sont distincts des obliques et des agents passifs. Les auteurs ont fait le choix d'utiliser des têtes fonctionnelles, hormis pour le traitement des auxiliaires et des déterminants. Des nœuds vides sont ajoutés pour gérer les cas d'ellipses. Ce *treebank* se démarque par le choix d'utiliser des dépendances secondaires pour annoter les sujets des verbes à l'infinitif et les dépendants partagés dans les cas de coordination. La structure

argumentale est donc plus riche, mais la représentation d'une phrase n'est plus un arbre. Cela rend tout traitement automatique de ce *treebank* plus complexe. Afin de préserver la cohérence de l'annotation, si plusieurs analyses sont possibles, c'est celle qui s'applique le mieux aux données les plus anciennes qui est choisie. Cela permet de faire des requêtes sur l'ensemble du corpus et d'obtenir des résultats fiables.

Il est généralement considéré que l'annotation de textes de langues anciennes est très coûteuse à cause de l'absence d'une réelle compétence langagière. Des ressources contextuelles ont été mises à disposition des annotateurs pour limiter ce coût (Bamman et Crane 2011). Malgré cela, l'annotation du latin reste lente. L'équipe *PROIEL* a mesuré le nombre de mots annotés par heure sur les textes de latin et a comparé ce rendement à celui d'autres campagnes. Pour l'anglais et le chinois, l'annotation est respectivement sept et trois fois plus rapide. Le latin, bien que pratiqué au delà de la période médiévale, est très marqué par le style des auteurs. Les idiosyncrasies rendent l'analyse des phrases ambiguës (Bamman, Passarotti et al. 2008), ce qui justifie la lenteur de l'annotation.

Une conversion au format *UD* est en cours. Cela implique en partie une perte d'informations. Les dépendances secondaires sont écartées, et une transition est faite pour un emploi de têtes lexicales. Ce format facilite l'utilisation du *treebank* dans des applications de traitement automatique et permet la comparaison avec d'autres ressources. Une deuxième ressource de latin, annotée avec le même guide que celle de *PROIEL*, a été convertie au format *UD* (Cecchini, Korkiakangas et Passarotti 2020).

1.4.4 Treebanks en constitution

Des campagnes d'annotation de corpus de langue ancienne sont en cours actuellement. Les équipes font des constats similaires en ce qui concerne la complexité de l'analyse de ces données. Pour réduire le coût de l'annotation, il est possible d'adapter des outils qui ont été créés pour des états de langue contemporains. C'est le choix qu'ont fait Estarrona et al. (2020) pour l'analyse morpho-syntaxique du basque moderne (15e-18e s.). L'annotation syntaxique n'a pas encore été présentée, mais leurs travaux incluent un phase de normalisation (cf. définition dans la partie 2.1.3).

La campagne d'annotation du *treebank* de latin médiéval *UDante* (Cecchini, Sprugnoli et al. 2020) est en cours, et elle permet de mettre à jour le guide d'annotation *UD* pour cette langue.

Constituer des corpus arborés pour des états de langue anciens présente donc des difficultés supplémentaires. Les équipes ont partagé le procédé de traitement de leurs corpus, depuis la segmentation en mots et en phrases jusqu'aux choix d'annotation syntaxique. Ces tâches sont souvent rendues plus complexes par la présence de dialectes et par l'absence de normes graphiques. Faire des choix d'annotation acceptables pour tous les états de langues, parfois éloignés dans le temps, est une tâche difficile. Ces travaux bénéficient déjà de pré-traitements, comme l'analyse morpho-syntaxique automatique, pour réduire les coûts des campagnes. Le développement de ces techniques s'est accéléré ces dernières années. Le prochain chapitre propose un état de l'art de ces outils, ainsi que la présentation de travaux préliminaires d'annotation morpho-syntaxique pour le FMed.

2 Approches statistiques

Cette thèse doit beaucoup à Isabelle Tellier, décédée le 1er juin 2018, qui en était l'une des directrices. Elle a participé à élaborer son sujet, et en a initié les premières expériences d'annotation automatique avec les champs aléatoires markoviens (*CRF*). Son expérience et son enseignement ont continué à guider mon travail.

Même si les *CRF* n'ont pas été retenus par la suite, leur découverte et leur utilisation m'ont permis, d'une part, de mieux comprendre les données et, d'autre part, d'établir une référence pour les expériences suivantes. Ce chapitre contiendra donc la présentation de cette méthode et des travaux effectués. Vers 2014, le domaine du traitement automatique du langage (TAL) a connu un bouleversement avec l'arrivée de l'apprentissage automatique profond (*deep learning*). Les modèles statistiques, jusqu'alors de taille plutôt modeste et raisonnablement interprétables, ont été peu à peu délaissés au profit de modèles neuronaux très coûteux avec un fonctionnement de "boîtes noires". Ces méthodes sont utilisées dans le cadre du projet ANR *Profiterole* (cf. partie 6.3.1).

Un travail actuel en TAL qui n'utilise pas des méthodes statistiques ou neuronales fait désormais figure d'exception et demande à être justifié. Notre but est de mettre à disposition des outils dédiés au FMed, constitué d'états de langue fortement soumis à la variation. Des méthodes non symboliques se heurtent à un plafond de performance qu'il n'est pas aisé de dépasser. Notre hypothèse est qu'une grammaire électronique pourra s'adapter aux différents états de langue et fournir des analyses pour des phénomènes complexes et peu représentés. Un tour d'horizon des différents travaux d'analyse statistique et de nos expériences s'impose tout d'abord.

2.1 Approches en TAL pour les langues anciennes

2.1.1 Analyse automatique de corpus hétérogènes

Du fait de leur empan chronologique et des multiples sources de variation, les corpus de langue ancienne sont généralement des corpus hétérogènes. La présence de différents états de langue les rend difficiles à traiter car cela ajoute de la variation dans les données. D'une part, la nature hétérogène des données pose problème aux systèmes statistiques, qui reposent sur l'hypothèse que les données d'apprentissage suffisent à généraliser des exemples pour en faire un modèle efficace. D'autre part, comme nous l'avons expliqué dans la partie précédente, l'annotation manuelle comporte des défis du fait de ces disparités dans les données. Or la cohérence de l'annotation est très importante pour produire des analyses de qualité (Fort 2012). Le traitement de ces corpus s'inscrit ainsi dans la problématique d'adaptation au domaine, qui est un défi pour les systèmes automatiques (Dredze et al. 2007). On peut en effet considérer qu'un corpus hétérogène regroupe plusieurs "domaines" (les états de langue). La différence entre la distribution des domaines dans le corpus de référence et dans les données à traiter rend le traitement automatique difficile (Ben-David et al. 2006).

Yu (2018) recense quatre approches pour cette tâche :

1. Approche semi-supervisée ou non-supervisée. Blitzer, McDonald et Pereira (2006) disposent de données non annotées des domaines source et cible et s'en servent pour apprendre des représentations communes appelées "traits pivots".
2. Approche sélectionnant les données (McClosky, Charniak et Johnson 2010 ; Plank et van Noord 2011). Cette démarche peut aussi s'appliquer à des méthodes utilisant des grammaires. Par exemple, Hara, Miyao et Tsujii (2005) utilisent un corpus arboré du domaine cible pour entraîner un modèle de désambiguïsation de leur grammaire *HPSG*.
3. Approche utilisant des ressources lexicales. Le but d'Aubin, Nazarenko et Nédellec (2005) est d'adapter le parseur de Sleator et Temperley (1993) pour annoter un corpus de biologie, qui présente des milliers de tokens inconnus, dont des séquences ADN. Les autrices procèdent à une étape de normalisation et au relâchement de certaines règles pour limiter les échecs d'analyse. Elles traitent aussi des mots inconnus par ajout manuel ou au moyen d'un module qui leur attribue une étiquette selon la terminaison du mot. La méthode de Cohen, Goldberg et Elhadad (2012) ne consiste pas seulement à ajouter des formes spécifiques, mais aussi à renseigner leur attachement syntaxique grâce à l'apprentissage par allocations de Dirichlet latentes (*Latent Dirichlet Allocation*). Rimell et Clark (2008) ré-entraînent l'analyseur morpho-syntaxique de leur parseur (à base de grammaire *CCG*) avec des données annotées manuellement, ce qui rejoint aussi l'utilisation de nouvelles ressources lexicales.
4. Approche utilisant un ensemble de parseurs (Nivre, J. Hall et al. 2007).

Le caractère hétérogène des données dans ces études, bien qu'ayant un impact important sur la qualité de l'analyse, ne vient généralement que du vocabulaire inconnu, et rarement de la morphologie ou de la syntaxe elles-mêmes. Les corpus de langue ancienne que nous souhaitons traiter présentent des difficultés supplémentaires, pour lesquelles des méthodes spécifiques ont dû être développées.

2.1.2 Le cas particulier des corpus de langue ancienne

Les corpus de langue ancienne sont un cas particulier d'hétérogénéité, en particulier s'ils sont diachroniques. Ils ne présentent alors pas seulement des éléments de vocabulaire différents d'un état de langue à l'autre, mais également des particularités morphologiques et syntaxiques (Heiden et Prévost 2002). Ces différences s'observent à la fois entre des ensembles de textes et au sein même des textes, si bien qu'il n'y a aucune garantie de la qualité de l'annotation d'un texte par un modèle entraîné sur un corpus aux caractéristiques externes similaires, alors que ce serait le cas en FC. Les causes de cas variations sont de nature diverse :

D'une part, certaines langues anciennes n'ont pas de graphie fixe. Le FMed en fait partie. La variation graphique est très présente en FMed (cf. section 1.2.3), y compris pour les mots grammaticaux, comme la forme *el*, qui peut être une variante graphique du pronom personnel *elle* ou une contraction de la préposition *en* et de l'article masculin singulier *le*. La segmentation d'un corpus en tokens, ou "tokenisation", doit donc traiter des cas ambigus. Le processus de grammaticalisation vient aussi bousculer les frontières de mots, transformant par exemple des locutions complexes en conjonction de subordination figées, comme *par/por (ce) que/ke* (en deux ou trois tokens), qui devient la chaîne *parce que / porce que* au cours de la période. L'usage des mots évolue, et ils changent donc parfois de catégorie morpho-syntaxique, comme

c'est le cas pour l'adjectif *moult* qui devient un adverbe. Certains acceptent deux étiquettes à la même époque, comme *dame*, qui peut être nom commun ou adjectif qualificatif (Prévost 2011). Les homonymes sont fréquents, y compris au sein de catégories morpho-syntaxiques différentes. Par exemple, *se* peut être un pronom personnel réfléchi, une conjonction de subordination (*si*), ou parfois un adverbe.

D'autre part, on observe une grande variation syntaxique en FMed. L'ordre des mots est très souple, et les distributions des constituants majeurs varie selon les textes et la période. L'expression du sujet est optionnelle, mais son usage s'intensifie dans le temps, en relation avec l'usage croissant de la prose à partir du 13e siècle. La valence des verbes évolue également, comme celle du verbe *morir* (mourir) qui accepte un objet direct en FMed, mais pas en FC.

Ces exemples de variation ne distinguent pas seulement les états de langue définis qui présentent des caractéristiques communes. Comme le FMed est très marqué par les idiolectes (Prévost 2015), chaque texte pourrait constituer un état de langue, et on trouve de la variation au sein d'un même texte. Le FMed ne recouvre, de fait, "aucune réalité langagière homogène" (Prévost 2008). Ces données sont donc très ambiguës, d'une part pour le segmenteur, et d'autre part pour le parseur. Malgré la forte variation observée, les différents états de langue intègrent bel et bien un *continuum* langagier. C'est la continuité des phénomènes dans les textes qui permet de modéliser le FMed (Marchello-Nizia 1995). L'existence d'une forte variation, qui gêne l'analyse automatique, est partagée par d'autres langues anciennes, comme l'allemand médiéval (Barteld, Biemann et Zinsmeister 2018) et le vieux-slave (Pedrazzini et H. M. Eckhoff 2021), pour lesquelles il existe plusieurs méthodes d'étiquetage automatique.

2.1.3 Approches privilégiées

L'ambiguïté inhérente aux corpus de langue ancienne rend leur traitement automatique plus difficile, ce qui se traduit par des scores inférieurs à ceux des langues contemporaines. Cependant, à condition de posséder suffisamment de données annotées, l'annotation automatique peut être envisagée. Par exemple, les premières expériences d'analyse syntaxique de l'AF (Stein 2014) ont donné des résultats encourageants sur le *SRCMF*, avec 82,62% de LAS [1]. Ce score a été comparé à celui du FC, 87,6% (Candito, Crabbé et Denis 2010). L'écart entre ces deux états de langue s'est considérablement réduit dans les expériences suivantes, comme le montrent les résultats de deux des meilleurs parseurs de la compétition *CoNLL 2018* (cf. tableau 2.1).

	French GSD		Sequoia		SRCMF	
	LATTICE	UDPipe	LATTICE	UDPipe	LATTICE	UDPipe
POS	96,52	96,32	98,12	97,56	95,75	96,22
UAS	89,50	88,74	91,81	90,07	91,35	91,72
LAS	86,89	85,74	86,17	88,04	85,51	87,12

Tableau 2.1 – Résultats des parseurs *LATTICE* (Lim et al. 2018) et *UDPipe Future* (Straka 2018) à la compétition *CoNLL 2018*

Toutefois, nous ne disposons pas de telles ressources pour entraîner des modèles dans toutes les langues anciennes, ce qui pousse les équipes à utiliser des techniques complémentaires pour rendre le traitement automatique possible. Pour adapter un parseur à de nouveaux états de langue, des aménagements sont

1. La qualité de l'annotation syntaxique est mesurée au moyen de deux indices : le taux de tokens correctement attachés à leur gouverneur (*Unlabeled Attachment Score, UAS*) et le taux de ces dépendances ayant reçu l'étiquette correcte (*Labeled Attachment Score, LAS*).

nécessaires. Certaines techniques concernent l'apport de nouvelles données pour pallier l'absence ou la faible quantité de données annotées, d'autres permettent d'harmoniser les annotations pour contrebalancer les effets de l'hétérogénéité des données.

Filtrer ou augmenter les données d'apprentissage ?

Généralement, le traitement des langues anciennes pose des défis similaires à celui de langues peu dotées. Seules quelques unes de ces langues disposent de ressources annotées (cf. sections 1.3 et 1.4). Il est donc primordial de tirer le meilleur profit possible des données à disposition, qu'elles soient annotées ou non.

Sélection des données Guibon, Tellier, Prévost et al. (2015) ont fait le choix de sélectionner les meilleurs ensembles de textes pour entraîner un modèle selon les caractéristiques externes du texte cible. On recherche ainsi à entraîner un modèle efficace sur un certain état de langue. Cette démarche a donné de bons résultats pour l'AF, en particulier pour gérer la variation dialectale. Cependant, Pedrazzini (2020) fait le constat inverse : représenter tous les dialectes de son corpus dans son ensemble d'apprentissage donne de meilleurs résultats pour des états anciens du slave. Limiter l'impact de l'hétérogénéité des données lors de l'apprentissage n'est donc pas une solution universelle. La technique de sélection de sous-ensembles de données pour l'apprentissage de modèles peut être coûteuse, et elle n'est pas à considérer pour tous les corpus de langue ancienne, ni pour tous les parseurs.

Approche par transfert Le traitement de corpus hétérogènes et de langue ancienne présente des ressemblances avec celui de langues peu dotées. Les différents états de langue traités peuvent être considérés comme des sous-domaines ayant trop peu de représentants pour être correctement généralisés par le modèle. L'analyse de nouveaux textes aux caractéristiques externes éloignées du corpus d'apprentissage est donc de moins bonne qualité que pour des textes plus proches, faute de données pour en représenter le vocabulaire et la syntaxe. On peut donc faire l'hypothèse qu'augmenter le volume des données disponibles aiderait à représenter ces états de langue et apprendre leur grammaire. Cependant, il est fréquent que les états de langue soient très peu, voire pas du tout, représentés.

On peut alors considérer une technique pour inférer une représentation à partir de données proches, ce qu'on appelle l'approche par "transfert", couramment utilisée pour traiter des langues peu dotées (Agić et al. 2016). Elle peut tirer bénéfice de corpus alignés (McDonald, Petrov et K. Hall 2011), mais il est possible de l'utiliser sans ces ressources, comme Scrivner et Kübler (2012) l'ont fait pour l'ancien occitan. Elles ont eu recours à des corpus arborés de catalan contemporain et d'AF pour entraîner un modèle pour l'ancien occitan. Les deux langues sources, typologiquement proches de la cible, présentent des caractéristiques communes, comme un ordre des mots souple et la possibilité de sujets nuls.

Pré-traitements avec des données brutes Les parseurs qui retiennent le plus d'attention actuellement sont ceux qui intègrent des réseaux de neurones. Ceux-ci peuvent prendre en entrée des plongements lexicaux, qui peuvent être obtenus sur des données brutes. On obtient ainsi une première représentation de la langue, sous forme de vecteurs, qui sert à orienter le parseur lors de l'apprentissage. Cela permet de tirer profit de grands volumes de données non annotées.

Cette technique s'avère très utile pour des états de langue pour lesquelles on dispose de peu de données, comme les langues anciennes. Afin d'améliorer les résultats de l'analyse morpho-syntaxique de l'allemand médiéval et moderne (11e-17e s.), Barteld, Biemann et Zinsmeister (2018) ajoutent des plongements lexicaux (ainsi que les annotations d'un autre analyseur) en amont de l'apprentissage du modèle. Mairey et Aouini (2021) procèdent de manière similaire, car ils manquent de données pour entraîner leurs modèles d'étiquetage morpho-syntaxique de moyen français, de moyen anglais et de latin médiéval. Les plongements lexicaux sont utilisés en entrée de leurs réseaux récurrents à portes bi-directionnels (Cho et al. 2014) et donnent de meilleurs résultats que *TreeTagger*, leur référence.

Modèles	**POS**	**UAS**	**LAS**
Straka, Straková et Hajič (2019)	96.26	91.83	86.75
mBERT	96.19	92.03	87.52
BERTrade-petit	96.60	92.20	87.95
BERTrade-mBERT	97.11	93.86	90.37
BERTrade-FlauBERT	97.15	93.96	90.57
BERTrade-CamemBERT	97.29	94.36	90.90

Tableau 2.2 – Résultats sur l'ensemble de test du *SRCMF-UD*

Plus récemment, des travaux de *parsing* neuronal ont pris place dans le cadre du projet ANR *Profiterole* à partir du parseur de Grobol et Crabbé (2021). Les données mentionnées dans la figure 1.4, qui représentent la grande majorité des textes de FMed disponibles, ont été rassemblées pour entraîner des plongements lexicaux qui ont servi au pré-entraînement du parseur. L'apport de ces données entraîne un gain de performance, en particulier en utilisant des modèles du FC, comme *FlauBERT* et *CamemBERT* (cf. tableau 2.2).

Même si le volume de données annotées à disposition est faible pour apprendre tous les états de langue souhaités, il est possible d'améliorer les résultats de l'analyse syntaxique. Les nouveaux parseurs permettent notamment d'utiliser de grands volumes de données non annotés pour pré-entraîner les modèles. Cependant, il est aussi possible de réduire la confusion du parseur face aux données annotées.

Harmoniser l'annotation

Du fait de leur échantillonnage, les corpus arborés de langues anciennes représentent difficilement la continuité linguistique. L'absence de graphie standardisée, la souplesse de l'ordre des mots et l'évolution langagière sont des exemples de paramètres de variation, qui s'observent à la fois entre textes et au sein même des textes. Or l'hypothèse du traitement automatique des langues est que la régularité du langage permet son traitement par un automate. Les systèmes actuels sont capables d'apprendre des modèles à partir de corpus hétérogènes à condition de posséder assez de données. Leurs résultats peuvent être très bons, comme ceux du parseur *UDPipe 2.0* à la compétition *CoNLL 2018*, qui obtient 87,12% de LAS sur le *SRCMF*. Par contraste, le score obtenu sur le *French Treebank* est de 86,13%. Néanmoins, ces modèles peinent parfois à capturer des connaissances linguistiques, et ils se heurtent à un plafond de performance difficile à dépasser. Cette limitation ne touche pas uniquement les corpus de langues anciennes, mais aussi les langues soumises à la variation dialectale et sans graphie standardisée (Millour et Fort 2019). Rapprocher les tokens de formes plus canoniques est une démarche intéressante pour les données historiques.

Bollmann (2018) la nomme “adaptation des données” (*data adaptation*), par opposition à l’adaptation au domaine (*domain adaptation*). Il en existe deux procédés : la lemmatisation et la normalisation.

Lemmatisation La lemmatisation est le rapprochement d’une forme avec sa “représentation standardisée” (Souvay et Pierrel 2009), c’est-à-dire généralement son lexème. Pour les langues anciennes, il s’agit d’une tâche difficile à cause du manque de données d’apprentissage, de la variation graphique et de la richesse de la morphologie, comme en moyen néerlandais (Kestemont et al. 2016) et en irlandais médiéval (Dereza 2019). Il ne s’agit pas seulement de “faire une généralisation à partir d’une forme fléchie” mais aussi de déduire une forme canonique à partir de variants graphiques (Stein 2014).

Outre son intérêt pour les linguistes, cet apport d’information permet de rattacher les formes fléchies à une forme plus standard et ainsi de faciliter l’apprentissage automatique de modèles d’analyse morpho-syntaxique. Selon Holgado, Lavrentiev et Constant (2021), parmi les quatre analyseurs morpho-syntaxiques disponibles pour l’AF qui sont aussi des lemmatiseurs, c’est LGeRM (Souvay et Pierrel 2009) qui produit les meilleurs résultats. Cet outil dispose en effet d’un vaste lexique et de règles de substitution pour prédire les lemmes. Sa référence de lemmatisation est le *Dictionnaire de Moyen français* (*DMF*), dont les entrées encore présentes en FC sont modernisées.

Cependant, en l’absence de lexique suffisant pour trouver les correspondances entre les formes des textes à traiter et leurs lemmes, la lemmatisation peut être une tâche trop ambitieuse. Bollmann (2018) propose la normalisation comme tâche intermédiaire.

Normalisation Pour annoter un corpus de langue ancienne, ou toute autre ressource écrite dans un état de langue éloigné d’un standard connu, il est possible d’avoir recours à des méthodes de normalisation (Bollmann et Søgaard 2016). Il s’agit d’ajouter une couche d’annotation aux données pour indiquer à quelle graphie canonique rattacher chaque forme. Le choix de cette référence est à la discrétion de chaque utilisateur. Il peut s’agir d’une forme moderne du mot (Baron et Rayson 2008) ou de la forme ancienne jugée la plus canonique parmi un ensemble de variants constitué automatiquement (Barteld 2017 ; Barteld, Schröder et Zinsmeister 2015).

Cette technique est aussi utilisée par Schauwecker et Stein (2018) pour l’anglo-normand. Les outils d’analyse syntaxique fonctionnent mieux avec l’AF continental, en partie parce que ses graphies présentent moins de variation que celles de l’anglo-normand. Les textes de l’*AND* (Rothwell et Trotter 2005) contiennent des phrases très complexes et des emprunts à d’autres langues, ce qui en fait un corpus difficile à traiter. Ajouter une couche de normalisation avec les variantes standards de l’AF améliore considérablement les résultats. Ce gain est moins important à mesure que les textes analysés sont plus récents, comme Gabay et al. (2020) le constatent pour le français classique et moderne (16e-18e s.). La normalisation peut être utile pour d’autres tâches, comme la détection d’entités nommées en français classique (Kogkitsidou et Gambette 2020).

2.2 Des outils pour le traitement du français médiéval

L’analyse automatique du FMed présente un défi à cause de la forte variation à laquelle les différents états de langues sont soumis. La présence de ressources annotées a cependant permis l’entraînement de modèles et la création d’outils dédiés pour l’analyse morpho-syntaxique, syntaxique et la lemmatisation.

2.2.1 Annotation morphosyntaxique et lemmatisation

Les analyseurs morpho-syntaxiques du FMed servent aussi à la lemmatisation des textes. Le développement de tels modèles a commencé dès la démocratisation des outils statistiques et se poursuit encore, avec des réseaux de neurones.

TreeTagger

TreeTagger (Schmid 1994) est très populaire, car il est simple d'utilisation et performant pour un coût computationnel modeste. C'est le premier outil utilisé pour l'AF. D'après les expériences de Stein et Kunstmann (2003), il est capable d'apprendre des phénomènes même s'ils sont peu fréquents dans l'ensemble d'apprentissage. La lemmatisation repose sur l'utilisation d'un lexique extrait du dictionnaire de Tobler et Lommatzsch et de règles morphologiques pour gérer les formes fléchies et les variants graphiques. Deux modèles d'AF sont disponibles en ligne [2]. *TreeTagger* est intégré à la plate-forme de textométrie *TXM* (Heiden, Magué et Pincemin 2010), ce qui le rend plus facile à utiliser pour des recherches sur corpus.

Une nouvelle version de cet outil utilise des réseaux de neurones à base de modèles à mémoire à court-terme persistante (ou *Bidirectionnal Long Short-Term Memory*, abrégé en *bi-LSTM*) en caractères (*character-based*) pour mieux gérer la variation dans les corpus de langue ancienne (Schmid 2019). Ce nouveau *TreeTagger* atteint plus de 90% d'exactitude sur l'analyse morpho-syntaxique de langues anciennes (moyen néerlandais, ancien anglais, ancien islandais...), dont le MF, pour lequel ce score est de 96,45%.

LGeRM

L'outil *LGeRM* (*Lemmes, Graphies et Règles Morphologiques*, Souvay et Pierrel (2009)) repose aussi sur l'utilisation de règles pour trouver le lemme d'une forme dans un lexique. Celui-ci est plus vaste, car il est composé du *Dictionnaire de moyen français* (R. Martin, Bazin et Cromer 2012), de formes complémentaires au singulier et au pluriel générées automatiquement, d'entrées des dictionnaires de Tobler et Lommatzsch et de Godefroy, qui couvrent l'AF, et de formes verbales fléchies de la nomenclature du *Trésor de la langue française* (*TLF*). Selon Holgado, Lavrentiev et Constant (2021), c'est l'outil qui fournit les meilleurs résultats pour le FMed, face à *TreeTagger* (deuxième place), *UDPipe* et *Pie*.

PIE

Manjavacas, Kádár et Kestemont (2019) ont recours à une architecture neuronale de type encodeur-décodeur qui utilise la prédiction des mots précédents et suivants pour utiliser le contexte de la phrase. Ils traitent indépendamment la lemmatisation et l'analyse morpho-syntaxique, ce qui entraîne des incohérences dans l'annotation des données, mais les résultats sont globalement meilleurs que ceux d'*UDPipe* (Holgado, Lavrentiev et Constant 2021).

Ce modèle est entraîné sur le corpus d'AF de l'*Ecole nationale des Chartes* (*ENC*) et intégré à la plate-forme *Deucalion* (Camps, Clérice et al. 2020), qui propose un service d'annotation morpho-syntaxique et de lemmatisation en ligne.

2. Les modèles *TreeTagger* de l'AF sont disponibles à cette adresse : `http://srcmf.org/`.

UDPipe

UDPipe (Straka 2018) est aussi une infrastructure d'apprentissage supervisé. Elle procède simultanément aux tâches de lemmatisation et d'analyses morpho-syntaxique et syntaxique. L'apprentissage est fait sur des données annotées et des plongements lexicaux pré-entraînés. L'architecture est un réseau de neurones récurrents bi-directionnels. Ce parseur obtient de bons résultats pour de nombreuses langues, mais il est un peu moins bon que les systèmes précédents pour l'annotation morpho-syntaxique du FMed.

PALM

La plate-forme *PALM* dispose d'un modèle d'analyse morpho-syntaxique et de lemmatisation pour le MF, le moyen anglais et le latin médiéval (Mairey et Aouini (2021), cf. section 2.1.3). Les corpus sont annotés avec un nouveau jeu de seize étiquettes morpho-syntaxiques. Les données d'apprentissage de MF étant très différentes de celles des autres systèmes, il n'est pas possible de comparer les performances de ce système avec les autres.

2.2.2 Annotation syntaxique

On ne trouve pas encore de parseurs dédiés au FMed, mais quelques systèmes atteignent une qualité d'annotation qui se rapproche de celle du FC, pour lequel on dispose de beaucoup de données, et qui présente moins de variation. Les dernières études utilisent le format *Universal Dependencies*.

Modèles statistiques

Plusieurs travaux sur le *SRCMF* utilisent le parseur à réécriture de graphes *mate tools* (Bohnet 2010). Les premières expériences (Stein 2014) valident la possibilité de l'analyse de textes de siècles différents par un même modèle et parviennent à un score satisfaisant (cf. section 2.1.3). Les expériences suivantes (Stein 2016) comparent ce parseur avec un parseur à transition (Bohnet et al. 2013). Ce type de parseur est réputé efficace pour les langues à morphologie riche (de Kok 2015), et il présente un gain de performance pour l'AF (LAS 85,96%). Les deux modèles sont disponibles en ligne [3].

Guibon, Tellier, Prévost et al. (2015) utilisent aussi le parseur *mate tools* (Bohnet 2010). Leurs expériences consistent à sélectionner des données d'apprentissage (cf. section 2.1.3) pour déterminer quelles caractéristiques externes influent le plus sur l'analyse syntaxique. Ils concluent à l'intérêt de développer une méthode de sélection de sous-ensembles de textes pour analyser de nouveaux jeux de données.

Modèles neuronaux

Les architectures neuronales sont particulièrement performantes, car elles peuvent tirer parti de données brutes (cf. section 2.1.3). Les différentes compétitions d'analyse syntaxique présentent des scores élevés, comme le parseur de Straka, Straková et Hajič (2019), qui atteint 86,75% de LAS en enrichissant le parseur UDPipe 2.0 (Straka 2018) de plongements lexicaux contextuels. Les développements plus récents ont eu lieu dans le cadre du projet ANR *Profiterole*, et ils sont présentés dans la partie 6.3.1.

3. Les modèles d'AF entraînés sur ces parseurs sont disponibles à cette adresse : https://sites.google.com/site/achimstein/research/resources.

2.3 Approche par *CRF*

2.3.1 Principes des champs aléatoires markoviens (*CRF*)

Les champs aléatoires markoviens (*CRF*, *Conditional Random Fields*, Lafferty, McCallum et Pereira (2001)) sont des modèles probabilistes pour l'annotation séquentielle. On cherche à attribuer une séquence d'étiquettes y à une séquence d'unités x, qui est une phrase dans notre tâche d'annotation morpho-syntaxique. Ces modèles sont dits "discriminants" car "ils modélisent la probabilité conditionnelle de la sortie par rapport à l'entrée, notée $p(y|x)$" (Dupont 2017). Ils appartiennent à la famille des modèles graphiques non dirigés. Nous reprenons la définition des *CRF* de Constant et al. (2011) :

Ils sont définis par X et Y, deux champs aléatoires décrivant respectivement chaque unité de l'observation x et son annotation y, et par un graphe $\mathcal{G} = (V, E)$ dont $V = X \cup Y$ est l'ensemble des nœuds (vertices) et $E \subseteq V \times V$ l'ensemble des arcs (edges). Deux variables sont reliées dans le graphe si elles dépendent l'une de l'autre. [...] chaque étiquette $[Y_i]$ est supposée dépendre de l'étiquette précédente $[Y_{i-1}]$ et de la suivante $[Y_{i+1}]$ et, implicitement, de la donnée x complète.

Les *CRF* sont basés sur ce modèle (Lafferty, McCallum et Pereira 2001) :

$$p_\theta(y|x) = \frac{1}{Z_\theta(x)} \prod_{c \in \mathcal{C}} exp\left(\sum_k \theta_k f_k(y_c, x) \right) \tag{2.1}$$

où $\mathcal{C}$ est l'ensemble des cliques c (c'est-à-dire les sous-graphes connectés) de $\mathcal{G}$ sur Y. La valeur de y sur la clique c est notée y_c. Les fonctions caractéristiques f_k (les traits K) sont définies à l'intérieur de chaque clique. Elles sont fournies par l'utilisateur et prennent la valeur 1 si la configuration choisie est observée, et 0 sinon. Dans notre exemple, f_k prend la forme $f_k(t, y_t, y_{t-1}, x)$. Ces fonctions sont pondérées par les paramètres du modèle, θ_k, qui sont estimés au cours de l'apprentissage. Le facteur de normalisation $Z_\theta(x)$ est ainsi calculé sur l'ensemble des séquences du corpus d'apprentissage :

$$Z_\theta(x) = \sum_y \prod_{t=1}^{T} exp\left(\sum_{k=1}^{K} \theta_k f_k(t, y_t, y_{t-1}, x) \right) \tag{2.2}$$

où T est la taille de la séquence courante, et t, la position courante dans la séquence et celle de y.

Tout comme Guibon, Tellier, Constant et al. (2014), nous utilisons l'implémentation *Wapiti* (Lavergne, Cappé et Yvon 2010), qui permet d'utiliser un format tabulaire pour indiquer les fonctions caractéristiques, ce qui revient à définir des patrons qui sont appliqués à l'ensemble des séquences. Ce format est explicité dans l'exemple 2.3. Les éléments de la séquence ont une ligne chacun, ils sont alignés dans la première colonne. Leur étiquette (y) est dans la dernière colonne. Les colonnes intermédiaires sont des informations ajoutées automatiquement, il s'agit de l'étiquette morpho-syntaxique et du lemme proposés par *TreeTagger*.

L'utilisateur doit décrire des patrons pour entraîner le modèle d'annotation morpho-syntaxique, par exemple, à la position t, il s'agit de consulter le token courant et les prédictions de *TreeTagger* ainsi que le token précédent et son étiquette morpho-syntaxique *TreeTagger*. Ainsi, lorsque t atteint *AUCASIN*, l'observation x contient les informations ci-dessous (en gras) :

C’	NOM	<nolem>	PROdem
EST	VER	ester|estre1_+M|+I	VERcjg
D’	PRE	de_+IS	PRE
AUCASIN	NOM	<unknown>	NOMpro
ET	CON:coord	et_ST	CONcoo
DE	PRE	de_+CIST	PRE
NICOLETE	NPR	nicolete_Z	NOMpro

Tableau 2.3 – Première séquence d’*Aucassin et Nicolette* (fin 12e s. ou début 13e s.)

D’	**PRE**	de_+IS	PRE
AUCASIN	**NOM**	**<unknown>**	NOMpro

Tableau 2.4 – Exemple de patron

2.3.2 Travaux précédents

Guibon, Tellier, Constant et al. (2014) ont fait de premières expériences d’annotation morpho-syntaxique avec des *CRF* sur une version antérieure du *SRCMF*. Outre l’annotation syntaxique du *treebank*, les données comptent une annotation morpho-syntaxique et une lemmatisation par *TreeTagger*. Il est alors possible d’exploiter ces informations avec les patrons déclarés pour l’apprentissage d’un modèle. Comme l’ajout d’un lexique permet d’améliorer la qualité de l’analyse (Constant et al. 2011), un lexique a été extrait de la *BFM*, lequel attribue à toutes les formes de cette base de textes une partie du discours. Cette méthode présentait un gain de qualité par rapport à l’état de l’art.

L’analyse morpho-syntaxique aide aussi à identifier les critères externes qui sont des facteurs de variation. Les écarts de performance entre les différentes configurations des ensembles d’apprentissage et de test permettent en effet de discriminer certains textes ou ensembles de textes qui partagent les mêmes spécificités. Guibon, Tellier, Prévost et al. (2015) concluent notamment que le facteur dialectal est la plus grande source d’hétérogénéité dans le corpus, même si le domaine, l’époque et la forme des textes ont aussi un impact.

2.3.3 Explorations dans le corpus Profiterole

Les premières expériences ont visé à exploiter les travaux précédemment effectués sur le *SRCMF* pour avoir un aperçu de la difficulté de traiter le FMed et mieux connaître les données. Nous avons appliqué la méthode de Guibon, Tellier, Constant et al. (2014) aux textes du corpus *Profiterole* dont l’analyse morpho-syntaxique a été vérifiée. Nos expériences ont confirmé que leurs patrons de *CRF* sont optimisés, nous les utilisons donc à notre tour.

Caractéristiques externes du corpus *Profiterole*

Comme les corpus *SRCMF* et *Profiterole* présentent des caractéristiques externes en partie différentes, les modèles d’apprentissage automatique sont confrontés au défi de l’adaptation au domaine. Les tableaux de données permettent de mettre en avant quelques spécificités :

— Majoritaires dans le *SRCMF* (57,1%), les données en vers deviennent minoritaires (38,1%). L’ajout de textes en MF explique en partie ce changement, car, sur cette période, seulement quatre textes sur vingt-neuf ne sont pas écrits en prose.

— Le nombre moyen de tokens par phrase passe de 12,78 à 14,09. Cette différence peut paraître faible, mais l'écart entre les textes d'AF et de MF est plus grande : la moyenne sur les données d'AF est de 12,17 mots par phrase, contre 16,94 pour le MF. L'usage de phrases très longues (au delà de cent tokens) se généralise à partir du 14e siècle, ce qui implique souvent une augmentation de l'ambiguïté pour le parseur, qui sélectionne toutes les analyses possibles pour chaque token et les combine afin de trouver la plus probable.

— La distribution des domaines est modifiée, limitant un peu l'influence des textes littéraires et religieux. Par exemple, le domaine historique passe d'un seul représentant à dix. Le domaine politique n'est pas présent dans la *BFM*, dont sont extraits la plupart des textes du corpus *Profiterole*.

	religieux	littéraire	didactique	historique	juridique	politique
SRCMF	27,03	57,50	9,34	12,69	0,04	-
Profiterole	16,50	45,37	15,33	16,53	4,03	0,16

Tableau 2.5 – Distribution des domaines dans les données du *SRCMF* et du corpus *Profiterole* (en pourcentage)

— La distribution des dialectes est également différente. De nombreux textes sont écrits dans un dialecte indéterminé, et quatre dialectes font leur entrée : le wallon, le parisien, le bourguignon et l'orléanais.

Notre corpus de référence pour l'annotation morpho-syntaxique contient les textes en bleu dans les tableaux 2.6 et 2.7. Il s'agit des textes du *SRCMF* et de ceux qui viennent de la *BFM* et dont l'annotation morpho-syntaxique a été vérifiée. Lorsque la taille moyenne des phrases dépasse vingt mots, elle est mise en gras. Les textes dont les phrases peuvent dépasser la taille de cent mots sont aussi mis en évidence. Nous ne pouvons donc pas apporter ici de conclusions définitives sur les caractéristiques externes causant les plus grands écarts de performance. Nos expériences nous permettent cependant d'aborder l'hétérogénéité du corpus, et d'en proposer des mesures indicatives.

Expériences

Nous reprenons les paramètres des expériences de Guibon, Tellier, Constant et al. (2014) (cf. section 2.3.2) pour pouvoir comparer les résultats. Nous avons plus de données, notamment des textes de MF annotés, et nous souhaitons savoir si les conclusions faites sur le *SRCMF* se confirment sur les données *Profiterole*, en vue de nos travaux sur l'analyse syntaxique.

Examen des caractéristiques externes Nous avons tout d'abord évalué l'influence des caractéristiques externes des textes sur la qualité de l'annotation morpho-syntaxique, ce qui nous donne une indication sur l'hétérogénéité de notre corpus. Le premier paramètre à tester est celui de l'époque. Nous disposons de textes d'AF annotés en syntaxe de dépendances, mais ce n'est pas le cas pour le MF. Cette expérience donne une idée de l'écart existant entre les deux états de langue.

Les scores du tableau 2.8 sont les pourcentages de tokens correctement annotés par les modèles *CRF*. Pour chaque ensemble de test, nous colorons le meilleur score en gras, et ceux qui s'en approchent (avec une différence en-dessous de 0,5 points). Les étiquettes qui causent le plus d'erreurs dans l'ensemble de test de MF (avec un modèle d'AF) correspondent essentiellement à des homographes. Le pronom impersonnel *il* n'est correctement analysé que dans 11% des cas dans notre expérience. La plupart du temps,

	Siècle	Textes	Domaine	Forme	Dialecte	Nb tokens	Nb phrases	Longueur médiane ph.	Longueur max. ph.
TAF	9	*Serments de Strasbourg*	juridique	prose	indéfini	131	3	53	60
	9	*Eulalie*	religieux	vers	indéfini	212	20	9	26
	11	*Passion de Clermont*	religieux	vers	franco-occ.	3 420	343	9	29
	11	*Saint Léger*	religieux	vers	franco-occ.	1 678	189	7	26
	11	*Saint Alexis*	religieux	vers	normand	5 536	529	8	53
AF	12	*Chanson de Roland*	littéraire	vers	normand	35 314	3 915	8	60
	12	*Comput*	littéraire	vers	anglo-nd	16 711	1 523	9	58
	12	*Descr. Engl.*	historique	vers	anglo-nd	1 509	171	7	23
	12	*Lapidaire*	didactique	prose	anglo-nd	5 516	509	9	58
	12	*Brut*, Wace	historique	vers	anglo-nd	18 037	1952	7	63
	12	*Eneas*	littéraire	vers	normand	40 547	3 714	8	70
	12	*Psautier de Cambridge*	religieux	prose	anglo-nd	5 156	502	9	47
	12	*Becket*	religieux	vers	ouest	22 556	1 989	10	66
	12	*Yvain*	littéraire	vers	champ.	47 995	3881	11	68
	12	*Tristan*, Béroul	littéraire	vers	franco-pic.	32 800	3 321	8	92
	12	*Li Dialoge Gregoire*	didactique	prose	wallon	17 295	1 050	14	76
	13	*Coll. miracles* Adgar	religieux	vers	anglo-nd	24 132	2 112	9	98
	13	*Quatre Livres des Rois*	religieux	prose	anglo-nd	45 131	3 934	8	**108**
	13	*Ami et Amile*	littéraire	vers	indéfini	29 946	2 852	8	65
	13	*Clari*	historique	prose	picard	38 088	2 347	13	**112**
	13	*Aucassin*	littéraire	mixte	picard	11 679	1 032	9	78
	13	*Tyolet*	littéraire	vers	nord-ouest	4 787	430	8	50
	13	*Miracles* de Coinci	religieux	vers	picard	19 794	1 442	11	**160**
	13	*Miracles* de Coinci NCA	religieux	vers	picard	5 364	409	11	57
	13	*Vie saint Eustache*	religieux	prose	indéfini	8 279	638	10	80
	13	*Chansons* Th. Champ.	littéraire	vers	champ.	25 393	1 980	11	50
	13	*Queste Graal*	littéraire	prose	indéfini	44 829	3 117	12	88
	13	*Lettre* J. Sarrasin	historique	prose	parisien	2 619	138	16	77
	13	*Châtelaine de Vergy*	littéraire	vers	normand	6 959	424	13	90
	13	*Récit ménestrel de Reims*	historique	prose	indéfini	22 543	1 868	10	**101**
	13	*Roman de la rose*	didactique	vers	indéfini	22 518	1 481	12	**189**
	13	*Bathilde 1*	religieux	prose	picard	11 175	571	17	72
	13	*Coutumes Beauvaisis*	juridique	prose	picard	22 637	947	21	**119**

TABLEAU 2.6 – Corpus *Profiterole* (1/2)

	Siècle	Textes	Domaine	Forme	Dialecte	Nb tokens	Nb phrases	Longueur médiane ph.	Longueur max. ph.
MF	14	*Roman de Fauvel*	didactique	vers	normand	21 775	1575	12	**101**
	14	*Fouke le Fitz Waryn*	littéraire	mixte	anglo-nd	29427	2372	10	85
	14	*Passion* (anonyme)	religieux	vers	bourg.	14059	1345	8	72
	14	*Chronique de Morée*	historique	prose	indéfini	21759	1 124	15	**153**
	14	*Gdes chron. de France* 9	historique	prose	indéfini	15377	736	17	**116**
	14	*Fortune*	littéraire	vers	champ.	29 096	1 748	13	**123**
	14	*Livre seyntz medicines*	religieux	prose	anglo-nd	21 861	932	18	**175**
	14	*De la erudition* J. Daudin	didactique	prose	parisien	22 794	1 328	14	**150**
	14	*Geneviève 2*	religieux	prose	indéfini	13 838	962	11	64
	14	*Chroniques Jean II et Ch. V*	historique	prose	indéfini	14 281	528	19	**162**
	14	*L'art de dictier*	didactique	mixte	champ.	6 421	291	15	**154**
	14	*Mélusine*	littéraire	prose	indéfini	26 661	1 787	11	**170**
	14	*Manières lang.* 1396	didactique	mixte	anglo-nd	18 188	1 428	10	**143**
	14	*Récit...* O. d'Anglure	littéraire	prose	champ.	27 923	1487	16	95
	14	*Manières lang.* 1399	didactique	mixte	anglo-nd	6 404	554	8	**228**
	15	*Quinze Joyes de mariage*	littéraire	prose	ouest	39 448	2 679	12	**131**
	15	*Cité Dames*	littéraire	prose	indéfini	22 838	1 253	14	**106**
	15	*Journal* 1 N. de Baye	juridique	prose	indéfini	24 111	767	22	**230**
	15	*Ballades* Ch. d'Orléans	littéraire	vers	orléanais	26 136	1 869	12	64
	15	*Manières lang.* 1415	didactique	mixte	anglo-nd	3 818	256	9	**108**
	15	*Bathilde* 3	religieux	prose	picard	1 246	88	11	49
	15	*Quadrilogue*	didactique	prose	indéfini	21 041	826	21	**112**
	15	J. Juvénal des Ursins	politique	prose	champ.	1 812	105	14	61
	15	*Chronique* Monstrelet	historique	prose	indéfini	32 757	1411	18	**174**
	15	*Bathilde* 2	religieux	prose	indéfini	12 517	638	16	90
	15	*Le Jouvencel* 1	didactique	prose	indéfini	22 756	1 423	15	98
	15	*Evangiles Quenouilles* 1	didactique	prose	picard	9 727	539	16	59
	15	*Roman J. de Paris*	littéraire	prose	indéfini	28 967	1 836	14	74
	15	*Mémoires* Ph. Commynes	historique	prose	ouest	25 251	1 300	15	**138**

TABLEAU 2.7 – Corpus *Profiterole* (2/2)

		Test	
		AF	MF
Train	AF	**95,90**	94,73
	MF	91,81	**96,89**

Tableau 2.8 – Expériences sur les époques (*Cattex*)

il est confondu avec le pronom personnel. Les erreurs sur l'étiquette *ADVsub* (adverbes subordonnants) sont causées par *comment*, confondu avec l'adverbe interrogatif, et *comme*, confondu avec une conjonction de subordination. Ces erreurs ne sont pas dues à des changements linguistiques entre AF et MF, mais à l'ambiguïté des formes.

En revanche, l'évolution linguistique est à l'origine de nombreuses erreurs d'analyse. Celles-ci restent néanmoins minoritaires face aux étiquettes correctes des catégories concernées. Par exemple, certaines prépositions peu fréquentes en AF sont analysées comme des adverbes, comme *après* (et sa variante graphique *aprés*) et *jusque* (aussi *jusques*), dont l'usage se généralise à partir du 13e siècle.

Le score d'étiquetage du MF avec un modèle de la même époque semble indiquer que ces données sont plus homogènes que celles d'AF. La taille inférieure de l'échantillon en est a priori la raison. Sur les cinq textes de MF utilisés pour ces expériences, trois sont *Manières de langage*, un ouvrage anglo-normand. Les scores sont meilleurs lorsque les ensembles d'apprentissage et de test ont les mêmes caractéristiques, mais l'écart des performances laisse penser qu'il n'est pas impossible de faire un système commun pour l'ensemble du FMed.

La plupart des erreurs d'annotation du modèle d'AF sur le MF viennent de la polysémie de certains tokens, notamment la conjonction de subordination *que* annotée comme pronom relatif ou interrogatif, ou de mots inconnus, en particulier des adverbes (*ADVgen*) et des noms communs. Ils sont alors étiquetés avec une autre catégorie d'une classe "ouverte" du lexique (adjectif, adverbe, nom, verbe). Les catégories qui ont un taux d'erreur supérieur à un tiers sont celles qui sont faiblement représentées (ex. interjections, contractions) et celles qui présentent des cas d'ambiguïté, comme *ADJcar* et *DETcar*, qui sont difficiles à différencier pour un analyseur morpho-syntaxique.

		Test		
		12e	13e	MF
Train	12e	**95,31**	94,54	94,20
	13e	94,38	**96,24**	94,54
	MF	90,78	92,80	**96,89**

Tableau 2.9 – Expérience sur les époques, avec comparaison des 12e et 13e siècles (*Cattex*)

L'AF couvre une période plus étendue, et a accueilli de nombreux changements. L'expérience 2.9 permet de comparer les scores obtenus sur les 12e et 13e siècles, qui regroupent le plus de textes annotés à notre disposition, et sur le MF.

A mesure que les époques des ensembles s'éloignent l'une de l'autre, les résultats perdent en qualité. Il semble que les textes les plus récents sont de mauvais candidats pour annoter les textes plus anciens, mais que l'écart est moins important dans le cas inverse, ce qui peut aussi être dû à la différence de volume de données.

Les différences de qualité d'annotation sont moins importantes pour le critère de forme des textes (vers ou prose, cf. tableau 2.10), contrairement à ce qui était attendu. Guibon, Tellier, Prévost et al. (2015)

faisaient état d'un écart notable dans les résultats de l'analyse morpho-syntaxique, mais pas dans ceux de l'analyse syntaxique. Il est possible qu'en utilisant un volume de données plus grand, l'ensemble soit rendu plus homogène, et que la part de mots inconnus ait diminué. Cette expérience permet aussi de mêler des textes d'époques et de dialectes différents, qui sont des critères plus discriminatoires que la forme, ce qui explique aussi ces nouveaux scores.

		Test	
		vers	prose
Train	vers	**95,54**	95,25
	prose	93,45	**96,65**

Tableau 2.10 – Expérience sur les sous-corpus de vers et de prose (*Cattex*)

Les écarts de qualité d'annotation sont, en revanche, plus visibles pour les expériences sur les dialectes (cf. tableau 2.11) et les domaines (cf. tableau 2.12), même s'ils sont inférieurs à ceux des premières expériences sur le *SRCMF*. Ne disposant que d'un texte en champenois, nous l'avons remplacé par l'anglo-normand.

		Test		
		picard	normand	anglo-nd
Train	picard	**95,59**	93,90	92,58
	normand	93,26	**95,71**	93,99
	anglo-nd	93,06	93,78	**95,98**

Tableau 2.11 – Expérience sur les dialectes (*Cattex*)

Les premières expériences ne disposaient que d'un texte historique. L'apport de nouvelles ressources, plus précisément des *Mémoires* de Ph. de Commynes (cf. tableau 2.7, dernière ligne), a permis d'améliorer considérablement les résultats du modèle entraîné sur ce domaine.

		Test				
		littérature	religion	didactique	historique	juridique
Train	littérature	**95,81**	93,74	92,20	95,72	92,75
	religion	94,19	**95,3**	92,28	95,56	93,38
	didactique	93,30	92,76	**94,52**	94,58	93,42
	historique	92,19	91,90	90,84	**96,76**	92,31
	juridique	91,16	90,98	90,33	92,84	**95,46**

Tableau 2.12 – Expérience sur les domaines (*Cattex*)

Influence du jeu d'étiquettes Malgré un degré de granularité moindre, les scores d'annotation avec les étiquettes *Universal Dependencies* (version 2.6) sont inférieurs à ceux obtenus avec le jeu *Cattex*. On le voit notamment sur l'expérience suivante (cf. tableau 2.13), où l'écart moyen est de deux points.

Cependant, on retrouve les tendances observées dans les expériences effectuées avec les annotations au format *Cattex*, par exemple avec les domaines (cf. tableau 2.14), où on trouve globalement les meilleurs scores lorsque l'ensemble d'apprentissage possède les mêmes caractéristiques que l'ensemble de test.

Les étiquettes les moins bien prédites sont celles dont les formes sont les plus soumises à la variation, c'est-à-dire les auxiliaires (*AUX*), les numéraux (*NUM*) et les contractions rares comme *ADV.PRON* (ex.

		Test		
		picard	normand	anglo-nd
Train	picard	**93,32**	90,80	92,66
	normand	89,71	**94,41**	92,42
	anglo-nd	89,27	91,62	**95,36**

Tableau 2.13 – Expérience sur les dialectes (*UD*)

		Test				
		littérature	religion	didactique	historique	juridique
Train	littérature	**93,51**	91,84	91,31	94,13	88,73
	religion	91,12	**95,2**	93,28	95,58	94,07
	didactique	90,36	92,29	**95,01**	94,88	93,71
	historique	92,77	93,07	92,04	**97,23**	93,56
	juridique	89,1	92,52	92,06	94,11	**96,35**

Tableau 2.14 – Expériences sur les domaines (*UD*)

nel, nen, sil, sis). Les formes annotées *AUX* ne couvrent pas uniquement les auxiliaires *estre* et *avoir*, mais également les modaux, ce qui les rend d'autant plus difficiles à distinguer des formes de l'étiquette *VERB*. Les interjections (*INTJ*) ont une exactitude irrégulière, selon les tokens présents dans l'ensemble d'apprentissage.

L'annotation de nouveaux textes de FMed se heurte non seulement au défi de l'adaptation au domaine, mais aussi au changement linguistique constant. L'évolution linguistique et la grande diversité des textes tendent à faire de chaque corpus de textes anciens une ressource hétérogène. Cela a motivé le développement de modèles et d'outils dédiés aux états de langue anciens, notamment pour le FMed. Nous gardons la distinction entre AF et MF, notamment à cause de l'évolution lexicale. D'après nos expériences préliminaires, il est envisageable de développer un outil commun pour annoter les textes du corpus *Profiterole*. Nous espérons ainsi produire une annotation cohérente entre états de langue, malgré l'hétérogénéité des données.

3 Adaptation d'une métagrammaire existante

Les travaux en traitement automatique du langage sont désormais majoritairement menés avec des méthodes statistiques. Cependant, les systèmes symboliques restent pertinents, notamment parce qu'ils permettent d'obtenir une annotation cohérente et de qualité (Brants et al. 2002).

L'analyse syntaxique du français médiéval, comme celle des autres langues anciennes, est rendue difficile par la forte variation linguistique. Il semble donc intéressant d'utiliser directement les connaissances grammaticales pour contraindre l'analyse des phrases, plutôt que de se fonder sur des corpus arborés, dont le volume est limité (du fait de la faible quantité de données disponibles au regard de celles qui le sont pour des états de langue contemporains), pour entraîner des systèmes statistiques. D'une part, un système symbolique garantit une annotation de qualité pour les phénomènes décrits. La couverture syntaxique de la grammaire grandit au fil des développements. D'autre part, il est possible de corriger directement les analyses erronées, grâce à la fouille d'erreurs en sortie de parseur.

Il existe plusieurs formalismes syntaxiques assez performants pour produire une grammaire électronique du français médiéval. Nous choisissons d'adapter la métagrammaire *FRMG* pour le français contemporain (Villemonte de la Clergerie 2005). Nous présentons donc dans cette partie les concepts sur lesquels elle s'appuie : les grammaires formelles, en particulier le formalisme grammatical *LTAG*, celui des métagrammaires, puis celui de l'instanciation *SMG* (*Simple Metagrammar*), propre à *FRMG*. Enfin, nous présentons notre démarche pour adapter cette métagrammaire.

3.1 L'approche (méta)grammaticale

3.1.1 Grammaires formelles

Nous désignons désormais par "grammaires" les grammaires formelles, et non les grammaires traditionnelles comme celle de Buridant (2000) pour l'ancien français. Elles forment un automate, ce qui rend leur implémentation possible. Ces systèmes reconnaissent un "langage" particulier, i.e. un ensemble infini de phrases à partir d'une description de taille finie, mais ils rejettent celles qui n'en font pas partie. Si la phrase est acceptée, le parseur en propose une analyse. Le système syntaxique (la grammaire formelle) est décrit et évalué par les développeurs, et il doit ressembler autant que possible à la langue naturelle (Chomsky 1959).

Principes de base

Les grammaires sont des descriptions de structures bien formées d'un langage donné (Fort 2016 ; Kallmeyer et al. 2017). Une grammaire électronique doit respecter quelques critères décrits par A. Abeillé (2002). D'une part, elle doit couvrir les phénomènes majeurs de la langue qu'elle traite, reconnaître les analyses possibles d'une phrase et rejeter celles qui ne sont pas correctes. D'autre part, elle doit être

efficiente sur la langue et être réutilisable sur de nouveaux corpus. Le langage utilisé pour décrire les contraintes doit posséder une syntaxe claire, et les contraintes décrites doivent être explicites. Une telle grammaire doit pouvoir utiliser une ressource lexicale suffisamment couvrante pour être utilisée sur un corpus. Dans un deuxième temps, ce système est amené à être paramétrable, réutilisable pour de nouvelles tâches et interfacé avec des modules phonétiques et sémantiques.

Types de grammaires

Les grammaires électroniques sont habituellement catégorisées en suivant cette hiérarchie (Chomsky 1959) :

— Type 3 : les grammaires régulières, qui sont reconnaissables par des automates à états finis

— Type 2 : les grammaires hors-contexte (ou non contextuelles), qui sont reconnaissables par des automates à pile non déterministes

— Type intermédiaire : les grammaires faiblement contextuelles (ou faiblement dépendantes du contexte). Quelques formalismes se situent entre les types 1 et 2 parce qu'ils permettent d'introduire un contexte limité, mais l'analyse reste d'une complexité acceptable, c'est-à-dire polynomiale en la longueur de la phrase à analyser.
ex. grammaires d'arbres adjoints (*TAG*), grammaires catégorielles combinatoires (*CCG*), grammaire minimaliste

— Type 1 : les grammaires contextuelles, qui sont l'équivalent d'automates linéairement bornés
ex. grammaires lexicales-fonctionnelles (*LFG*)

— Type 0 : les grammaires générales, qui sont l'équivalent des machines de Turing
ex. grammaires syntagmatiques guidées par les têtes (*HPSG*), grammaires transformationnelles

3.1.2 Les grammaires d'arbres adjoints lexicalisées (*LTAG*)

Principe

Les grammaires d'arbres adjoints (ou *Tree Adjoining Grammars*, désormais *TAG*, Joshi, Levy et Takahashi (1975)) sont un formalisme mathématique adapté aux langages naturels. Joshi et Schabes (1997) en donnent cette description formelle avec le quintuplet <N, T, I, A, S> où :

— N est un ensemble de non-terminaux correspondant à des catégories syntagmatiques,

— T est un ensemble de terminaux (aussi appelés "ancres"),

— I est un ensemble d'arbres initiaux

— A est un ensemble d'arbres auxiliaires, qui ont chacun un nœud feuille appelé "nœud pied", symbolisé par une étoile (*) et étiqueté par un non-terminal de même catégorie que le nœud-racine de cet arbre

— S ∈ N est l'axiome de la grammaire.

Les arbres élémentaires d'une *TAG* sont l'ensemble des arbres initiaux et auxiliaires. Ils se combinent au moyen des deux opérations ci-dessous, à savoir la substitution et l'adjonction. Un arbre dérivé par ces opérations est dit "complet" lorsqu'il ne reste plus de nœuds feuilles étiquetés par des non-terminaux. L'ordre des nœuds feuilles reflète celui des mots de la phrase. L'ensemble des opérations appliquées est la

dérivation. Elle est représentée dans un "arbre de dérivation" qui enregistre l'historique de l'assemblage des arbres élémentaires (Kallmeyer 2010) : les arêtes de l'arbre représentent les opérations.

1. L'opération de substitution consiste à insérer un arbre initial ou dérivé sur un nœud-feuille dans un arbre élémentaire ou dérivé (marqué par le symbole N↓). C'est une opération non contextuelle, similaire à la substitution dans les grammaires hors-contexte.

   (1) Sainz Innocenz ert idunc apostolie
   ADJqua NOMpro VERcjg ADVgen NOMcom
   'Saint Innocenz était donc un apôtre.'
   Vie de Saint Alexis, v. 301 (ca. 1050)

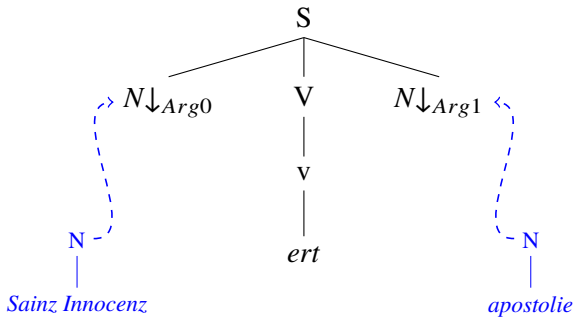


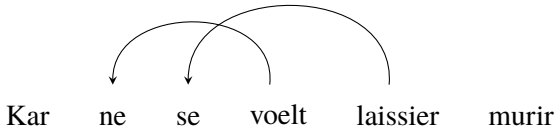
Figure 3.1 – Exemple de substitution du sujet (Arg0) et de l'attribut du sujet (Arg1)

   Dans cet exemple, les nœuds du sujet et de l'attribut du sujet sont représentés dans l'arbre élémentaire du verbe d'état *estre* ("être") et sont amenés à être substitués par leurs réalisations syntaxiques.

2. L'opération d'adjonction (cf. figure 3.2) consiste à insérer un arbre auxiliaire au niveau d'un nœud interne de même catégorie, c'est-à-dire qu'on ouvre un nœud d'un arbre de catégorie *N* pour y insérer un arbre auxiliaire dont la racine est aussi de catégorie *N*. La partie sous le premier nœud est raccrochée au niveau du nœud pied de l'arbre auxiliaire, également de catégorie *N*. Cette opération est facultative et peut être répétée. Elle peut être rendue obligatoire ou interdite, mais elle ne peut avoir lieu sur un nœud de substitution ou un nœud pied. Elle est faiblement contextuelle, rendant ainsi la grammaire faiblement contextuelle. En effet, l'ordre linéaire des feuilles des arbres *TAG* est conservé lors de l'analyse.

Les *TAG* appartiennent à la classe des grammaires faiblement contextuelles (Joshi, Vijay-Shanker et Weir 1990), ce qui leur permet de générer le langage $L = \{ww/w \in \{a,b\}*\}$ (Joshi 1985), qui n'est pas contextuel. Les dépendances sérielles croisées (Shieber 1985) peuvent ainsi être traitées, comme dans l'exemple suivant, avec le langage $L = \{ww/w \in \{a,b\}*\}$, qui est ici un équivalent du premier :

(2) Kar ne se voelt laissier murir
CONcoo ADVneg PROper VERcjg VERinf VERinf
'Car [il] ne veut pas se laisser mourir'
Lais de Marie de France, v. 128 (ca. 1160)



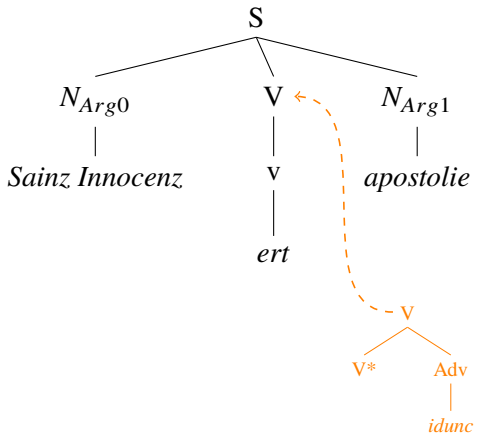


FIGURE 3.2 – Exemple d'adjonction d'un adverbe

Cet exemple renferme le langage $L = (ne)^m(se)^n(voelt)^m(laissier)^n$, soit un exemple de la forme $L' = a^m b^n c^m d^n$. Il permet d'illustrer la possibilité du croisement, mais pas celle de la récursion. Analyser un tel langage ne serait pas possible pour une grammaire générant un langage contextuel. En revanche, *TAG* ne peut pas générer le langage de triple copie $L = \{www/w \in \{a,b\}*\}$ (KALLMEYER (2010), p. 63).

Les *TAG* sont donc plus puissantes que des grammaires hors contexte, mais elles restent exploitables pour l'analyse syntaxique automatique de corpus. La complexité de l'analyse dépend de la longueur n de la phrase, et les *TAG* ont une complexité en temps qui est polynomiale, en l'occurrence $O(n^6)$, dans le pire des cas. La complexité en espace est de $O(n^4)$. Les *TAG* ont cette particularité grâce à "la factorisation de la récursivité et au domaine de localité [étendu]" (JOSHI 1985). Ce type de grammaire est utilisé pour traiter de nombreuses langues, en particulier le français (A. ABEILLÉ 1998 ; CRABBÉ 2005 ; VILLEMONTE DE LA CLERGERIE 2005), l'anglais (XTAG RESEARCH GROUP 2001), l'allemand (FRANK 2001), le coréen (HAN et al. 2000) et l'arabe (BEN FRAJ 2011).

Certaines grammaires *TAG* sont des équivalents de grammaires hors-contexte, ce qui limite leur coût. Pour cela, on y interdit l'opération d'adjonction, obtenant ainsi des grammaires à base de substitution (*Tree Substitution Grammars, TSG*, JOSHI (2004)). Cependant, le développement d'une telle grammaire est plus complexe d'un point de vue linguistique, car il est plus facile de décrire les modifieurs avec des arbres auxiliaires.

Les grammaires d'insertion (*Tree Insertion Grammar, TIG*, SCHABES et WATERS (1995)) sont un type de *TAG* qui autorisent l'opération d'adjonction, mais qui la restreignent à un usage particulier pour limiter le coût de l'analyse. En effet, les *TIG* n'autorisent pas les arbres auxiliaires englobants, mais seulement ceux qui insèrent du matériel en frontière gauche ou droite. Cela fait des *TIG* un équivalent de grammaire hors-contexte. Ce type d'arbre permet de couvrir de nombreux phénomènes. Une grammaire *TAG* peut donc être totalement ou partiellement *TIG*. C'est cependant cette opération d'adjonction englobante qui fait la force du formalisme *TAG*, car cela permet l'adjonction de structures complexes.

Structures de traits

Enrichir une *TAG* de structures de traits sur les nœuds permet d'introduire des contraintes sur l'unification des nœuds (VIJAY-SHANKER 1987). L'accord en personne et en nombre peut, par exemple, être exigé entre le verbe et son sujet. En ouvrant le nœud sur lequel a lieu l'adjonction, on crée deux structures de

traits : une en amont (*top*) et l'autre en aval (*bottom*). Elles doivent s'unifier au terme de la dérivation, aboutissant ainsi à une nouvelle structure. Sur les autres nœuds (de substitution et nœuds pieds), elles sont unifiées par défaut. Dans une grammaire lexicalisée (voir section 3.1.2), ces structures peuvent être intégrées aux entrées lexicales pour contraindre l'analyse syntaxique des arbres associés à ces entrées (Vijay-Shanker et Schabes 1992).

Version lexicalisée (LTAG)

Pour produire des analyses linguistiquement valides, quatre principes sont appliqués à ces grammaires (A. Abeillé 1993 ; Kroch et Joshi 1985) :

1. principe d'ancrage lexical (ou lexicalisation) : un arbre élémentaire doit avoir au moins une ancre lexicale non vide (ex. arbre 3.1, ancré par *ert*),
2. principe de cooccurrence prédicat-arguments : chaque prédicat intègre ses arguments à sa structure élémentaire (ex. *ert* requiert deux arguments, Arg0 et Arg1),
3. principe de consistance sémantique : un arbre élémentaire ne peut être sémantiquement "vide", les éléments fonctionnels n'étant pas perçus comme autonomes,
4. principe de non-compositionnalité : un arbre élémentaire correspond à une seule unité sémantique.

Le principe d'ancrage lexical introduit la version lexicalisée des *TAG*, les *LTAG* (Schabes, A. Abeillé et Joshi 1988). Ces grammaires sont composées d'un lexique morphologique et syntaxique et d'un ensemble de schémas d'arbres, i. e. des arbres dont il manque l'ancre lexicale (Candito 1996). Ces arbres sont lexicalisés lors de l'analyse syntaxique, lors de laquelle un lexique attribue un ensemble défini de structures à chaque item lexical. On considère même qu'un arbre élémentaire est la projection de cette ancre, comme l'arbre 3.1, qui serait une projection de l'entrée *ert*. Une phrase sélectionne donc un ensemble fini de structures, qui peuvent se combiner un nombre fini de fois, limitant efficacement l'ambiguïté de l'analyse (Schabes et Joshi 1991).

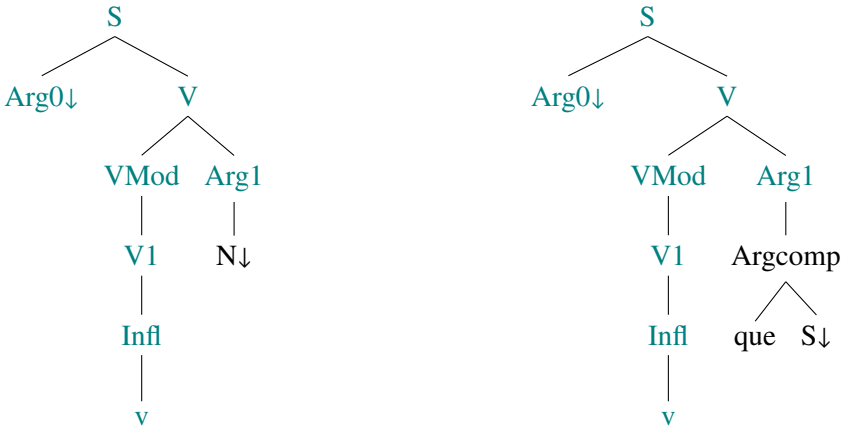
Cet ancrage lexical est possible dans une grammaire d'arbres adjoints, car son domaine de localité étendu permet une description locale des phénomènes syntaxiques, comme l'attachement des arguments à leur prédicat (deuxième principe). Cela permet de simplifier la description linguistique (Kroch et Joshi 1985) et justifie la formation des arbres (Joshi et Schabes 1991). L'utilisation d'un lexique est bénéfique pour l'analyse syntaxique, car il permet de limiter le nombre d'arbres considérés pour une phrase en fonction des mots présents. Par exemple, une phrase qui ne contient pas d'adverbe n'inclut pas les arbres de modifieurs adverbiaux dans les recherches d'analyse, excluant ainsi des erreurs d'analyse flagrantes, ce qui améliore aussi l'efficacité de l'analyse.

Kallmeyer (2010) remarque cependant que seuls les deux premiers principes sont généralement observés, le dernier demandant une prise de position claire sur ce qu'est une unité sémantique. Il est possible de trouver dans une grammaire des arbres dont l'ancre ou la co-ancre est un élément fonctionnel, excluant ainsi en partie la contrainte de consistance sémantique. La grammaire *FRMG* (cf. partie 3.2) comporte ce type de description, ainsi que des arbres non ancrés, notamment pour la description de la ponctuation et l'accroche des propositions relatives. En revanche, le principe de non-compositionnalité y est respecté autant que possible. Les quatre principes sont donc des directions générales pour les grammaires, mais ils peuvent être nuancés selon les besoins.

D'un point de vue computationnel, ces principes, en ajoutant des contraintes de description, limitent le coût de l'analyse. De plus, le domaine de localité étendu des *TAG* permet d'éviter la propagation de structures de traits pour faire les unifications nécessaires aux substitutions, car elles ont lieu à l'intérieur de l'arbre (Carroll, Nicolov et al. 1999). Cela permet de limiter l'utilisation de la mémoire, rendant l'analyse plus efficace.

3.1.3 Métagrammaires

Une grammaire *TAG* à large couverture est nécessaire pour traiter une langue naturelle, ce qui implique la présence d'un grand nombre d'arbres. Par exemple, dans la grammaire *LTAG* de l'anglais de Abeille et al. (1990), huit familles d'arbres sont nécessaires pour traiter les verbes ayant des arguments nominaux et pronominaux, et vingt-et-une familles d'arbres pour les verbes ayant des arguments phrastiques. Chaque famille contient entre trois et douze arbres. Pour traiter tous les verbes, il faut notamment ajouter les familles d'arbres qui décrivent les semi-auxiliaires, les verbes à particules et les idiomes. Cependant, un grand nombre d'arbres fait de la grammaire un objet complexe (Gómez-Rodríguez, Alonso et Vilares 2006) et difficile à maintenir. Ceux-ci peuvent être factorisés pour rassembler l'information commune et éviter de faire les mêmes descriptions dans de multiples arbres.



(a) Verbe transitif avec un objet nominal

(b) Verbe transitif avec un objet phrastique

Figure 3.3 – Points communs entre les arbres élémentaires ancrés par des verbes transitifs (en vert) dans *FRMG*

La figure 3.3 illustre la descriptions d'arbres de la même famille dans *FRMG*, en l'occurrence celle des verbes transitifs. Les deux arbres possèdent la même épine dorsale (de S à v), mais des réalisations différentes de l'argument objet Arg1. Le syntagme V prévoit plusieurs nœuds d'accroche pour les éléments du syntagme verbal, notamment les auxiliaires de temps et de passivation (Infl) et les pronoms conjoints (V1). Il est intéressant de partager ces descriptions aux arbres ancrés par des verbes transitifs. C'est un formalisme plus abstrait qui permet de générer ces arbres factorisés à partir d'une description modulaire de la syntaxe : les métagrammaires.

Définition

Une métagrammaire (Candito 1999) fournit une description modulaire et hiérarchique d'une langue, organisée en classes. Celles-ci héritent des contraintes des classes parentes et elles peuvent ajouter des

contraintes supplémentaires. Les informations contenues dans une métagrammaire sont censées être organisées en trois domaines différents, appelés les "dimensions" (Candito 1999) :

1. Sous-catégorisation initiale des arbres élémentaires : les fonctions grammaticales des composants de l'arbre doivent être renseignées.
   ex. Dans la figure 3.3, on précise que Arg0 est le sujet de l'ancre (v), et Arg1 son objet.
2. Redistributions possibles : il s'agit des changements de diathèse.
   ex. Si un verbe transitif peut apparaître à la voix passive, il est analysé avec un arbre élémentaire pour ce phénomène.
3. Réalisations des fonctions grammaticales : il s'agit des formes que peuvent prendre les éléments mentionnés dans la première dimension.
   ex. Dans l'arbre 3.3a, l'objet du verbe (Arg1) est réalisé par un syntagme nominal, et celui de l'arbre 3.3b, par une complétive introduite par "que".

Dans une métagrammaire comme *FRMG*, les frontières entre ces trois dimensions sont plus floues. Les métagrammaires peuvent générer différents types de grammaires (Clément et Kinyon 2003), mais nous ne nous intéresserons ici qu'aux grammaires d'arbres adjoints (TAG).

Les métagrammaires générant des TAG

Avec une métagrammaire générant une *LTAG*, comme *FRMG*, la description syntaxique d'un énoncé dépend du lexème qui le compose. Un lexique donne accès à des informations, comme la valence et la catégorie grammaticale, organisées en structures de traits appelées *hypertags* (Kinyon 2000).

La sous-catégorisation initiale d'un prédicat est représentée dans l'arbre élémentaire qu'il ancre. Cette première dimension implémente le principe de cooccurrence prédicat-arguments des TAG. Par exemple, il existe, dans le lexique, deux entrées pour le verbe *estre*, car elles n'ont pas le même cadre de valence : la première entrée est le verbe d'état et l'autre est l'auxiliaire. L'exemple de la figure 3.1 peut être analysé avec l'arbre élémentaire de la première entrée, qui attend deux arguments : un sujet et un attribut du sujet.

La deuxième dimension apporte l'information de la sous-catégorisation des redistributions. Dans le cas de la voix passive, Arg1 devient généralement le sujet, et Arg0 devient un complément optionnel introduit par la préposition *par*. La plupart des verbes transitifs offrent cette possibilité, comme le verbe *trover* ("trouver") dans l'exemple ci-dessous.

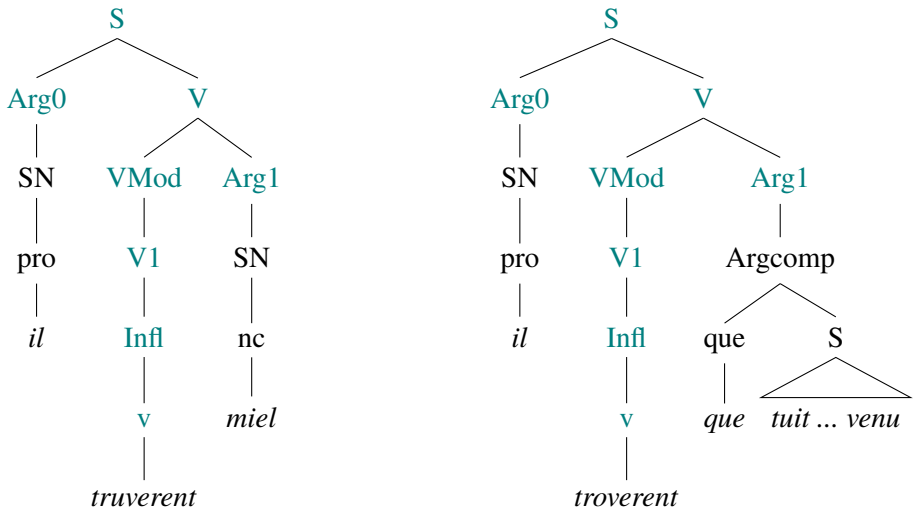
(3) é dit m' ad qu' il sunt truve
CONcoo VERppe PROper VERcjg CONsub PROper VERcjg VERppe
'et [il] m'a dit qu'ils ont été trouvés'
Quatre Livres des Rois, p. 19 (ca. 1190)

Enfin, les fonctions syntaxiques peuvent être réalisées de différentes manières, selon l'hypothèse distributionnelle, ce qui constitue la dernière dimension contrôlant la construction des arbres. Ces différents phénomènes syntaxiques sont décrits individuellement dans les classes de la métagrammaire. La fonction d'objet du verbe *trover* peut ainsi être réalisée par un certain type de constituant : un syntagme nominal ou une proposition complétive.

(4) ex. objet nominal

Vindrent li plusur en une lande ú il truverent miel
VERcjg DETdef PROind PRE DETind NOMcom PROrel PROper VERcjg NOMcom
'Ils vinrent à plusieurs dans une lande où ils trouvèrent du miel'
Quatre Livres des Rois, p. 26 (ca. 1190)

(5) ex. objet phrastique
il troverent que tuit li compaignon de la table reonde
PROper VERcjg CONsub PROind DETdef NOMcom PRE DETdef NOMcom ADJqua
furent venu
VERcjg VERppe
'Ils trouvèrent que tous les compagnons étaient venus'
Queste del saint Graal, p. 161b (ca. 1225 ou 1230)



(a) *trover* + objet nominal (b) *trover* + complétive

FIGURE 3.4 – Analyse simplifiée de deux phrases

Ces deux réalisations sont décrites dans des classes différentes, l'une traitant les propositions subordonnées complétives, et l'autre les syntagmes nominaux. Les verbes ayant le même cadre de valence que le verbe *trover* ancrent plusieurs arbres élémentaires, l'un appelant une complétive pour remplir Arg1, l'autre appelant un objet nominal.

Une métagrammaire permet de factoriser l'information, et ainsi de limiter la redondance dans les descriptions. Les principes communs à plusieurs éléments grammaticaux peuvent ainsi être rassemblés, puis hérités par les classes filles. Par exemple, les verbes modaux, les auxiliaires et les autres verbes du lexique partagent des caractéristiques comme l'accord avec le sujet et une partie de leur structure d'arbre. Néanmoins, ils n'ont pas la même sous-catégorisation et n'ancreront donc pas les mêmes arbres. Modulariser ainsi l'information facilite la compréhension et la maintenance de la métagrammaire.

3.2 FRMG

French Metagrammar [1] (*FRMG*) est avant tout une métagrammaire du français contemporain, développée par VILLEMONTE DE LA CLERGERIE (2005). Elle rassemble 451 classes qui produisent 381 arbres,

1. La documentation de *FRMG* est disponible en ligne à cette adresse : http://alpage.inria.fr/frmgwiki/.

garantissant à la grammaire générée une large couverture. Elle est utilisée dans la chaîne de traitement installée par *alpi* (*Alpage Installer*) pour faire de l'analyse syntaxique sur corpus.

3.2.1 Formalisme

Description de la métagrammaire

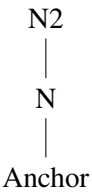
FRMG est à la fois le nom d'une métagrammaire, d'une grammaire TAG générée à partir de cette métagrammaire, et d'un parseur compilé à partir de cette grammaire (Villemonte de la Clergerie 2013). La métagrammaire s'appuie sur le formalisme *Simple Metagrammar* (*SMG*, Thomasset et Villemonte de la Clergerie (2005)), qui offre de nombreuses possibilités d'expression pour décrire une langue.

D'une part, la description d'un phénomène peut être répartie sur plusieurs classes, grâce à un système d'héritage. Il est aussi possible de mettre en place des ressources, que ces classes créent ou consomment. Par exemple, la ressource de l'accord verbal est requise dans les classes décrivant le syntagme verbal canonique, les auxiliaires et les verbes modaux.

D'autre part, il est possible d'appliquer des contraintes sur les nœuds des arbres, en les exprimant par les décorations des nœuds ou par la classe. Pour cela, on dispose des opérateurs de base d'une grammaire, qui sont l'égalité, la dominance (directe et indirecte) ou la précédence entre nœuds. Dans la classe générique pour les noms et les pronoms (*noun*), on trouve des contraintes de dominance (cf. figure 3.5).

```
N2 >> N;
N >> Anchor;
```

(a) Description de contraintes dans *FRMG*



(b) Arbre partiel correspondant

Figure 3.5 – Exemple de contrainte de dominance entre nœuds dans *FRMG*

Les contraintes de dominance peuvent être indirectes, ce qui permet de décrire des nœuds intermédiaires dans d'autres classes. Dans une classe fille de *noun* qui vise à décrire les noms communs, on trouve une contrainte d'égalité (cf. figure 3.6).

```
N2 >> Nc;
Nc = Anchor;
```

(a) Description de contraintes dans *FRMG*

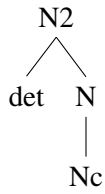
(b) Arbre partiel correspondant

Figure 3.6 – Exemple de contrainte d'égalité entre nœuds dans *FRMG*

Cette classe fille hérite des contraintes de la classe *noun*, dont l'arbre précédent. Cette description établit que le syntagme nominal peut être ancré par un nom commun, car le nœud Nc prend les particularités du nœud Anchor. Cette classe présente aussi des contraintes de précédence (cf. figure 3.7).

```
N2 >> det;
det < N;
```

(a) Description de contraintes dans *FRMG*



(b) Arbre partiel correspondant

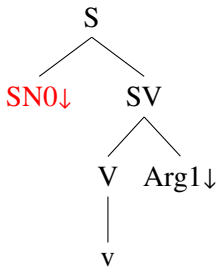
FIGURE 3.7 – Exemple de contrainte de précédence entre nœuds dans *FRMG*

De la métagrammaire à la grammaire

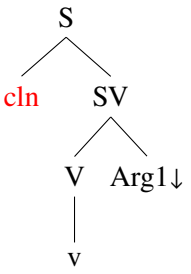
La compilation de la métagrammaire en une grammaire est effectuée par *mgcomp* (*metagrammar compiler*, THOMASSET et VILLEMONTE DE LA CLERGERIE (2005)), développé en *DyALog*. Cette compilation se déroule en trois étapes (VILLEMONTE DE LA CLERGERIE et al. 2009) : tout d'abord, les classes terminales héritent des classes parentes, puis les ressources sont consommées, ce qui permet d'obtenir un ensemble de classes neutres (i. e. qui ne fournissent ou ne demandent plus aucune ressource), enfin, on génère un ensemble d'arbres factorisés minimaux à partir de chaque classe neutre. Les nœuds intermédiaires supplémentaires sont donc supprimés.

La grammaire ainsi générée est constituée d'arbres factorisés, qui le sont au moyen d'opérateurs réguliers (VILLEMONTE DE LA CLERGERIE 2010) :

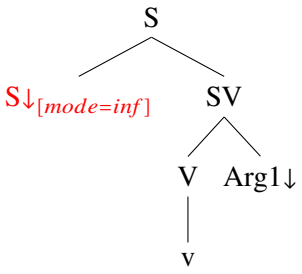
— La disjonction permet d'introduire des alternatives dans la dérivation. Par exemple, on peut faire figurer les différentes réalisations possibles du sujet canonique (cf. figures 3.8a, 3.8b et 3.8c) dans un même arbre [2] (cf. figure 3.8d).



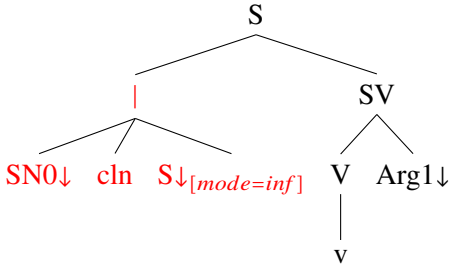
(a) Sujet nominal



(b) Sujet pronominal (clitique)



(c) Infinitive sujet



(d) Expression des différentes réalisations possibles du sujet

2. Cet exemple est tiré de la documentation en ligne de *FRMG*, disponible à cette adresse : http://alpage.inria.fr/frmgwiki/node/10604.

— Les gardes permettent la gestion des nœuds optionnels. Dans *FRMG*, les verbes sans sujet sont les verbes à l'infinitif, à l'impératif ou au participe passé. Cette contrainte est exprimée avec le symbole => (cf. figure 3.9).

```
Arg0 =>
  node(V).top.mode = value(¬inf|imp|part);
¬ Arg0 =>
  node(Arg0).bot = value(inf|imp|part);
```

FIGURE 3.9 – Description de garde dans la métagrammaire

En l'absence de Arg0 (i.e. ¬ Arg0), dont les réalisations possibles sont sous la disjonction, le mode du verbe est une des trois possibilités précédentes (en rouge dans l'arbre 3.10), sinon c'est un verbe conjugué à un autre mode (en vert).

— La répétition est possible grâce à l'étoile de Kleene, qui est surtout utilisée pour la coordination. Cela permet de répéter plusieurs fois un élément, mais aussi de ne pas le laisser apparaître. L'arbre 3.11 permet d'analyser l'énumération de l'exemple suivant en s'adjoignant sur le syntagme nominal *la pierre*.

(6) Item, doit avoir et prendre en ladite forest la
ADVgen VERcjg VERinf CONcoo VERinf PRE DETdef.ADJqua NOMcom DETdef
pierre, le caillou, la maille, la mousse et l'
NOMcom DETdef NOMcom DETdef NOMcom DETdef NOMcom CONcoo DETdef
argile
NOMcom
'De même, il doit avoir avec lui et prendre dans ladite forêt la pierre, le caillou, la maille, la mousse et l'argile'
Coutumier des forêts d'Hector de Chartres, p. 76 (1398–1409)

— Enfin, l'entrelacement (symbolisé par ##) permet un ordre libre entre nœuds frères. Dans la métagrammaire du français médiéval, on représente l'ordre libre entre constituants majeurs grâce à cet opérateur (cf. figure 3.12).

Dans cet exemple, l'arbre permet d'analyser les séquences suivantes :

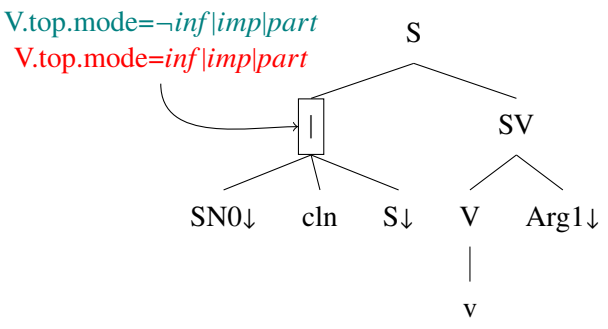


FIGURE 3.10 – Exemple de garde dans un arbre de *FRMG*

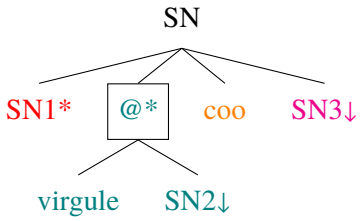


FIGURE 3.11 – Exemple d'énumération dans *FRMG*

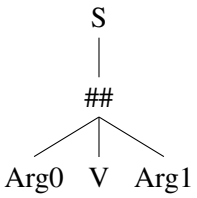


FIGURE 3.12 – Exemple d'entrelacement dans *MetaMOF*

— Arg0 V Arg1
— Arg0 Arg1 V
— Arg1 V Arg0
— Arg1 Arg0 V
— V Arg0 Arg1
— V Arg1 Arg0

Ces opérateurs permettent d'étendre le formalisme *TAG*, mais ils n'ont pas d'impact sur sa complexité ou son expressivité. En revanche, ils réduisent grandement la taille de la grammaire (THOMASSET et VILLEMONTE DE LA CLERGERIE 2005), rendant son usage plus efficace.

De la grammaire au parseur

Dans sa version initiale, *DyALog* compile la grammaire *TAG* en un analyseur tabulaire (*chart parser*) basé sur l'algorithme d'EARLEY (1970) qui fonctionne comme un automate à deux piles. L'objectif est de représenter les traversées possibles de chaque arbre élémentaire (de gauche à droite), ce qui est modélisé "par une série de suspensions/reprises au niveau des nœuds de substitution, nœuds d'adjonction et nœuds pied" (VILLEMONTE DE LA CLERGERIE et al. 2009). L'analyse se fait de manière descendante et les portions d'arbre analysées sont stockées dans une pile (VILLEMONTE DE LA CLERGERIE 2001), ce qui permet de détecter rapidement des erreurs grâce à la propriété de la validité du préfixe (ALONSO et al. 1999 ; SCHABES 1991). La forme tabulaire de l'analyseur permet de stocker les calculs et de les partager, allégeant ainsi le coût computationnel de l'analyse syntaxique. La version actuelle de *DyALog* a évolué vers les *Thread Automata* (VILLEMONTE DE LA CLERGERIE 2002). Lors de l'analyse syntaxique, on peut suivre plusieurs *threads* (fils), mais un seul est actif à la fois, car il est possible de les suspendre. Cela permet de traiter des constituants discontinus. La complexité peut donc être plus grande, mais ce système est plus efficace en pratique.

La complexité maximale (i. e. dans le pire des cas) des *TAG* est $O(n^6)$, ce qui est donc le cas dans l'implémentation de *FRMG* (VILLEMONTE DE LA CLERGERIE 2002). En pratique, une majorité des phrases ne l'atteint pas. Cependant, l'analyse syntaxique est souvent accélérée, car la plupart des arbres de *FRMG* sont *TIG*, c'est-à-dire qu'ils sont assez contraints pour garantir une complexité cubique ($O(n^3)$), équivalente à celle des grammaires hors-contexte (VILLEMONTE DE LA CLERGERIE et al. 2009). La plupart des arbres de *FRMG* sont *TIG*, et seuls certains phénomènes restent *TAG*, comme la construction comparative, lors-

qu'elle a une partie droite. Il est difficile de contrôler cela lors de la description des contraintes. Pour s'assurer que le nombre d'adjonctions englobantes est limité, il faut faire une analyse automatique de la grammaire pour détecter ces cas spécifiques et les contrôler manuellement [3].

Le parseur retourne toutes les analyses complètes qui sont possibles pour une phrase. Puis un module de désambiguïsation s'appuie une heuristique définie manuellement pour sélectionner la meilleure analyse (Villemonte de la Clergerie 2013). Ces règles ont des poids qui peuvent être adaptés par apprentissage artificiel sur un corpus arboré, comme le *French Treebank* (*FTB*, A. Abeillé, Clément et Toussenel (2003)). Les schémas d'annotation du parseur et des ressources sont différents, mais il est possible d'apprendre des poids sur des sous-arbres. Si le parseur ne peut pas retourner au moins une analyse complète de la phrase, il en fait une analyse "robuste", qui combine des analyses partielles. On obtient ainsi des sous-arbres qui ne sont pas liés à une seule tête.

3.2.2 Fonctionnement de la chaîne de traitement

Les différents composants de la chaîne de traitement permettent de traiter des textes bruts. Le segmenteur *Sxpipe* (Sagot et Boullier 2008) propose tout d'abord une segmentation en mots et en phrases. Il transforme ces phrases en treillis de mots (ou *Directed Acyclic Graph*, *DAG*), ce qui permet de conserver l'ambiguïté jusqu'à l'analyse syntaxique.

Les tokens ainsi dégagés doivent être analysables pour être utilisés par le parseur, c'est-à-dire qu'ils doivent être présents en tant qu'entrées dans le lexique ou avoir une structure décomposable pour être traitée par les modules gérant les affixes et les mots composés. S'il s'agit d'un token inconnu, une correction orthographique est proposée pour l'identifier en le rapprochant d'une entrée du lexique. Le segmenteur peut aussi reconnaître les préfixes et les suffixes d'une liste pré-établie, puis rapporter la racine à une entrée du lexique. Des règles reconnaissent les chiffres arabes et romains, qui ne sont pas renseignés individuellement dans le lexique. Certaines traitent des formes spéciales, constituées de plusieurs tokens, comme les dates.

Dans un deuxième temps, les transitions dans les *DAG* sont enrichies d'informations morphologiques et syntaxiques issues d'un lexique, comme *Lefff* pour le français contemporain (Villemonte de la Clergerie et al. 2009). Ces structures de traits sont structurées sous forme d'*hypertags* (Kinyon 2000), qui sont compatibles avec ceux de la métagrammaire. Garder toutes les segmentations possibles pour une phrase et toutes les interprétations possibles des tokens permet de préserver l'ambiguïté et de déléguer le tri à l'analyse syntaxique.

Le parseur procède alors à l'analyse syntaxique et produit une forêt d'analyses à partir de chaque *DAG* (Villemonte de la Clergerie 2013). Elle contient toutes les analyses possibles de la phrase traitée. Dans le meilleur des cas, le parseur propose des analyses complètes des phrases, mais, en cas d'échec, il passe en mode "robuste". Les analyses sont converties dans des formats dépendenciels en utilisant les ancres des arbres élémentaires comme sources et cibles de dépendances syntaxiques. C'est le module de désambiguïsation qui sélectionne la meilleure analyse à partir des scores attribués à chaque arc dans la forêt d'analyse.

3. Dans le cas de *FRMG*, la liste des adjonctions englobantes est générée dans le fichier *tig_header.tag*

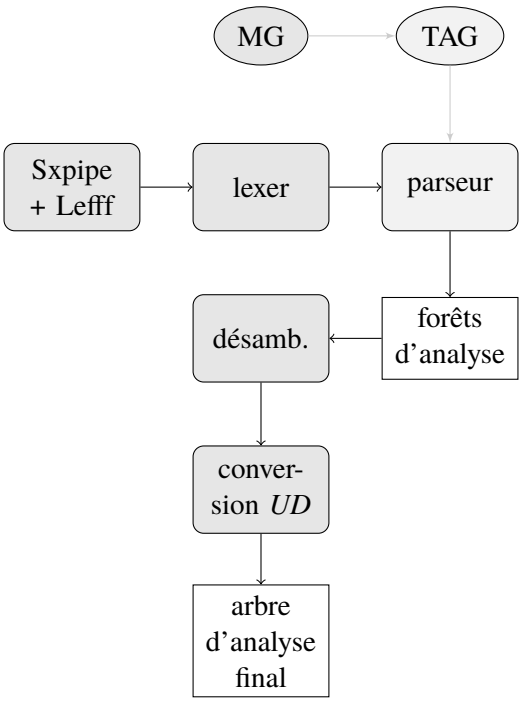


FIGURE 3.13 – Architecture de la chaîne de traitement développée par BOULLIER et al. (2005)

3.2.3 Langues de spécialité

Dans *FRMG*, la description syntaxique des constituants majeurs est contenue dans un petit nombre de classes. En revanche, les modifieurs ont des réalisations très diverses, et ils peuvent s'ancrer à différents endroits dans la phrase, ce qui fait de leur description un enjeu majeur du développement VILLEMONTE DE LA CLERGERIE (2012). De plus, ils constituent des facteurs de variation entre corpus.

FRMG permet de traiter non seulement des états de langue "standards", mais aussi des domaines particuliers comme des écrits journalistiques et juridiques (VILLEMONTE DE LA CLERGERIE et al. 2009 ; VILLEMONTE DE LA CLERGERIE et al. 2009). Cette métagrammaire a été progressivement adaptée pour faire l'analyse syntaxique de nouveaux types de données, notamment politiques et scientifiques, comme celles du *Parlement européen*, de *Wikipédia* et de l'*Agence Médicale européenne* (EMA) [4]. *FRMG* a aussi été dotée d'une extension pour traiter des corpus de botanique (ROLE, GAVILANES et VILLEMONTE DE LA CLERGERIE 2007). Elle a été adaptée pour annoter des corpus de français parlé, notamment pour traiter les disfluences (GERDES et al. 2019). Ces changements ne concernent pas uniquement le français contemporain. Dans le cadre du projet *TIME US*, les états de langue couverts par *FRMG* ont été étendus au français moderne (17e-20e s.) pour analyser des archives industrielles (CHAGUÉ et al. 2019). L'adaptation du parseur ne dépend donc pas seulement du lexique, mais aussi des structures syntaxiques. De nouvelles descriptions sont souvent nécessaires, mais l'essentiel de la grammaire est conservé à chaque changement de corpus, car il est rare qu'un phénomène disparaisse entièrement. Au contraire, les constructions syntaxiques apparaissent généralement dans tous les registres, et parfois à des époques différentes, mais à des fréquences différentes.

4. Les résultats du traitement de ces corpus sont disponibles sur le site de FRMG : http://alpage.inria.fr/frmgwiki/wiki/corpus-disponibles.

3.3 Adaptation d'une métagrammaire pour des états de langue anciens

3.3.1 Motivations de la démarche

Nous avons fait le choix de faire une seule grammaire pour traiter toute la période du français médiéval, car il ne nous est pas possible de décrire individuellement chaque état de langue. En effet, la délimitation nette des états de langue reste difficile, en particulier parce que les états successifs s'inscrivent dans un *continuum* de langue et possèdent un socle de caractéristiques communes, même si les différents phénomènes syntaxiques n'y sont pas également représentés. Construire plusieurs grammaires pour des états de langues serait, par ailleurs, contraire aux principes donnés par A. Abeillé (2002) car un tel découpage empêcherait l'ensemble des grammaires de bénéficier des descriptions faites dans l'une d'entre elles.

FRMG est une métagrammaire suffisamment versatile pour être l'objet d'un tel projet (cf. partie 3.2.3). Dans le cadre du projet ANR *Profiterole*, un lexique du français médiéval est développé sur le modèle du *Lefff* (Sagot 2019), qui permet l'analyse syntaxique de textes d'ancien français et de moyen français.

3.3.2 Implications du choix d'une grammaire *TAG*

Le FMed présente quelques défis à l'implémentation dans une grammaire, notamment à cause de l'ordre libre des constituants majeurs. *FRMG* dispose d'un opérateur utile à cet égard : l'entrelacement (##), qui permet un ordre libre entre nœuds frères [5]. En laissant l'ordre entre deux nœuds sous-spécifié dans la description métagrammaticale, cet opérateur apparaît dans l'arbre de la grammaire. L'implémentation *XLE* du formalisme *LFG* dispose d'un opérateur semblable en amont, dans les règles syntagmatiques qui permettent de générer des arbres. Il s'agit de l'opérateur de brassage, *shuffle* (Kaplan et Maxwell III 1994), qui est représenté par une virgule et qui entraîne la génération des variations de l'arbre avec les différents ordres possibles. En *HPSG*, les règles de dominance (*ID*) sont séparées de celle de linéarisation (*LP*), qui donnent des contraintes explicites d'ordre ou le sous-spécifient (Müller et al. (2021), chap. 10). En français, les constituants majeurs sont généralement représentés au même niveau (Abeillé et Godard 1999), ce qui permet non seulement d'obtenir un ordre libre, mais aussi d'ajouter un trait de poids aux constituants, apportant ainsi des contraintes d'ordre plus souples. Il existe donc diverses façons de traiter l'ordre libre dans les grammaires.

TAG présente aussi des différences avec les autres formalismes, par exemple dans le traitement de la relative. Dans *FRMG*, il s'agit d'un modifieur qui s'adjoint à un syntagme nominal. La proposition relative a une structure différente de la phrase canonique, car un élément en est extrait et repris par le pronom relatif. En *HPSG*, dans les cas les plus courants, il s'agit aussi d'une adjonction (Müller et al. (2021), chap. 14), et la relative contient un *gap* qui correspond à l'argument extrait. Comme avec *TAG*, la propagation des traits avec la tête permet d'utiliser des contraintes d'accord. Müller et al. (2021) indique cependant que le traitement des relatives n'est pas unifié dans le formalisme *HPSG*, et que les dépendances à longue distance et la complémentation d'un élément nominal reçoivent une analyse différente. Dans *FRMG*, une solution a été trouvée pour analyser le premier cas (cf. figure 4.31) et il est possible d'attribuer des arguments aux noms, ce qui permet de traiter la complémentation par une substitution.

5. De plus, les branchements des nœuds ne sont pas binaires, ce qui permet à plusieurs nœuds d'apparaître dans divers ordres.

3.3.3 Méthodologie de l'adaptation d'une métagrammaire

Travaux existants

Il est possible d'adapter des grammaires à visée générique pour traiter des domaines spécifiques, comme cela a été fait dans le cadre du projet *XLE* (Kaplan, King et Maxwell III 2002) avec la grammaire *ParGram STANDARD* de l'anglais (formalisme *LFG*). Dans un premier temps, elle a été développée pour traiter des états de langue "standards". Pour annoter deux nouveaux corpus, l'un composé d'écrits techniques, et l'autre d'articles de journaux (*UPenn Wall Street Journal treebank*, Marcus et al. (1994)), elle a été déclinée en deux nouvelles grammaires : *EUREKA* et *WSJ*. Les développeurs ont souhaité limiter les modifications pour préserver le cœur de la grammaire. Des mécanismes ont été mis en place grâce à un fichier de configuration pour appeler les parties modifiées nécessaires à un nouveau corpus. Cela permet d'avoir une seule architecture pour des ensembles d'analyse différents.

Les travaux de Rocio et al. (2003) prouvent qu'il est possible d'adapter une grammaire existante en la prenant comme noyau pour traiter des états de langue plus anciens (cf. partie 1.4.2). Il s'agit alors d'intégrer des règles spécifiques à la nouvelle langue considérée, en l'occurrence le portugais médiéval, et de développer un nouveau lexique. Ces étapes sont envisageables pour le français médiéval, car la structure de la métagrammaire *FRMG* facilite les modifications de contraintes et permet de générer une nouvelle grammaire.

Certains projets couvrent plus d'une langue, comme celui de la *LINGO Grammar Matrix* permet de générer de nouvelles grammaires sur-mesure à partir d'un questionnaire qui ajoute des analyses issues de librairies à une "grammaire-cœur" (*core grammar*, Bender, Flickinger et Oepen (2002)). Avec ce *kit* de démarrage, les utilisateurs accélèrent considérablement le développement initial. Ce projet traite désormais plus de 120 langues [6], issues de familles différentes (indo-européenne, celtique, afro-asiatique, sino-tibétaine, austronésienne, créoles...).

Adaptation globale de *FRMG*

Notre objectif est d'utiliser *FRMG* comme une grammaire de base et de ne modifier que ce qui est nécessaire. D'une part, il faut identifier les phénomènes fondamentaux qui sont différents. Les premiers efforts doivent porter sur la structure verbale et les arguments du verbe. Ces phénomènes représentent le cœur de la grammaire et ils sont traités par un petit nombre d'arbres (Villemonte de la Clergerie 2012). L'une des spécificités majeures du français médiéval réside dans l'ordre souple des constituants majeurs, qui demande des changements importants dans la grammaire.

D'autre part, il faut identifier tous les changements qui interviennent au niveau des modifieurs. Ces modifications peuvent être effectuées en parallèle des premières, mais elles concernent plus de classes dans la métagrammaire, ce qui rend leur identification et leur traitement plus lent. On peut citer en exemples le complément déterminatif et le détachement des relatives.

Les interactions avec le lexique ne sont pas identiques à celles entre *FRMG* et le *Lefff*, car certaines unités ne sont pas figées en français médiéval. Il convient donc de décrire syntaxiquement certains phénomènes, comme les conjonctions de subordination complexes (ex. *afin que*, *por ce que*).

6. La liste des langues couvertes par la *Grammar Matrix* sont disponibles à cette adresse : https://wiki.ling.washington.edu/bin/view.cgi/Main/LanguagesList.

Pour une personne non spécialiste de français médiéval, l'adaptation d'une métagrammaire existante et la description de nouveaux phénomènes conduit souvent à suivre ces étapes :

1. catégoriser les phénomènes (présents dans *FRMG* ou non, actants ou circonstants (terminologie de Tesnière)...)
2. consulter leur description dans des grammaires traditionnelles et des travaux de linguistique
3. faire des relevés en corpus pour en mieux comprendre la structure (points d'insertion dans des arbres élémentaires, ordre des éléments dans l'arbre, possibles adjonctions intermédiaires...), mais aussi en déterminer les limites (emploi sur un lexique fermé, fréquence d'utilisation, nécessité de créer diverses descriptions pour des cas incompatibles au sein de la grammaire, compatibilité avec le formalisme TAG)
4. décider des changements à adopter : relâchement de contraintes, nouvelles classes, ajout de contraintes syntaxiques, lexicales ou sémantiques... et de la forme de ces changements, en fonction des possibilités offertes par le formalisme

4 Description syntaxique

Des travaux sur l'adaptation de grammaires ont prouvé qu'il était possible d'adapter un système existant à des états de langue plus anciens (Rocio et al. 2003) ou des langages de spécialité (Kaplan, King et Maxwell III 2002). Notre hypothèse est que le français contemporain et le français médiéval, bien que très différents, possèdent assez de points communs pour permettre l'adaptation de la métagrammaire *FRMG* à des états de langue plus anciens. La plupart des phénomènes syntaxiques de la langue moderne étaient déjà présents au Moyen Âge, mais dans des proportions souvent différentes.

Pour ce faire, nous proposons une première étape d'étude des phénomènes syntaxiques majeurs pour comparer, d'une part, leurs structures en français médiéval, et, d'autre part, leur forme en français contemporain et leur traduction dans la métagrammaire *FRMG*. Nous modifions alors ce qui semble nécessaire. Pour faire cet examen, nous avons besoin d'exemples de constructions dans les corpus disponibles, ce qui nous permet de constituer un ensemble de test de non-régression [1] à mesure que les différents phénomènes sont abordés. L'évaluation sur corpus se fait dans un deuxième temps et elle permet d'affiner la compréhension de la syntaxe des textes à analyser (Baldwin et al. 2008).

4.1 Le syntagme nominal

4.1.1 Les noms et les pronoms

Aucune altération n'a été apportée aux classes de base des noms propres et des noms communs. Leur fonctionnement est similaire en français contemporain et en français médiéval, malgré la présence d'une déclinaison bicasuelle. Celle-ci comporte un cas sujet (pour les sujets, les attributs du sujets, leurs appositions et les vocatifs) et un cas régime, pour toutes les autres fonctions. On constate cependant que, très tôt, de nombreux textes ne respectent pas parfaitement la déclinaison, comme en témoignent des textes en normand et en anglo-normand. Son abandon progressif s'effectue dans un mouvement d'ouest en est, les dialectes du nord et de l'est (picard, wallon, lorrain et bourguignon) la conservant en effet jusqu'au MF (Carlier, Guillot-Barbance et al. 2020). Le statut de cette déclinaison est donc très variable d'un texte à l'autre, ce qui ne permet pas de l'utiliser pour l'analyse syntaxique.

La classe des pronoms a connu quelques évolutions, mais le comportement syntaxique et l'autonomie des pronoms disjoints sont assez similaires à ceux du FC (Marchello-Nizia et Prévost (2020), p. 1102-1120) pour conserver les classes de *FRMG*. Ils ont en effet la même autonomie que le substantif. On peut ainsi les trouver :

— en début de phrase (séparés du verbe par d'autres constituants)

1. Ce type de tests est habituel en développement logiciel. Nous utilisons une plate-forme de versionnage pour partager les dernières mises à jour de la métagrammaire et les conserver. Pour ne pas compromettre les développements précédents, il convient de s'assurer qu'on obtient toujours les résultats acquis. Pour cela, nous constituons un ensemble de phrases simples qui sont parsées avant la mise en ligne d'une nouvelle version de la métagrammaire (cf. partie 6.1).

(1) **Moi** et Yseut, que je voi ci, En beümes :
PROper CONcoo NOMpro PRorel PROper VERcjg ADVgen PROadv VERcjg
demandez li !
VERcjg PROper
‘Moi et Yseult, que je vois ici, en avons bu : demandez-lui !’
Yvain de Chrétien de Troyes, v. 988 (1177–1181), cité par Hasenohr et Raynaud de Lage (1993)

— introduits par une préposition

(2) Entr’ **ax** est li diax si granz
PRE PROper VERcjg DETdef NOMcom ADVgen ADJqua
‘Parmi eux, l’affliction est si grande.’
Yvain de Chrétien de Troyes, v. 988 (1177–1181), trad. Hasenohr et Raynaud de Lage (1993)

Le statut des pronoms conjoints est expliqué dans la partie 4.3.3.

4.1.2 Les déterminants

Cas géneral Tout comme les noms, les déterminants sont soumis à une déclinaison en AF. Cependant, elle n’est pas toujours respectée non plus. Il est fréquent de trouver des occurrences de l’article régime en position de sujet, comme dans cet exemple (nous utilisons les étiquettes morpho-syntaxiques *Cattex*[2]) :

(3) si boen servise fist **cel** saint homo
ADVgen ADJqua NOMcom VERcjg **DETdem** ADJqua NOMcom
‘ce saint homme fit un si bon service’
Vie de Saint Alexis, v. 612 (ca. 1050)

La forme *cel* est une forme du cas régime, qui s’illustre aussi pas l’absence de *-s* au singulier. On trouve parfois l’article sujet (ex. *li*) en position régime :

(4) Li chien **li** cerf sivent qui fuit
DETdef NOMcom **DETdef** NOMcom VERcjg PRorel VERcjg
‘Les chiens suivent le cerf qui fuit.’
Tristan de Béroul, v. 1 706 (fin 12e s.)

La déclinaison n’est donc pas prise en compte pour les déterminants, mais l’accord en genre et en nombre est gardé dans les contraintes d’unification. Par ailleurs, en AF, l’article est souvent omis. Il n’est présent que lorsque l’entité doit être actualisée (Hasenohr et Raynaud de Lage 1993). En MF, sa présence devient plus fréquente. *FRMG* prévoit déjà ce cas, en plaçant une garde sur la présence du déterminant dans le syntagme nominal. Aucun changement n’est nécessaire dans ce sens.

L’article partitif Il existe plusieurs manières d’exprimer le partitif, c’est-à-dire “le prélèvement d’une portion sur une étendue de même nature plus vaste” (Buridant (2000), § 83). La première est l’absence de détermination devant un nom massif au singulier (cf. exemple 5).

2. L’inventaire des étiquettes *Cattex* est disponible en annexe (cf. tableau 1).

(5) Si mengierent pain et cervoise
ADVgen VERcjg NOMcom CONcoo NOMcom
'Ils mangèrent du pain et [burent] de la cervoise.'
Queste del saint Graal, p. 190c (ca. 1225 ou 1230), trad. Buridant (2000)

Cette première structure est concurrencée très tôt par une tournure *quantifieur* + *de*, comme en FC. *FRMG* possède déjà cette construction, qui s'adjoint sur les noms communs et les pronoms. Il s'agit d'un déterminant constitué d'un pronom (ex. *peu*) ou d'un équivalent comme l'adverbe *suffisamment* et d'une préposition (*de*). Certaines constructions sont lexicalisées sous forme de déterminants complexes dans le *Lefff* (Sagot 2010), comme *assez de/des/du, assez de la, beaucoup de/des/du, beaucoup de la*. Les mots qui peuvent occuper la position de quantifieur sont renseignés dans un fichier annexe du lexique, *missing.lex*. Ils reçoivent l'étiquette *predet*. Nous comparons les entrées du FC avec les résultats de requêtes dans la *BFM* pour avoir un inventaire adéquat. Nous obtenons notamment des variants graphiques de *peu* : *pois* et *poi*. Ces résultats peuvent être comparés à ceux qu'on obtient en cherchant un patron en dépendances dans le *SRCMF* (cf. figure 4.1). Cette requête est complémentaire, car elle autorise la présence d'autres éléments entre les tokens recherchés, ce qui permet de s'assurer de la bonne description de la construction dans la métagrammaire.

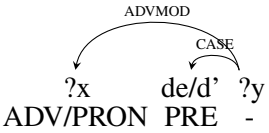


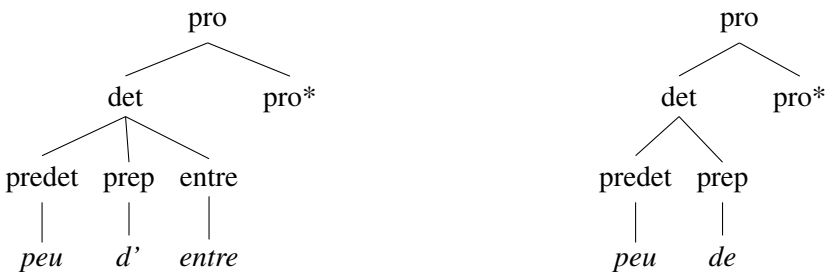
Figure 4.1 – Patron d'article partitif

En FC, cette structure peut même être un peu plus complexe devant un pronom, en comprenant potentiellement la préposition *entre* (cf. figure 4.2a) : *peu d'entre eux*. Comme elle n'est pas attestée dans la *BFM*, elle a été retirée de la métagrammaire (cf. figure 4.2b). Le jeu d'étiquettes des métagrammaires et des lexiques est en annexe (cf. tableau 2).

(6) Et de la partie des assègans fut mort ung gentil
CONcoo PRE DETdef NOMcom PRE.DETdef NOMcom VERcjg VERppe DETcar ADJqua
homme, nommé Gauthier de Pavant, et **peu d'** aultres avec luy.
NOMcom VERppe NOMpro PRE NOMpro CONcoo PROind PRE ADJqua PRE PROper
'Et, parmi les assiégeant, un jeune homme nommé Gauthier de Pavant mourut, et avec lui, peu d'autres personnes.'
Chronique d'Enguerrand de Monstrelet, p. 86 (1441-1444)

Le partitif sans quantifieur apparaît au 12e siècle. Il permet d'extraire "une fraction indéterminée d'un tout spécifique parfaitement déterminé" (Buridant 2000). Il n'est constitué que de la préposition *de* devant un substantif spécifié (cf. figure 4.3), c'est-à-dire précédé d'un article défini (cf. exemple 7) ou possessif (cf. exemple 8).

(7) Done moi **de la** ceue de ton destrier
VERcjg PROper **PRE DETdef** NOMcom PRE DETpos NOMcom
'Donne-moi un morceau de la queue de ton destrier.'
Aiol, v. 2 893 (fin 12e s.), trad. Buridant (2000), §84



(a) Prédéterminant adjoint sur un pronom en FC, ex. "Peu **d'entre eux** sont arrivés jusqu'ici."

(b) Prédéterminant adjoint sur un pronom en FMed

FIGURE 4.2 – L'article partitif ("prédéterminant") dans les métagrammaires

(8) Encontré a de son seignor
VERppe VERcjg **PRE DETpos** NOMcom
'[il] a trouvé les traces de son seigneur'
Tristan de Béroul, v. 1 498 (fin 12e s.), trad. TILANDER (1963)

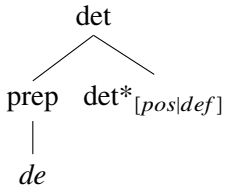


FIGURE 4.3 – Partitif sans quantifieur

L'expression de la possession Le déterminant possessif existe dès l'AF, mais on trouve aussi la séquence avec un déterminant défini, indéfini ou démonstratif suivi d'un adjectif possessif (CARLIER, GUILLOT-BARBANCE et al. 2020). Cet adjectif est traité comme un adjectif qualificatif dans la métagrammaire.

(9) ne vient pas de la moye merite mais de la tue
ADVneg VERcjg ADVneg PRE DETdef ADJpos NOMcom CONcoo PRE DETdef ADJpos
merci
NOMcom
'[Ma liesse] ne vient pas de mon mérite (du mien mérite), mais de ta pitié (la tienne pitié)'
Commentaire en prose sur les psaumes I-XXXV, p. 164 (ca. 1163)

4.1.3 Les modifieurs simples du nom

Les adjectifs

Comme en FC, les adjectifs épithètes peuvent s'adjoindre à gauche ou à droite du nom qu'ils modifient. L'adjonction antéposée se fait sur *N* pour garantir d'apparaître après le déterminant, et l'adjonction postposée se fait à un niveau au-dessus, sur le syntagme nominal (*N2*), permettant ainsi l'adjonction d'éléments intermédiaires (ex. 10).

Pour le FMed, on préserve la structure des arbres et l'accord en genre et en nombre, mais on ignore à nouveau le cas.

(a) Adjectif qualificatif postposé (b) Adjectif qualificatif antéposé

FIGURE 4.4 – Les adjectifs qualificatifs dans *FRMG*

(10) Avoit oblié ne sai qui .I. **peigne** d' ivoire **doré**
VERcjg VERppe ADVneg VERcjg PROint DETcar **NOMcom** PRE NOMcom **ADJqua**
'Avait oublié je ne sais qui un peigne d'ivoire. (= Je ne sais qui avait oublié...)'
Chevalier de la Charrette ou Lancelot de Chrétien de Troyes, v. 1 351 (ca. 1177–1181)

FRMG permet aussi d'analyser les adjectifs apposés, qui sont repris tels quels dans la métagrammaire du FMed.

Les titres

Les titres sont décrits dans *FRMG* comme une co-ancre s'ils sont antéposés (ex. *Madame* la présidente), et comme une apposition sans ponctuation s'ils sont postposés (ex. Guillaume *le conquérant*).

(a) Titre antéposé, ex. "*Madame* la présidente" (b) Titre postposé, ex. "Guillaume *le conquérant*"

FIGURE 4.5 – Les titres dans *FRMG*

L'arbre 4.5a n'est pas modifié, car on observe une structure semblable en FMed (ex. 11). L'inventaire des noms qui peuvent ancrer le nœud *title* est donné explicitement dans un fichier annexe au lexique (*missing.lex*), chargé par le *lexer* au moment de l'analyse. On obtient 293 entrées de titres grâce à de la fouille d'erreur sur corpus et à des requêtes dans la *BFM*.

(11) Li **reis** **Marsilie** esteit en Sarraguce.
DETdef **NOMcom.title** **NOMpro** VERcjg PRE NOMpro
'Le roi Marsile était dans Saragosse.'
Chanson de Roland, v. 10 (ca. 1100), trad. Petit de Julleville

Pour l'arbre 4.5b, nous ajoutons une contrainte sémantique sur l'ancre de *N2*, le trait *titlepost*, pour éviter la confusion avec le complément déterminatif non introduit (cf. partie 4.4.1). Nous reprenons l'inventaire des formes possibles. Tout comme pour la liste des titres antéposés, il s'agit essentiellement de noms de fonction (ex. 12) et de filiation (ex. 13).

(12) Jehennete **La** **Chapeliere**
NOMpro **DETdef** **NOMcom.titlepost**
'Jeannette la chapelière'

Registre criminel du Châtelet, p. 249 (1389–1392)

(13) Ihesu Criz **li filz** de deu
NOMpro NOMpro **DETdef NOMcom.titlepost** PRE NOMpro
'Jésus Christ fils de Dieu'
Li sermon saint Bernart, p. 24 (fin 12e s.)

4.2 Les arguments du verbe

Le cadre valenciel du verbe est divisé en trois arguments essentiels (*Arg0, Arg1, Arg2*) dans *FRMG*. Dans une phrase dite "canonique" – i.e. à la voix active et sans phénomène de mise en relief – *Arg0* est le sujet, *Arg1* est l'objet direct ou l'attribut du sujet et *Arg2*, un complément d'objet second (souvent introduit par *à* ou *de* en français). Ils ne sont pas tous nécessairement réalisés.

4.2.1 Le cadre valenciel

FRMG a une représentation traditionnelle de la phrase canonique (voir Fig. 4.6a). Pour représenter l'ordre libre du FMed, nous avons choisi de rendre cette représentation plus plate (voir Fig. 4.6b), comme le recommandent A. Abeillé, Clément et Toussenel (2003). Nous tirons profit du formalisme *FRMG* qui autorise l'ordre libre entre nœuds frères [3], ce qui permet de simplifier la description. Par exemple, la grammaire utilise le même arbre verbal pour les phrases 4.7a et 4.7b, où les arguments apparaissent dans un ordre différent pour conserver les énoncés (voir partie 4.3 pour la description de la dorsale verbale). Il est ainsi possible de comparer des réalisations différentes de structures profondes semblables, en l'occurrence des verbes transitifs directs. En revanche, garder la représentation de *FRMG* nous aurait obligés à créer des nœuds à plusieurs endroits dans la phrase pour permettre aux arguments du verbe de s'y attacher, ce qui aurait rendu l'arbre de la phrase très complexe.

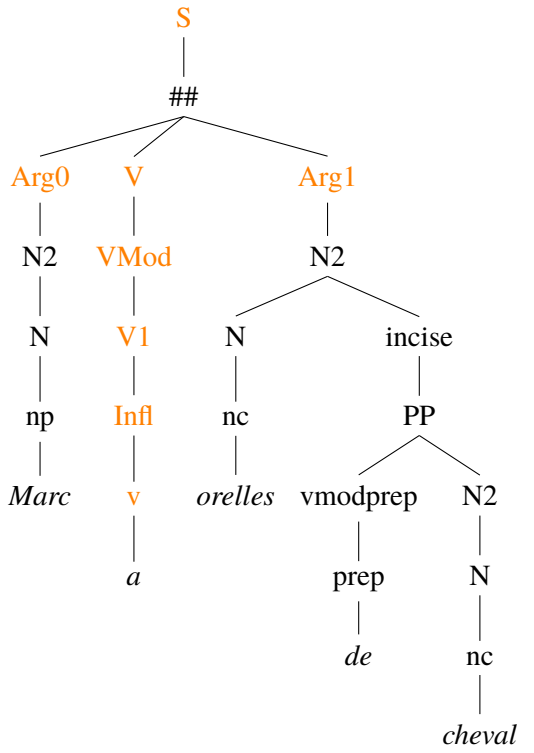
(a) Verbe transitif avec un objet nominal (b) Métagrammaire du français médiéval (*MetaMOF*)

Figure 4.6 – Organisation des constituants principaux dans les métagrammaires

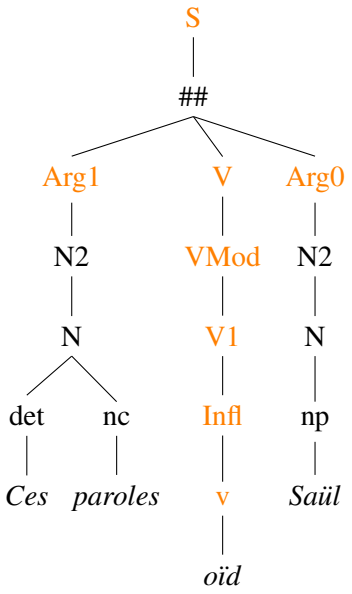
Lors de l'analyse de la phrase, seuls les arguments qu'autorise l'entrée lexicale du verbe principal sont considérés. Ainsi, pour un verbe intransitif, comme *dormir*, le parseur ne cherchera pas de réalisation pour *Arg1* et *Arg2*.

Le lexique *OFrLex* propose quatre types d'arguments non essentiels (cf. partie 5.2.4) : les obliques (*Obl* et *Obl2* [4]), le complément locatif (*Loc*, ex. *aller quelque part*) et le complément "délocatif" (*DLoc*,

3. Pour obtenir cet ordre libre, on n'indique pas de contrainte de précédence entre nœuds dans la métagrammaire. Dans les arbres *TAG*, cela se traduit par l'opérateur d'ordre libre *##*, comme dans la figure 4.6b.
4. On remplit d'abord *Obl* (ex. *lutter **pour***), puis *Obl2* (*lutter **contre***).



(a) “Marc a orelles de cheval”, *Tristan* de Béroul, trad. *Marc a des oreilles de cheval*



(b) “Ces paroles oïd Saül”, *Quatre Livres des rois*, trad. *Saül entendit ces paroles*

FIGURE 4.7 – Arbre verbal et ses arguments sujet et objet direct

ex. *partir de quelque part*). Ils participent du sens du verbe, ce qui justifie leur présence dans le cadre de valences. Leurs réalisations possibles sont aussi listées dans les entrées du lexique.

4.2.2 Réalisations des arguments principaux

Le lexique indique non seulement quels arguments sont attendus pour chaque verbe, mais aussi quelle forme ils peuvent prendre, c’est-à-dire s’il s’agit de syntagmes nominaux (ou pronominaux) ou verbaux, et s’ils peuvent être introduits par une préposition. Il faut alors indiquer laquelle. Dans le lexique, ces prépositions ont un trait qui permet de lier les variants graphiques (ex. *de*, *d’* et *d*). Ces informations permettent de contraindre l’analyse syntaxique.

Le sujet d’une phrase canonique, *Arg0*, peut être réalisé de diverses manières. Il s’agit le plus souvent d’un syntagme nominal ou pronominal, mais certains verbes acceptent un infinitif (ex. figure 4.8) ou une phrase (ex. figure 4.9). Ces descriptions sont reprises directement de la métagrammaire *FRMG*.

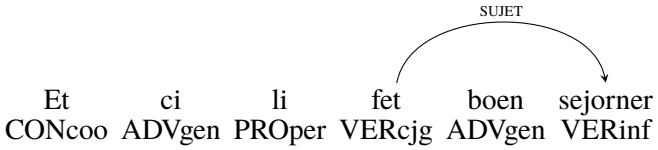


FIGURE 4.8 – “Et [il] lui est agréable de séjourner ici”, *Yvain*, v. 1 395 (1177–1181)

Certaines contraintes pèsent sur la réalisation possible de l’argument dans la zone préverbale en AF (RAINSFORD et al. 2012). Le verbe occupe généralement la deuxième position (ordre V2 (verbe second)

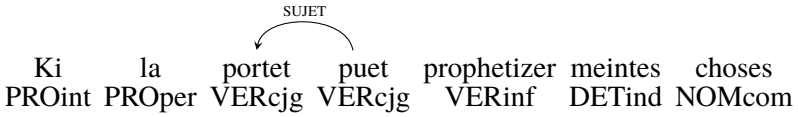


FIGURE 4.9 – "[Celui] qui la porte peut prophétiser de nombreuses choses", *Lapidaire en prose*, p. 104 (milieu 12e s.)

canonique), et le premier élément doit être tonique, selon la loi Tobler-Mussafia. Cette contrainte prosodique, qui exclut les pronoms personnels atones en première position, disparaît à partir de 1200. Cependant, introduire une contrainte de ce type pour la période concernée obligerait à vérifier le statut de chaque début de phrase. Toutes les structures pouvant y figurer devraient alors faire remonter un trait de tonicité à l'énoncé. Cela ne concerne pas seulement les arguments essentiels du verbe, mais aussi les modifieurs de phrase, qui sont très nombreux. L'effet de cette règle est en outre limité dans le temps, ce qui rend le coût supplémentaire trop élevé pour le faible gain dans les textes concernés.

Le complément *Arg1* est décrit en ordre libre avec les autres constituants majeurs, sauf s'il s'agit d'une complétive, auquel cas une contrainte d'ordre est ajoutée pour que cet argument soit nécessairement à droite du verbe.

4.2.3 Expression des arguments

Le lexique indique les arguments possibles de chaque verbe et leurs réalisation, et il spécifie explicitement si ces arguments sont obligatoires. Cela permet, par exemple, d'analyser une phrase ancrée par un verbe transitif, que l'objet soit présent ou non. En FC, l'objet direct et l'objet second sont souvent optionnels, contrairement au sujet. En FMed, le sujet aussi est optionnel.

Les sujets exprimés (S) et non-exprimés (S0) coexistent, mais S progresse très tôt, au détriment de S0 (MARCHELLO-NIZIA et PRÉVOST (2020), p. 1077-1079). On observe un tournant dans l'usage au 13e siècle, en partie lié au développement de la prose dans les écrits, la prose ayant été pionnière dans la progression des sujets exprimés (MARCHELLO-NIZIA et PRÉVOST (2020), p. 1059).

La phrase de l'exemple 4.10 n'a pas de sujet. Aucune réalisation de Arg0 n'est donc attendue dans l'arbre 4.11a. À partir de l'analyse de la grammaire, notamment à partir de l'arbre de dérivation, il est possible d'obtenir l'analyse en dépendances (cf. arbre 4.11b).

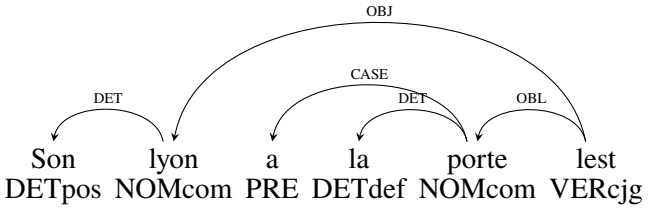
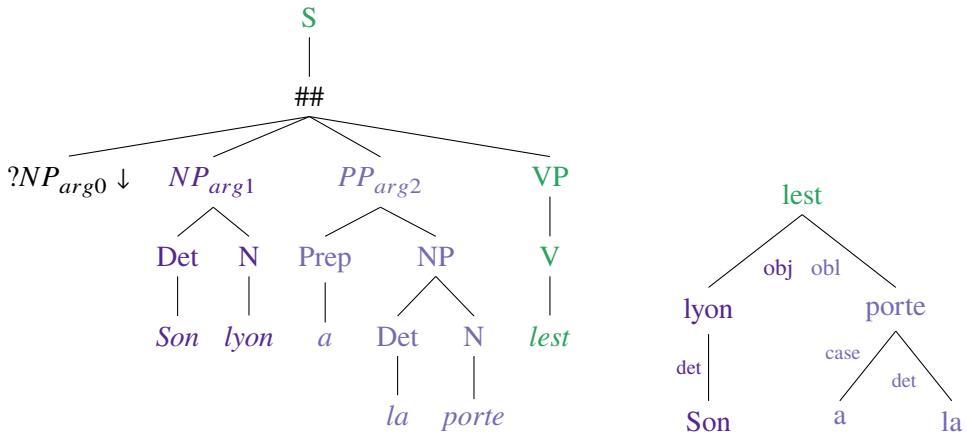


FIGURE 4.10 – "[Il] laisse son lion à la porte", *Yvain* de Chrétien de Troyes, v. 3784 (1177–1181)

L'analyse des phrases ancrées par un verbe transitif est donc rendue ambiguë non seulement à cause de l'ordre souple des constituants principaux, mais aussi par la possibilité de sujets nuls ainsi que d'objets nuls, même si cette dernière est très rare en AF (MARCHELLO-NIZIA et PRÉVOST (2020), p. 1130). La contrainte d'accord peut parfois aider à sélectionner le bon argument, mais c'est surtout le modèle de désambiguïsation entraîné sur corpus qui intervient pour ce choix.



(a) Analyse par la grammaire (b) Arbre de dépendances

FIGURE 4.11

4.2.4 Voix active et voix passive

Par convention, la voix active est considérée comme la voix canonique. Les auxiliaires *estre* (*être*) et *avoir* y sont utilisés pour les temps du passé. Dans le cas de la voix passive, l'auxiliaire du passif apparaît sous *SV*, ancré par un participe passé passif. Dans le lexique, les cadres de valence des verbes passivables font l'objet d'une transformation pour les entrées de participes passés passifs : l'objet direct (Arg1) devient le sujet (Arg0) et le sujet devient le complément d'agent, qui est optionnel. Celui-ci est introduit par la préposition *par*, comme en français contemporain. On trouve quelques occurrences du variant graphique *per*, comme dans cet exemple :

(14) Et cil li fu renduz **per** un Greu de la ville
CONcoo PROdem PROper VERcjg VERppe PRE DETind NOMpro PRE DETdef NOMcom
'Et celui-ci lui a été rendu **par** un Grec de la ville.'
Conquête de Constantinople de Geoffroi de Villehardouin, p. 88 (début 13e s.)

Les variants *por* et *pour* sont, en revanche, rarissimes, et nous avons choisi de ne pas les intégrer à la description du passif, pour éviter d'augmenter l'ambiguïté de l'analyse.

4.3 Le syntagme verbal

Nous n'avons pas gardé le syntagme verbal traditionnel, mais il est nécessaire de préserver une structure verbale pour décrire le placement des auxiliaires, des modaux, des pronoms conjoints et des modifieurs verbaux. Ces éléments sont organisés le long d'une "dorsale", conservée de l'implémentation de *FRMG*.

Le nœud *V* permet l'adjonction des verbes modaux, et le nœud *VMod V1*, celle des incises flottantes dans la phrase. Le nœud *V1* autorise l'accroche des pronoms négatifs postposés (ex. *pas, point, guère*). Les auxiliaires s'adjoignent sur *Infl*, et *v* domine l'ancre du SV.

4.3.1 Le verbe et ses formes conjuguées

Les auxiliaires s'adjoignent sur le nœud *Infl* ancré par un verbe au participe passé. En français contemporain, l'auxiliaire apparaît avant le participe passé. L'ordre est plus souple en FMed. On souhaite pouvoir

FIGURE 4.12 – La dorsale verbale dans *FRMG*

faire apparaître les auxiliaires au même niveau que la tête du syntagme verbal et les arguments du verbe, car ils peuvent être intercalés entre ces éléments.

(15) Quant la beste **at** ço **fait**
CONsub DETdef NOMcom **VERcjg** PROdem **VERppe**
'Quand la bête **a** *fait* cela'
Bestiaire de Philippe de Thaon, v. 1 693 (1121–1135), exemple de l'ordre *Arg0 – Aux – Arg1 – Vppe*

L'auxiliaire peut être aussi postposé.

(16) Li mien baron, **nurrit** vos **ai** lung tens.
DETdef ADJpos NOMcom **VERppe** PROper **VERcjg** ADJqua NOMcom
'Mon baron (le mien baron), je vous **ai** *nourri* longtemps.'
Chanson de Roland, v. 3 374 (ca. 1100), exemple de l'ordre *Vppe – Arg1 – Aux*

4.3.2 Les verbes modaux

Traditionnellement dans les *TAG*, les verbes modaux sont adjoints au syntagme verbal, et ils n'en sont pas la tête, même s'ils portent les marques de conjugaison. Le verbe à l'infinitif permet de conserver l'ancre sémantique d'une phrase et c'est lui qui détermine le cadre valenciel. Dans *FRMG*, ces verbes s'adjoignent sur le nœud *V* (cf. figure 4.13), ce qui autorise la présence d'auxiliaires et de clitiques entre le modal et le verbe à l'infinitif. Ils apparaissent avant ces autres éléments (à gauche de la dorsale). On peut donc analyser des exemples de ce type :

— *Il peut venir.*

— *Il peut le voir.*

— *Il peut l'avoir vu.*

Dans notre métagrammaire, nous introduisons un ordre libre entre ces deux nœuds pour permettre aux modaux d'apparaître à gauche ou à droite des éléments sous *V*, tout en préservant la structure sous *V*.

(17) puis qu' il ne **peut** autrement *estre*
CONsub PROimp ADVneg **VERcjg** ADVgen *VERinf*
'puisqu'il ne peut en être autrement'

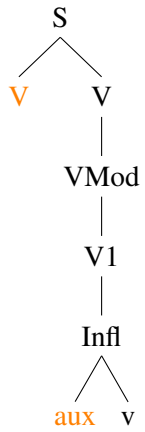


FIGURE 4.13 – Intégration des modaux au syntagme verbal dans *FRMG* (nœuds adjoints symbolisés en orange)

Mélusine de Jean d'Arras, p. 23 (1392), exemple de l'ordre *Vmodal – Vinf*

(18) quant *antrer* **peut** en la cuisine
CONsub *VERinf* **VERcjg** PRE DETdef NOMcom
'quand [il] peut entrer dans la cuisine'
Roman de la rose de Jean de Meun, v. 21 528 (1269–1278), exemple de l'ordre *Vinf – Vmodal*

4.3.3 La gestion des formes de compléments conjoints

Dans *FRMG*, les pronoms conjoints, dits "clitiques" sont ordonnés en une séquence fixe qui s'insère sous *V1*. Ils apparaissent groupés et à proximité du verbe. On trouve, dans l'ordre (cf. figure 4.14) : le clitique négatif (*clneg*, réalisé par *ne*), l'adverbe négatif (*advneg*, réalisé par *pas, point...* et qui peut aussi être adjoint à droite du verbe), le clitique réfléchi (*clr*, réalisé par *se*), et les séquences de clitiques (*clseq*) antéposée et postposée (cf. figure 4.15).

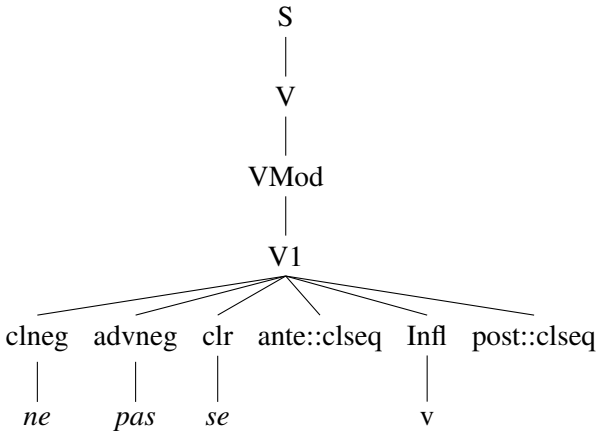


FIGURE 4.14 – Séquence de pronoms conjoints

Les séquences de clitiques comprennent les clitiques accusatifs (*cla*), datifs (*cld12* pour P1, P2, P4 et P5 et *cld3* pour P3 et P6), locatifs (cll *y*) et génitifs (clg *en*). Cette représentation permet de limiter le nombre d'analyses possibles, contrairement à une séquence laissée en ordre libre.

En FC, les clitiques datifs des premières et deuxièmes personnes (ex. *me, te*) précèdent le clitique accusatif.

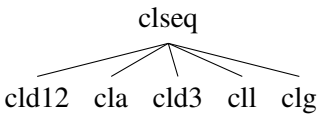


FIGURE 4.15 – Organisation générale des clitiques

ex. Je **te/vous** *l'*ai demandé. Il **me/nous** *l'* a demandé.

En revanche, c'est l'inverse pour les clitiques datifs à la troisième personne.

ex. Je *le* **lui/leur** ai demandé.

En AF, nous n'utilisons pas la dénomination de "clitique". On trouve des formes de pronoms toniques et atones (CARLIER, GUILLOT-BARBANCE et al. (2020), p. 685-686), mais il s'agit d'une distinction phonétique, et les formes toniques ne sont pas toutes disjointes. Nous conservons toutefois l'appellation de "clitiques" pour les pronoms régimes conjoints dans la métagrammaire, afin de garder une continuité entre les deux systèmes. Il est en effet pertinent de traiter les pronoms conjoints en séquence à proximité du verbe, comme en FC. On observe cependant une différence dans l'organisation interne de cette séquence : quelle que soit la forme du pronom datif, il apparaît majoritairement après le pronom accusatif. Pour nous en assurer, nous avons fait des requêtes dans la *BFM* sur des formes pronominales qui ne sont pas ambiguës. Les ordres accusatif – datif et datif – accusatif ont été recherchés avec ces formes conjointes :

— pronoms accusatifs : *le, la, les*

— pronoms datifs : *li, me, te* (par élimination, ces pronoms personnels sont datifs s'ils sont employés avec les pronoms ci-dessus)

Les séquences ont été contrôlées individuellement avec Sophie Prévost. L'ordre accusatif – datif représente l'écrasante majorité des résultats : 634 cas contre 10 pour l'ordre datif – accusatif. Les tableaux de concordances sont en annexe (cf. tableaux 4, 5 et 6). Parmi les dix exceptions (cf. tableau 7), on trouve une anomalie dans les données :

(19) Jel te le di
PROper-cln.**PROper-cla** PROper-cld **PROper-cla** VERcjg
'Je te le dis'
Aucassin et Nicolette, v. 12 (fin 12e s. ou début 13e s.)

Dans cet extrait, le pronom accusatif est présent une première fois dans *jel* (forme contractée de *je* et *le* avec enclise), puis il est répété, ce qui constitue un hapax. Ces phrases ne peuvent donc pas être couvertes par notre grammaire, qui décrit le comportement régulier de la langue.

On modifie donc l'ordre des pronoms conjoints dans *MetaMOF*. Ce n'est qu'en MF que l'ordre commence à s'inverser, au profit de l'ordre *datif – accusatif* lorsque le pronom datif est à la 1ère ou 2e personne (*me/te le*) (MARCHELLO-NIZIA 1979). Cependant, le nombre d'occurrences de cet ordre est trop peu élevé pour motiver un mécanisme de sélection de l'époque pour l'analyse syntaxique dans cette première version de la métagrammaire.

Dans *FRMG*, une contrainte pèse sur la postposition de la séquence de clitiques : le verbe doit être à l'impératif et ne pas recevoir une négation. Pour permettre plus de souplesse dans l'analyse des phrases, cette contrainte est supprimée.

En plus de ces deux points d'ancrage de la séquence de pronoms conjoints, on observe, en FMed (et jusqu'au 17e siècle), une montée du clitique quasiment systématique lorsqu'un modal apparaît, et ce, qu'il soit à droite ou à gauche du verbe à l'infinitif.

(20) Lores **le** *volt* mangier Si le prent a bechier
ADVgen **PROper** *VERcjg* VERinf ADVgen PROper VERcjg PRE VERinf
'Alors il veut le manger, il se met à le frapper du bec.'
Bestiaire de Philippe de Thaon, v. 1 789 (1121–1135)

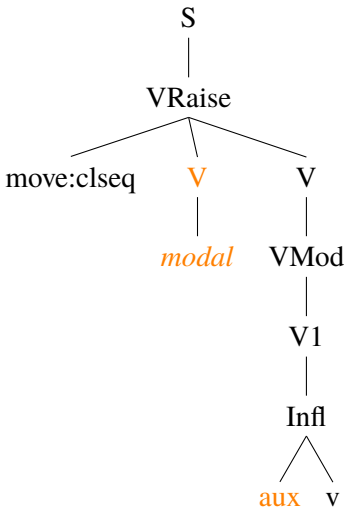


FIGURE 4.16 – Intégration des modaux au syntagme verbal dans *MetaMOF* (nœuds adjoints symbolisés en orange)

Un nœud supplémentaire, *VRaise*, a donc été ajouté (cf. arbre 4.16 et extrait de la description dans la figure 4.17), pour permettre l'ancrage des pronoms conjoints plus haut dans l'arbre. Ils peuvent donc être analysés à gauche du modal, mais remplir le cadre valenciel du verbe à l'infinitif. Cette position exclut les autres (*ante* et *post*).

Il est rare que la séquence de pronoms conjoints ne monte pas, mais cela reste possible, et elle peut alors se réaliser en position *ante* ou en position *post*, comme dans l'exemple 4.18.

4.3.4 Le pronom personnel sujet

Le pronom personnel sujet est aussi appelé "clitique" (nominatif) dans *FRMG*. Ce nom est gardé, mais le pronom personnel sujet a une plus grande autonomie en FMed, sans toutefois se comporter comme un pronom disjoint. Il peut en effet :

```
- move::clitic_sequence;
    VRaise >> move::clseq;
    move::clseq < V;
    move::clseq =>
        node(move::clseq).excl = value(move);
    ~ move::clseq =>
        node(move::clseq).excl = value(-|ante|post);
```

FIGURE 4.17 – Extrait de la description de la séquence clitique

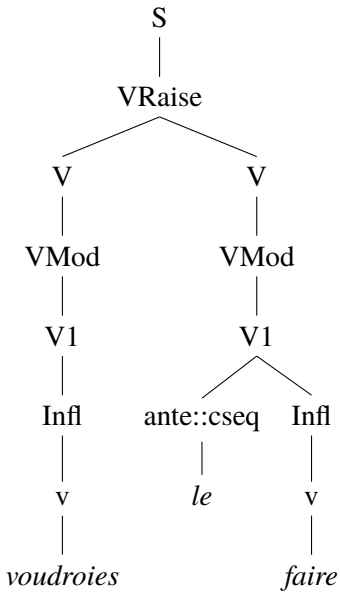


FIGURE 4.18 – “Rois, voudroies le faire issi ?”, *Tristan* de Béroul, trad. *Rois, voudrais-tu le faire ici ?*

— être séparé du verbe, surtout en position préverbale (MARCHELLO-NIZIA et PRÉVOST (2020), p. 1 104),

(21) Ne **il** le roi ne **desfia**
CONcoo **PROper** DETdef NOMcom ADVneg **VERcjg**
‘Et il ne défia pas le roi’
Roman de Thèbes, v. 7 933 (ca. 1150), trad. MARCHELLO-NIZIA et PRÉVOST (2020), p. 1 104

— être coordonné,

(22) **je** **et** les autres prelas dou roiaume vos
PROper CONcoo DETdef ADJind NOMcom PRE.DETdef NOMcom PROper
conseillerons volentiers.
VERcjg ADVgen
‘Moi et les autres prélats du royaume vous conseillerons volontiers.’
Continuation de Guillaume de Tyr, p. 50 (ca. 1200)

— être modifié (le FC garde une trace de cette construction avec la formule *je, soussigné(e) XY, certifie que...*)

(23) lor conmençai ge bien a faire, **je** **qui** onques mes
ADVgen VERcjg PROper ADVgen PRE VERinf **PROper PROrel** ADVgen ADVgen
bien ne fis
ADJqua ADVneg VERcjg
‘alors commençai-je à bien faire, je (=moi) qui jamais n’avais fait de bien’
Roman de Renart, v. 10 710 – 10 711 (début 13e s.), trad. CARLIER, COMBETTES et al. (2020), p. 1 002

Nous ajoutons donc quelques descriptions pour autoriser ces analyses, comme la relative attachée au pronom personnel sujet.

4.3.5 Les modifieurs du verbe

FRMG est déjà équipé de modifieurs du syntagme verbal, comme les adverbes, qui peuvent apparaître à gauche et à droite du verbe. Ils disposent de plusieurs points d'ancrage possibles, ce qui leur permet de séparer ou de laisser groupés les éléments du syntagme. Nous conservons ces descriptions, car le FMed présente des cas similaires.

— Adverbe antéposé

(24) **Sovent** *escumbatirent* mei dès la meie juvente
ADVgen *VERcjg* PROper PRE DETdef PROpos NOMcom
'Ils me combattirent souvent, dès mes jeunes années.'
Psautier d'Oxford, p. 205 (début 12e s.)

— Adverbe devant l'auxiliaire

(25) **Ben** *ad* parlet li dux.
ADVgen *VERcjg* VERppe DETdef NOMcom
'Le duc en a bien parlé.'
Chanson de Roland, p. 243 (ca. 1100)

— Adverbe intercalé entre l'auxiliaire et le participe passé

(26) Genz sanz seignur *sunt* **malement** *bailli*
NOMcom PRE NOMcom *VERcjg* **ADVgen** *VERppe*
'Les gens sans seigneur sont mal dirigés !'
Chanson de Guillaume, v. 287 (ca. 1140)

— Adverbe antéposé et séparé du verbe

(27) Contra ·ls afanz que an a padir toz sos fidels
PRE DETdef NOMcom PROrel PROper PRE VERinf PROind DETpos NOMcom
ben en *garnid*
ADVgen PROper *VERcjg*
'Contre les tourments qu'ils ont à souffrir, tous ses fidèles bien [il] fortifia (=il fortifia bien tous ses fidèles).'
Passion de Clermont, p. 112 (ca. 1000), trad. Jacques-Joseph Champollion-Figeac

— Adverbe postposé

(28) ce fut qu' il me *dist* **soubdainement**
PROdem VERcjg CONsub PROper PROper *VERcjg* **ADVgen**
'C'est ce qu'il m'a soudainement dit.'
Mémoires de Philippe de Commynes, p. 118 (ca. 1490–1505)

— Adverbe postposé et séparé du verbe

(29) Or *sai* jo **veirement** Que hoi murrum
CONcoo *VERcjg* PROper **ADVgen** CONsub ADVgen VERcjg
'Je sais clairement [...] que nous mourrons aujourd'hui.'
Chanson de Roland, v. 1 935 (ca. 1100), trad. Jean Gillequin

4.4 Les modifieurs non phrastiques

Décrire les modifieurs est une tâche majeure du développement d'une métagrammaire (Villemonte de la Clergerie 2012). Dans une phrase, ils peuvent s'insérer à de multiples endroits, et leur construction dépend parfois de leur type sémantique. Leur intégration est un travail à long terme qui bénéficie de la fouille d'erreur sur corpus.

4.4.1 Le complément déterminatif

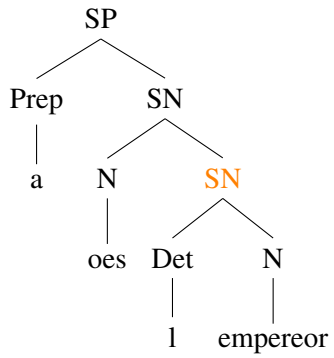


Figure 4.19 – *a oes l'empereor*, *Chanson de Roland*, v. 3 678 (ca. 1100), trad. "au profit [de] l'empereur"

Ce complément nominal restreint l'extension du nom qu'il modifie (Riegel, Pellat et Rioul 1994). En FMed, les compléments déterminatifs ne sont pas nécessairement précédés d'une préposition, comme dans la figure 4.19. En FC, on trouve une construction similaire, par exemple "un département recherche" ou "objectif lune". On a aussi hérité de locutions, comme "Fête Dieu" et "hôtel-Dieu", désormais figées. Dans *FRMG*, le deuxième nom est considéré comme un modifieur qui apporte une dénomination. Il s'adjoint sur le premier nom.

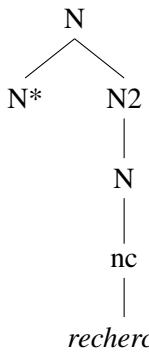


Figure 4.20 – Complément nominal non-introduit en FC, ex. "un département recherche"

En règle générale, nous préférons adjoindre le complément déterminatif sur un syntagme nominal, et non directement sur un nom, pour permettre à des modifieurs d'apparaître entre ces deux éléments. La description de *FRMG* n'est donc pas retenue pour ce cas. En FMed, ces compléments peuvent être placés à gauche ou à droite de leur gouverneur s'ils sont précédés d'une préposition. Nous gardons la description du complément du nom de *FRMG* (cf. figure 4.21a) et lui ajoutons l'équivalent pour le complément antéposé (cf. 4.21b).

(a) Complément du nom dans *FRMG* (b) Equivalent antéposé

FIGURE 4.21 – Arbres de compléments déterminatifs antéposés et postposés, avec une préposition

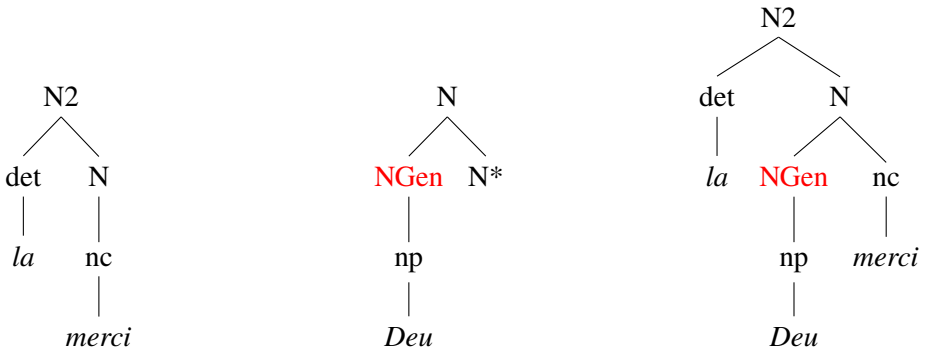
En revanche, en l'absence d'une préposition, le siège de l'adjonction peut changer. Le complément postposé s'adjoint au syntagme nominal (*N2*), mais le complément antéposé s'adjoint sur le nom parce qu'il apparaît à droite du déterminant (cf. figure 4.22). Lorsqu'il est antéposé, il est en outre soumis à une contrainte lexicale. En effet, il ne peut être réalisé que par un de ces mots :

— *Dieu* et ses variantes graphiques *Deu, Deus*, ainsi que d'autres dénominations de divinité, comme *Dé hé* et *Damedieu* (on a d'ailleurs conservé l'expression "Dieu merci" en FC)

— *Jesus*

— *autrui*

— *roi*, non cité dans les grammaires, mais que nous avons découvert dans *Tristan* de Béroul (cf. exemple 30), cela signifie que la liste est peut-être incomplète.

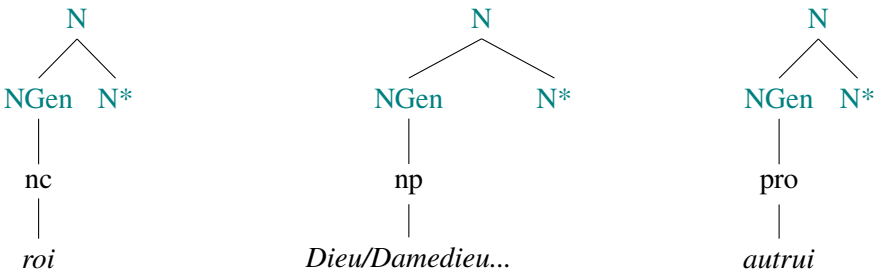


(a) Syntagme nominal : déterminant + nom (b) Arbre auxiliaire du complément déterminatif antéposé et non-introduit (c) Syntagme nominal complet

FIGURE 4.22 – Complément à gauche : *Por Deu merci*, *Melion*, trad. *pour la miséricorde de Dieu*

(30) Fus tu donc pus a la roi cort ?
VERcjg PROper ADVgen ADVneg PRE DETdef NOMcom NOMcom
'N'es-tu donc plus allé à la cour [du] roi ?'
Tristan de Béroul, v. 2 498 (fin 12e s.)

On contraint autant que possible l'adjonction de ce complément déterminatif (cf. figure 4.23).



(a) Nom commun (b) Nom propre (c) Pronom

FIGURE 4.23 – Arbres de compléments déterminatifs antéposés sans préposition ayant différentes ancres

4.4.2 Comparatif et superlatif

Les constructions comparatives sont généralement en deux parties. La première (cf. figure 4.24), obligatoire, consiste en l'adjonction d'un adverbe ou l'utilisation d'un adjectif spécifique (ex. *meilleur*). La seconde (cf. figure 4.25), optionnelle, est liée à la première, et fournit un comparant. Comme elle rend l'adjonction englobante (de part et d'autre de l'ancre), c'est elle qui sort l'arbre adjoint final du cadre *TIG*, majoritaire dans *FRMG*. Cependant, les arbres restent *TAG*.

Le premier arbre peut s'adjoindre aux éléments suivants :

— syntagme adjectival, ex. *si **joli***

— syntagme prépositionnel, ex. *plus **à propos***

— syntagme nominal, ex. *c'est plus **son rôle***

— verbe, ex. *il a tellement **travaillé** (qu'il dort debout)* [5]

— adverbe, ex. *plus loin*

— proposition subordonnée, ex. *il vient plus **parce que tu lui demandes** (que pour l'argent)* [6]

La deuxième partie de la comparaison peut être constituée des éléments suivants :

— syntagme nominal (*N2*), ex. *c'est plus son rôle que **le mien***

— syntagme prépositionnel (*PP*), ex. *moins pour la clarté que **pour l'audience***

— proposition subordonnée (*CS*), ex. *il vient plus parce que tu lui demandes que **parce qu'il en a envie***

— phrase (*S*), ex. *il a tellement travaillé qu'**il dort debout***

— adverbe, ex. *mieux vaut tard que **jamais***

— adjectif, ex. *plus joli que **pratique***

En FMed, on trouve aussi cette construction, avec une deuxième partie également optionnelle (BURIDANT (2000), §177).

En AF, on trouve aussi des formes synthétiques, issues du latin. Il s'agit du préfixe *-or* (ou de son variant *-ur*), qui est ajouté à un adjectif (cf. exemple 31). Cependant, ces formes "perdent leur degré comparatif ou superlatif pour être prises au degré positif" (BURIDANT (2000), §175). Nous ne proposons donc pas

5. Cet exemple est disponible sur *FRMG Wiki* : http://alpage.inria.fr/frmgwiki/sentences/il-tellement-travaill%C3%A9-quil-dort-debout.

6. Cet exemple est disponible sur *FRMG Wiki* : http://alpage.inria.fr/frmgwiki/sentences/il-vient-plus-parce-que-tu-lui-demandes-que-pour-largent.

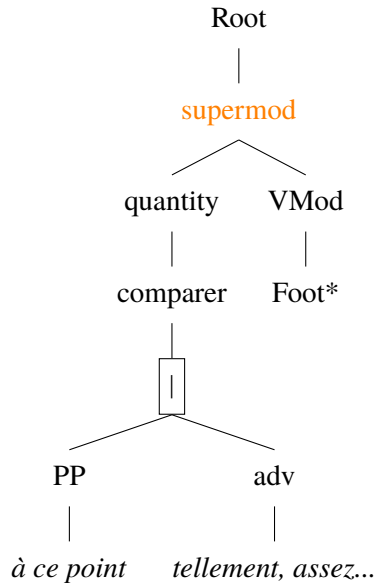


FIGURE 4.24 – Première partie de la comparaison (“|” symbolise une disjonction)

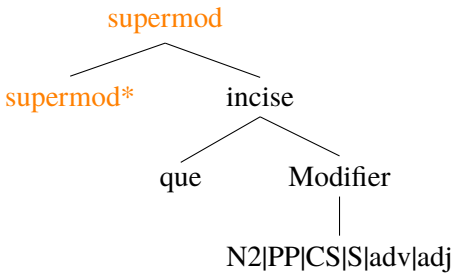


FIGURE 4.25 – Deuxième partie de la comparaison

d’analyser ces adjectifs comme des exemples de comparaison, à l’exception de *meilleur* et *pire*, qu’on a gardés en FC.

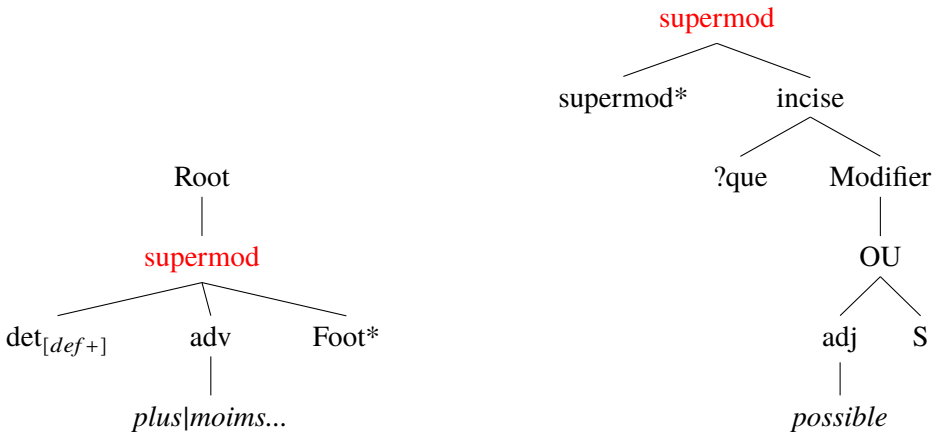
(31) Païen chevalchent par cez **greignurs** valees
NOMcom VERcjg PRE DETdem **ADJqua** NOMcom
‘Les païens chevauchent à travers les **immenses** vallées’
Chanson de Roland, v. 710 (ca. 1100), trad. BURIDANT (2000)

La structure du superlatif est semblable à celle du comparatif. La première partie (cf. figure 4.26a) peut s’adjoindre sur un adjectif ou un adverbe. Le déterminant est nécessairement défini. La deuxième partie est optionnelle, comme pour le comparatif. Elle peut être réalisée par l’adjectif *possible* (précédé ou non de *que*) ou une subordonnée.

4.4.3 Modifieurs de phrase

Les adverbes et les compléments prépositionnels

Les adverbes peuvent être à proximité du verbe (cf. partie 4.3.5), mais également à divers endroits dans la phrase, y compris aux extrémités de la phrase et “flottants” entre des constituants majeurs. On parle



(a) Première partie du superlatif

(b) Première partie du superlatif

FIGURE 4.26 – Arbres de superlatif

alors d'adverbes de phrase. *FRMG* prévoit de multiples points d'ancrages pour les modifieurs, afin de pouvoir les adjoindre à la tête de phrase directement, et non le long de la dorsale verbale.

(32) Mult **gentement** li emperere chevalchet
ADVgen **ADVgen** DETdef NOMcom VERcjg
'Très doucement, l'empereur allait à cheval'
Chanson de Roland, v. 3 120 (ca. 1100)

(33) A ces deux venoit le royaulme **justement** et **loyaulment**.
PRE DETdem ADJcar VERcjg DETdef NOMcom **ADVgen** CONcoo **ADVgen**
'Le royaume venait à eux deux de manière juste et loyale.'
Mémoires de Philippe de Commynes, p. 181 (ca. 1490–1505)

Les compléments prépositionnels peuvent aussi s'insérer à divers endroits de la phrase. Il peut s'agir d'une invocation (ex. *par Dieu, par mun chef*). Ces modifieurs peuvent aussi être des modifieurs de temps, de manière, de lieu etc.

(34) Danz Alexis an Alsis la citét sert sun seinur **par**
NOMcom NOMpro PRE NOMpro DETdef NOMcom VERcjg DETpos NOMcom **PRE**
bone volentét
ADJqua NOMcom
'Dan Alexis, dans la cité d'Alsis, sert son seigneur de bonne volonté.'
Vie de Saint Alexis, p. 159 (ca. 1050)

Les syntagmes nominaux

Certains syntagmes nominaux peuvent constituer des modifieurs de phrase sans être introduits par une préposition. Ils peuvent s'ancrer à divers endroits dans la phrase. Comme en français contemporain, on compte les compléments de temps (ex. *le matin, le lendemain*) et les inaliénables (ex. il avance *dos courbé*). Ces descriptions sont déjà présentes dans *FRMG*. Nous ajoutons les équipements dans la catégorie des inaliénables. Même si leur sens n'est pas exactement celui d'une entité inaliénable, leur comportement syntaxique est similaire. En voici quelques exemples :

(35) Ki la veit **le matin**, ja le jor n' iert
PROrel PROper VERcjg **DETdef NOMcom** ADVgen DETdef NOMcom ADVneg VERcjg
vencuz en bataille ne en nule afere.
VERppe PRE NOMcom CONcoo DETind NOMcom
'[Celui] qui la voit le matin, il n'est jamais vaincu à la bataille ou en aucune affaire le jour.'
Lapidaire en prose, p.108 (milieu 12e s.)

(36) il li acorurent **les braz tenduz**
PROper PROper VERcjg **DETdef NOMcom VERppe**
'ils accoururent vers lui, les bras tendus'
Queste del saint Graal, p. 166b (ca. 1225 ou 1230)

(37) **L' espee nue** an la loge entre
DETdef NOMcom ADJqua PRE DETdef NOMcom VERcjg
'L'épée nue, il entre dans la loge.'
Tristan de Béroul, v. 1987 (fin 12e s.)

De telles analyses ne sont possibles que si ces noms communs ont un trait sémantique renseigné par le lexique. Nous décrivons la tâche d'extraction des expressions et d'enrichissement du lexique dans la partie 5.2.3.

4.5 La coordination

4.5.1 Coordination et alternative

Les conjonctions de coordination du FMed sont un peu différentes de celle du FC (Hasenohr et Raynaud de Lage 1993). Tandis que celles du FC ne présentent pas d'ambiguïté, certaines conjonctions du FMed peuvent, d'une part, être confondues avec des homonymes. Par exemple, *ne* est une conjonction de coordination (*ni*) et un "clitique" négatif. Par ailleurs, certaines formes peuvent être utilisées comme conjonctions de subordination (*car*) ou adverbe (*car* en appui de l'impératif et *mais*). Leur emploi peut être plus souple, comme celui de *ne*, qui peut coordonner deux propositions ou termes négatifs, mais aussi une proposition positive et une proposition négative. Nous ne gardons pas toutes les conjonctions de coordination du FC, notamment celles qui contiennent plusieurs tokens, comme *et alors*, mais nous ajoutons les graphies du FMed (*ne*) et *ou soit* (*soit* en FC).

La coordination fonctionne cependant comme en FC. Nous reprenons donc la description de *FRMG* (cf. figure 4.27), qui autorise un nombre illimité d'éléments coordonnés (cf. partie 3.2.1 pour l'opérateur étoile de Kleene), c'est-à-dire que *Seq* peut être répété autant de fois que nécessaire (ou ne pas être réalisé) avant de passer à la dernière partie de la coordination, *Seqlast*. Cela permet notamment d'analyser les énumérations, qui peuvent être très longues en MF, comme dans cet extrait de *Manières de langage* (1399, p. 50) :

Et maintenant il fara bon d'ensaigner les enfans a conpter. Pour ce comencez ainsi enpreuf : un, deux, trois, quatre, cinq, six, sept, uuit, neuf, dis, onse, dous, tresze, quatorse, quinse, sesze, dis et sept, dis et uuyt, dis et neuf, vint, vint et un, vint et deux, vint et trois, vint et quatre, vint et cinq, vint et six, vint et sept, vint et uuyt, vint et neuf, trente, trente et un, trente et deux, et cætera.

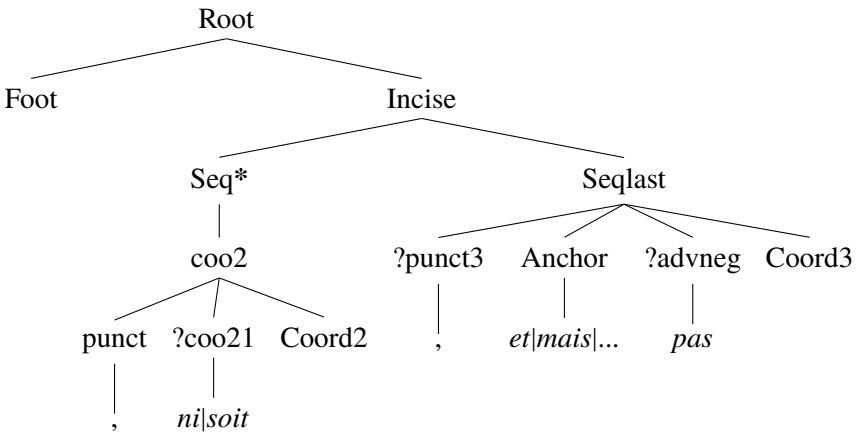


FIGURE 4.27 – La coordination dans *FRMG*

FRMG prévoit donc aussi les constructions en deux parties (ex. *ni X ni Y*, *soit X soit Y*). Si *coo21* est réalisé, la deuxième partie de la coordination doit l'être aussi. Ces constructions sont présentes dès l'AF (AMIOT et al. (2020), p. 944–952).

4.5.2 La polysyndète

L'emploi de conjonctions de coordination en tête de phrase est fréquent en FMed, notamment avec *et*, qui rend les échanges plus vifs en situation de dialogue (ex. 38). En FC, la polysyndète apparaît surtout dans la poésie.

ex. "Et l'unique cordeau des trompettes marines", *Chantre*, Guillaume Apollinaire

(38) Et vos, sire, fait il au roi, metez terme de
CONcoo PROper NOMcom VERcjg PROper PRE.DETdef NOMcom VERcjg NOMcom PRE
la bataille
DETdef NOMcom
'Et vous, sire, dit-il au roi, mettez un terme à la bataille.'
Eneas, v. 7 767 (ca. 1155)

Comme en FC, le gouverneur de la conjonction de coordination est le verbe, dans cette construction particulière. Nous reprenons la description de *FRMG* (cf. figure 4.28), qui en fait une adjonction sur le nœud de phrase, et la contraint pour apparaître au début (*start*).

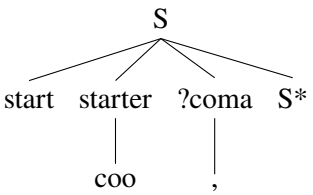


FIGURE 4.28 – La polysyndète dans *FRMG*

4.6 Les propositions subordonnées

4.6.1 Les propositions subordonnées complétives et infinitives

Dans *FRMG*, les propositions subordonnées complétives et infinitives sont décrites sous *VMod* (cf. figure 4.29). *SArg* peut être ancré par un verbe conjugué ou à l'infinitif.

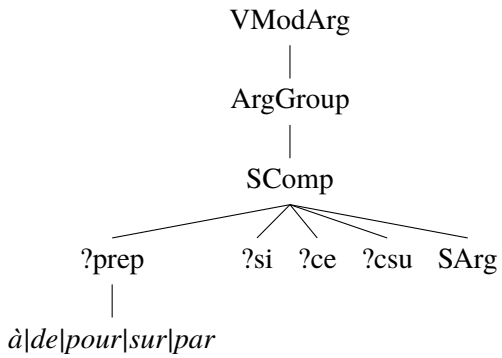


FIGURE 4.29 – La complétive objet dans *FRMG*

Cette structure peut être la réalisation du sujet phrastique, de la complétive objet (direct ou indirect) et de l'attribut du sujet. Ces propositions peuvent être précédées d'une séquence composée d'une préposition et parfois du pronom *ce* (ex. *demander* ***à ce que*** + *SArg*), mais ce n'est pas obligatoire. Elles peuvent être des arguments du verbe, mais aussi des arguments de noms et d'adjectifs (cf. exemple 39). C'est le lexique qui décrit les cadres de valence des entrées et précise les réalisations possibles des arguments.

(39) Le prince ou le juge juge chelui qu' il a
DETdef NOMcom CONcoo DETdef NOMcom VERcjg PROdem PROrel PROper VERcjg
condempné **indigne** d' **estre** pugni par sa main
VERppe **ADJqua** PRE **VERinf** VERppe PRE DETdef NOMcom
'Le prince ou le juge le juge indigne d'être puni par sa main'
De la erudition de Jean Daudin, p. 184 (1360)

En FMed, le subordonnant est fréquemment omis (cf. exemple 40). Cette structure est aussi présente en FC avec des verbes "recteurs faibles" comme *croire* et *penser* dans un registre particulier (FERREIRA (2019), ex. *je crois je vais partir*). Ces deux constructions ne semblent pas liées, mais nous utilisons la description dans *FRMG*, qui rend le subordonnant optionnel.

(40) Qar bien sevent Tristran s' en vet
CONcoo ADVgen VERcjg NOMpro PROper PROper VERcjg
'Car ils savent bien [que] Tristan s'en va'
Tristan de Béroul, v. 1 123 (fin 12e s.)

4.6.2 Les propositions subordonnées relatives

On trouve en FMed les subordonnants simples *qui, que, cui, dont, ou* et les subordonnants que BURIDANT (2000) qualifie de "lourds" : *li quels, le quel*. La forme contemporaine lie le déterminant au pronom relatif. Les deux composés suivent la flexion de leur catégorie (déterminant et pronom).

En FMed, la proposition relative peut s'adjoindre sur le pronom personnel sujet, comme dans l'exemple 23. Cette construction est donc ajoutée à la métagrammaire.

En AF, le pronom relatif n'est pas nécessairement exprimé (Buridant (2000), §475) :

— dans l'expression de l'hyperbole

(41) Jamais ne verés home plus volentiers le face
ADVgen ADVneg VERcjg NOMcom ADVgen ADVgen PROper VERcjg
'Jamais vous ne verrez d'homme qui le fasse aussi spontanément'
Aiol, v. 10 194, cité par Buridant (2000), §475

— dans la relative déterminative

(42) Al seignur alerent retrere / Le cunseil lur aveit doné
PRE.DETdef NOMcom VERcjg VERinf / DETdef NOcom PROper VERcjg VERppe
'Ils allèrent rapporter à leur maître le conseil qu'il leur avait donné'
MFce, Fables, v. 15–16, cité par Buridant (2000), §475

Au 13e siècle, on trouve des hypercorrections pour *qui* et *qu'il* : l'un est écrit à la place de l'autre. C'est le lexique qui documente cette variation. On observe également le recours plus fréquent au subjonctif.

La relative explicative

Selon Hasenohr et Raynaud de Lage (1993), ce qui différencie la relative explicative de la relative déterminative est son lien lâche avec son antécédent. C'est une apposition, les modes des verbes sont ceux des propositions indépendantes. Dans *FRMG*, les relatives apparaissent à proximité du nom qu'elles modifient. Cependant, il existe un arbre pour les relatives détachées, en relation avec le sujet de la phrase et introduites par un pronom sujet, comme dans la figure 4.30, qui est un exemple extrait de *La Porte étroite* d'André Gide [7].

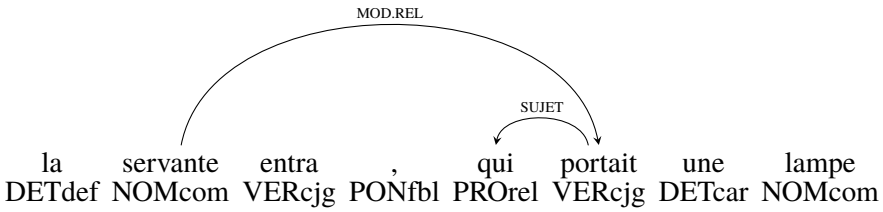


Figure 4.30 – Exemple de relative détachée

En FMed, on trouve des relatives détachées introduites par *que*, mais il s'agit d'une graphie alternative de *qui* (cf. exemple 43). Dans le *SRCMF*, on ne trouve pas de détachement d'autres types de relatives : nous n'avons donc pas besoin d'arbres supplémentaires pour attacher la relative à un syntagme nominal occupant une autre fonction.

(43) D' enz de sale uns **veltres** avalat, **Que** vint a Carles
PRE ADVgen PRE NOMcom DETcar **NOMcom** VERcjg **PROrel** VERcjg PRE NOMpro
lé galops e les salz
DETdef NOMcom CONcoo DETdef NOMcom
'Du fond de la salle dévale un lévrier ; qui court vers Charles au galop et par bonds'
Chanson de Roland, v. 730 (ca. 1100), trad. Joseph Bédier

7. Cet exemple est disponible sur *FRMG Wiki* : http://alpage.inria.fr/frmgwiki/wiki/jeudi-7-janvier-2016.

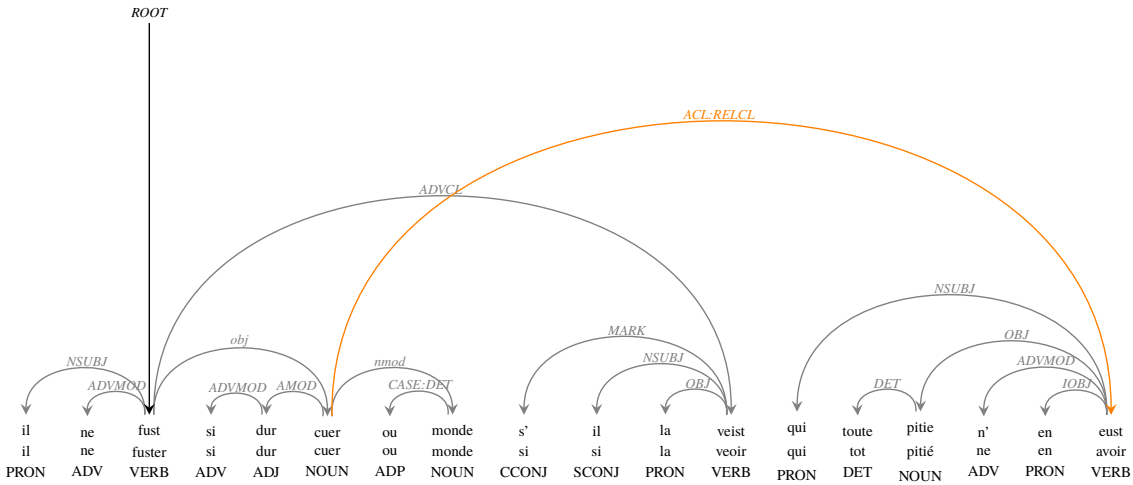


FIGURE 4.31 – "il ne fut pas de cœur si dur, ou de personne, qui n'en eût pitié s'il la vit", *Queste del saint Graal*, p. 165 (ca. 1225 ou 1230)

Cependant, plus longue est la distance entre le gouverneur syntaxique et la relative, plus l'analyse semble difficile à obtenir. Par exemple, la grammaire peut fournir l'analyse *UD* suivante (figure 4.31, en sélectionnant toutefois la mauvaise interprétation de *fust*), mais elle n'est pas prioritaire dans les sorties proposées.

Le complément déterminatif

BURIDANT (2000) appelle "complément déterminatif" le pronom *cui* lorsqu'il introduit une relative, précédé d'un déterminant défini.

(44) Artus li boens rois de Bretaingne / **La cui** proesce nos
NOMpro DETdef ADJqua NOMcom PRE NOMpro / **DETdef PROrel** NOMcom PROper
enseigne / Que nos solens preu et cortois
VERcjg / CONsub PROper VERcjg ADJqua CONcoo ADJqua
'Le noble roi Arthur de Bretagne, la [prouesse] de qui (=dont la prouesse) nous enseigne à être vaillants et courtois'
Yvain de Chrétien de Troyes, v. 1–3 (1177–1181), trad. David Hult

On ne trouve que trois occurrences de cette construction dans la *BFM*, dans les sous-corpus des 12e et 13e siècles. Une entrée est créée dans le lexique pour analyser ce pronom relatif composé.

4.6.3 Les propositions subordonnées circonstancielles

Les propositions circonstancielles en FMed s'insèrent sur le nœud de phrase, comme en FC. En revanche, les relateurs sont différents, et certaines locutions demandent des descriptions supplémentaires. Certaines conjonctions de subordinations sont en un seul bloc, comme en FC. C'est le cas, notamment, de *si*, *que* et *com* (*comme*). Le lexique *Lefff* renseigne ce type de conjonction de la même manière que des locutions comme *parce que* et *aussi longtemps que*, qui sont figées en FC. *FRMG* peut donc les utiliser comme réalisation de *csu* (conjonction de subordination) (cf. figure 4.32).

En FMed, en revanche, les conjonctions de subordination complexes ne sont pas encore figées, et il convient de les décrire dans la grammaire. Il est en effet possible de trouver des tmèses, c'est-à-dire

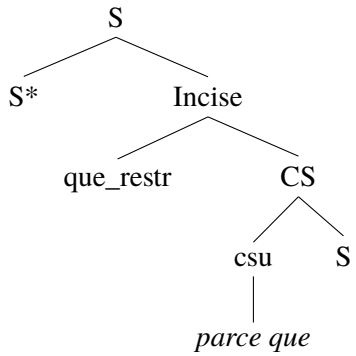
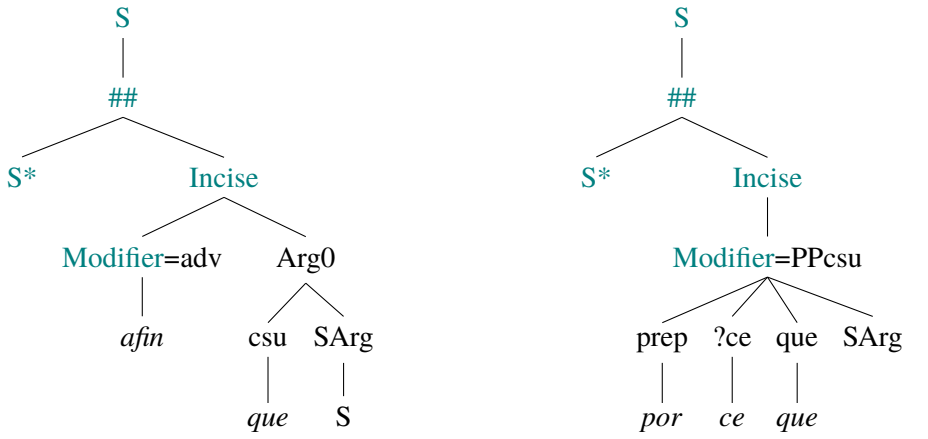


FIGURE 4.32 – Proposition subordonnée circonstancielle dans *FRMG*

des éléments qui s'insèrent entre les composants de la conjonction, ce qui empêche l'ajout d'entrées lexicales similaires à celles du français contemporain (locutions figées). On décrit deux structures dans la métagrammaire : *préposition + conjonction* (cf. figure 4.33a) et *adverbe + conjonction* (cf. figure 4.33b). On en trouve un inventaire chez AMIOT et al. (2020) (p. 909–911).



(a) Circonstancielle *adverbe + conjonction*, ex. *afin que*

(b) Circonstancielle *préposition + conjonction*, ex. *por (ce) que*

FIGURE 4.33 – Arbres de propositions subordonnées circonstancielles avec des conjonctions complexes

Dans l'arbre 4.33b, le pronom démonstratif *ce* est optionnel. Par exemple, il n'est pas exprimé dans l'exemple 45, mais il l'est dans 46.

(45) Oï ot feire menssion Del roi Artus [...] **Par qu'**
VERppe VERcjg VERinf NOMcom PRE.DETdef NOMcom NOMpro [...] **PRE CONsub**
estoit dotee sa corz Et renomee par le monde.
VERcjg VERppe DETpos NOMcom CONcoo VERppe PRE DETdef NOMcom
'Il avait entendu mentionner le roi Arthur [...] parce que sa cour était respectée et renommée de par le monde.'
Cligès de Chrétien de Troyes, v. 70 (1176), trad. GINGRAS (2004)

(46) Marie Magdalene, cui mult de pechiét furent pardonét **par ce ke**
NOMpro NOMpro PROrel ADVgen PRE NOMcom VERcjg VERppe PRE PROdem CONsub
ele amat mult
PROper VERcjg ADVgen
'Marie-Madeleine, à qui beaucoup de péchés furent pardonnés parce qu'elle a beaucoup aimé'

Li sermon saint Bernart sor les Cantikes, p. 130 (fin 12e s.)

Peu à peu, ces conjonctions de subordination se figent. Nous conservons cependant la même analyse à travers les textes, en traitant le nouveau token *parce* comme une contraction de *par* et *ce*. Celle-ci est renseignée dans le lexique.

(47) **parce que** pluseurs de conseillers dudit Parlement avoient obtenu lettres de
CONsub PROind PRE NOMcom PRE.DETcom NOMpro VERcjg VERppe NOMcom PRE
gages à vie
NOMcom PRE NOMcom
‘parce que plusieurs conseillers du Parlement avaient obtenu des lettres de gages à vie’
Journal de Nicolas de Baye, p. 150 (1400–1417)

4.7 Les structures non canoniques ou “marquées”

4.7.1 Les propositions interrogatives

En FMed, l’interrogation simple totale (ex. 48) est marquée, à l’écrit, par un point d’interrogation et l’inversion simple du sujet, quand il est exprimé (Combettes, Marchello-Nizia et Prévost (2020), p. 1221–1222). On ne trouve presque pas l’ordre *SV* (Combettes, Marchello-Nizia et Prévost (2020), p. 1227). En revanche, il peut être présent dans la deuxième partie d’une alternative (ex. 49). L’interrogation partielle est, elle aussi, caractérisée par l’inversion du sujet (ex. 50).

(48) Vous aida nuls ?
PROper VERcjg PROind PONfrt
‘N’avez-vous pas été aidé de personne ?’
Remede de Fortune de Guillaume de Machaut, v. 3 726 (1341)

(49) Voldréz jehir ou voz voldréz combatre ?
VERcjg VERinf CONcoo PROper VERcjg VERinf PONfrt
‘Voudrez-vous avouer ou voudrez-vous combattre ?’
Ami et Amile, v. 774 (ca. 1200), trad. Combettes, Marchello-Nizia et Prévost (2020)

(50) Et de qui se doubte li rois ?
CONcoo PRE PROint PROper VERcjg PROdef NOMcom PONfrt
‘Et de qui doute le roi ?’
Chroniques de Jean Froissart, p. 82 (14e s.)

Dans l’interrogation complexe, le syntagme nominal en tête de phrase est repris par un pronom conjoint (ex. 51). Son origine n’est pas certaine (Combettes, Marchello-Nizia et Prévost (2020), p. 1224), mais elle se développe surtout au 14e siècle et devient majoritaire au 15e siècle pour l’interrogation simple. Le développement de l’interrogation complexe dans l’interrogation partielle est plus tardif, et il n’entre pas dans le cadre de notre étude.

(51) **Ta besongne** est **elle** bien faicte ?
DETdef NOMcom VERcjg **PROper** ADVgen VERppe PONfrt
‘Ton travail est-il bien fait ?’
La Farce de maître Pierre Pathelin, v. 1658 (1456–1460)

Dans le *SRCMF*, les interrogatives ne font pas l'objet d'une annotation particulière, mais uniquement de surface. Nous ne proposons donc pas d'arbre supplémentaire pour analyser les interrogations simples totales, dont l'ordre *VS* (et l'ordre *SV* des alternatives) peut être analysé par *MetaMOF* grâce à sa structure plus plate que celle de *FRMG*. En FC, cette inversion n'est possible qu'avec un sujet clitique.

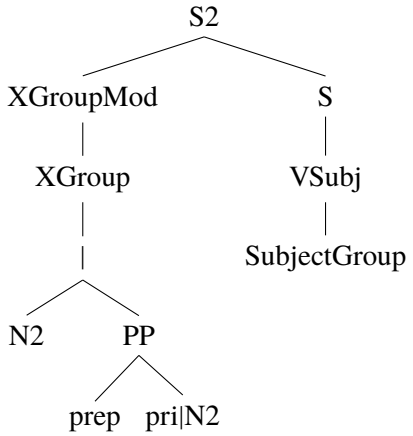


Figure 4.34 – La structure interrogative dans *FRMG*

Pour les autres cas d'interrogation, nous reprenons l'arbre dédié dans *FRMG* (cf. figure 4.34), qui regroupe les cas d'interrogation partielle, avec un pronom interrogatif (*pri*), précédé ou non d'une préposition, et les cas d'interrogation totale complexe, où la tête de *N2* n'est pas un pronom interrogatif.

Quant aux interrogations indirectes, elles sont introduites par un pronom interrogatif ou *se* (*si*) pour les interrogations totales. Elles sont introduites par un verbe ou un locution d'interrogation, de parole, de connaissance sensorielle ou intellectuelle (Combettes et Glikman (2020), p. 1357). Nous reprenons aussi les descriptions de *FRMG*, car ces structures s'apparentent à des complétives, comme en FC. L'ordre habituel dans ces propositions est le suivant : "terme interrogatif – sujet (s'il n'est pas l'objet de l'interrogation) – verbe" (Combettes et Glikman (2020), p. 1358), mais d'autres ordres sont possibles. Le mode majoritaire est l'indicatif, mais on trouve aussi des verbes au subjonctif, par imitation du latin.

4.7.2 La dislocation

On trouve des exemples de dislocation dès les premiers textes d'AF (Combettes, Marchello-Nizia et Prévost (2020), p. 1315–1320). Il s'agit du détachement d'un élément à gauche (ex. figure 4.35) ou à droite (ex. figure 4.36) de la phrase. Il est repris sous forme de pronom dans la proposition régissante (celle qui régit la dislocation). Cette construction connaît une évolution au cours de la période. Les premiers cas concernent les syntagmes nominaux et les relatives sans antécédents. La dislocation de pronoms est rare jusqu'au 14e siècle, et celle des syntagmes prépositionnels est, elle aussi, tardive. Combettes, Marchello-Nizia et Prévost (2020) expliquent cette évolution par la nature argumentative de ces constructions, qui est caractéristique des textes plus tardifs.

FRMG possède déjà ces descriptions (cf. figures 4.37). L'élément détaché est un syntagme nominal, ancré par un nom ou un pronom, et celui-ci peut être modifié. Les adjectifs détachés ne sont pas analysés de cette manière, mais comme simple adjonction sur la phrase. Cependant, nous n'observons pas d'adjectif détaché en corpus pour le FMed, et il n'en est pas question dans les grammaires, comme dans celle de Buridant (2000). On peut donc trouver un pronom suivi d'une proposition relative en position

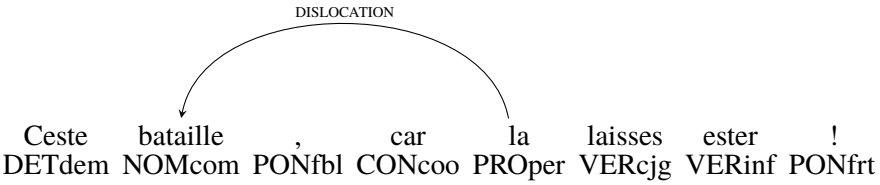


Figure 4.35 – Exemple de dislocation à gauche, extrait de *Chanson de Roland*, v. 3 902 (ca. 1100)
“Ce duel, renonces-y !”, trad. Combettes, Marchello-Nizia et Prévost (2020)

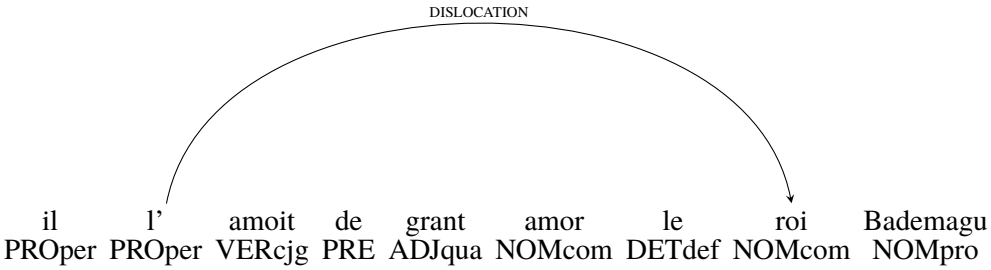


Figure 4.36 – Exemple de dislocation à droite, extrait de *Queste del saint Graal*, p. 222 (ca. 1225 ou 1230)
“car il l’aimait d’une profonde affection, le roi Bademagus”, trad. Combettes, Marchello-Nizia et Prévost (2020)

détachée (ex. ***Celui qui*** *est bon [...],* ***il*** *aura la terre promise*). La dislocation à droite a une contrainte supplémentaire : seuls les référents nominatifs (*dcln*) et accusatifs (*dcla*) sont autorisés dans la proposition régissante.

(a) Dislocation à gauche

(b) Dislocation à droite (avec contrainte sur le référent dans la proposition régissante)

Figure 4.37 – Dislocation dans *FRMG*

Nous gardons ces descriptions, car il semble que les éléments détachés à droite suivent les mêmes contraintes en AF, comme on peut le voir dans le *SRCMF*. On ajoute cependant la possibilité que l’élément détaché soit une proposition relative sans antécédent (cf. figures 4.38). L’élément détaché à droite est également repris par un pronom nominatif ou accusatif (cf. exemple 52).

(52) ne jo mie **nel** sai, **Liquels** d’ els dous en
ADVneg PROper ADVneg ADVneg.**PROper** VERcjg **PROrel** PRE PROper ADJcar PROadv
fut li plus isnels
VERcjg DETdef ADVgen ADJqua
‘et je ne le sais, lequel d’eux deux en fut le plus rapide’
Chanson de Roland, v. 1 387 (ca. 1100)

(a) Dislocation à gauche

(b) Dislocation à droite (avec contrainte sur le référent dans la proposition régissante)

FIGURE 4.38 – Dislocation d'une relative sans antécédent dans *MetaMOF*

Il existe aussi une construction en *sujet* + *si* + *verbe* (COMBETTES, MARCHELLO-NIZIA et PRÉVOST (2020), p. 1320–1321, cf. exemple 53), mais nous n'ajoutons pas d'arbre supplémentaire pour l'analyser. L'adverbe *si* est considéré comme un simple modifieur du verbe.

(53) **Li tierz si est** mult renumé, Watlingestrate est apelé
DETdef ADJqua ADVgn VERcjg ADVgen VERppe NOMpro VERcjg VERppe
'La troisième est renommée, [elle] est appelée Route de Watling'
Description d'Engleterre, v. 241–242 (ca. 1139), trad. COMBETTES, MARCHELLO-NIZIA et PRÉVOST (2020)

4.7.3 Les propositions clivées

On trouve des constructions clivées dès les premiers textes d'AF (ROUQUIER 2014). Comme en FC, elles sont de forme *C'est X Qu-* + *Verbe*. La première clause (*C'est X*) contient les trois premiers éléments et présente le focus (*X*), c'est-à-dire l'information nouvelle (COMBETTES, MARCHELLO-NIZIA et PRÉVOST (2020), p. 1309–1315). Toutes les fonctions sont clivables (et peuvent occuper *X*). La deuxième clause est *Qu-* + *Verbe*, soit l'information ancienne ou inférable. *Qu-* représente toutes les réalisations du pronom relatif.

En AF, le focus peut être un pronom sujet (ex. 54), contrairement au FC, qui n'accepte que le pronom régime (ex. *c'est **moi** qui...*). L'accord du verbe *estre* se fait alors avec le pronom sujet, et non le pronom démonstratif. Cette possibilité viendrait de la souplesse de l'ordre des mots. Cet emploi perdure jusqu'au 15e siècle.

(54) Ce sui **je** qui alai chacier
PROdem VERcjg **PROper** PROrel VERcjg VERinf
'C'est moi qui suis allé chasser'
Guingamor, v. 615 (fin 12e s.)

Cependant, la zone préverbale est de plus en plus occupée par le pronom sujet, avec la prévalence de l'ordre *SVO* et le recul de *S0*, ce qui a un impact sur cette structure. En effet, l'usage du pronom sujet en position de focus est peu à peu abandonnée, au profit du pronom régime (ex. 55).

(55) c' est luy qui determine de telz procès
PROdem VERcjg PROper PROrel VERcjg PRE ADJind NOMcom
'c'est lui qui détermine de tels procès'
Mémoires, de Philippe de Commynes, p. 32 (ca. 1490–1505)

En revanche, on ne trouve pas de structures où le pronom relatif est suivi par un syntagme nominal seulement (ex. *c'est lui le chef*). Cette construction est présente dans *FRMG*, elle doit donc être écartée.

Actuellement, la description des clivées, présente dans *FRMG*, n'a pas été reprise dans notre métagrammaire, à cause du nombre peu élevé des occurrences en FMed. Combettes, Marchello-Nizia et Prévost (2020) en recensent entre 41 et 42 occurrences sur tout leur corpus, ce qui est très peu. Nous ne disposons pas d'indices pour distinguer ces structures de phrases où le troisième élément est modifié par une relative. Étant donnés la confusion que l'implémentation de ce phénomène entrainerait et les progrès qu'il reste à faire (cf. chapitre 6), nous préférons écarter ces arbres pour le moment, et privilégier l'analyse avec une relative.

Les adaptations faites à partir de la métagrammaire *FRMG* sont de nature diverse. Certaines ont profondément modifié la structure de la grammaire, comme la disposition des constituants majeurs au même niveau. D'autres changements sont locaux, et ne consistent qu'à rendre certains éléments optionnels ou à changer les ancres lexicales. La plupart des phénomènes ont été conservés pour le FMed, mais certains étaient superflus, comme la description de références bibliographiques, ou inadaptés, comme le modifieur de nom non introduit. Il a aussi fallu ajouter quelques classes, par exemple pour le complément du nom non introduit.

5 Un lexique d'ancien français

Le lexique est une part importante de la chaîne de traitement, car il permet de constituer, avec la méta-grammaire, une grammaire d'arbres adjoints lexicalisée (ou *LTAG*, Schabes, A. Abeillé et Joshi (1988)). Les arbres élémentaires peuvent être décrits comme des projections des entrées du lexique. Pour produire une analyse syntaxique, le parseur utilise les diverses informations linguistiques des entrées correspondant aux mots de la phrase. Cet accès est rendu possible par leur organisation en structures de traits, qui permettent une grande finesse dans les descriptions. Elles ne peuvent cependant être validées que par des expériences d'analyse sur corpus.

Le lexique d'AF *OFrLex* (Sagot 2019) a été développé sur le modèle du lexique *Lefff* (Sagot 2010). Pour les besoins de l'analyse syntaxique, nous l'avons modifié et enrichi, en y ajoutant notamment des cadres de valence verbale et de nouvelles entrées.

Dans un premier temps, nous présentons le formalisme du lexique et son intégration à la chaîne de traitement. Nous dressons ensuite un état du lexique, et nous décrivons les modifications apportées. Enfin, nous donnons les premiers éléments d'évaluation et les cas d'usages possibles.

5.1 Formalismes et intégration à la chaîne de traitement

5.1.1 Chaîne de traitement

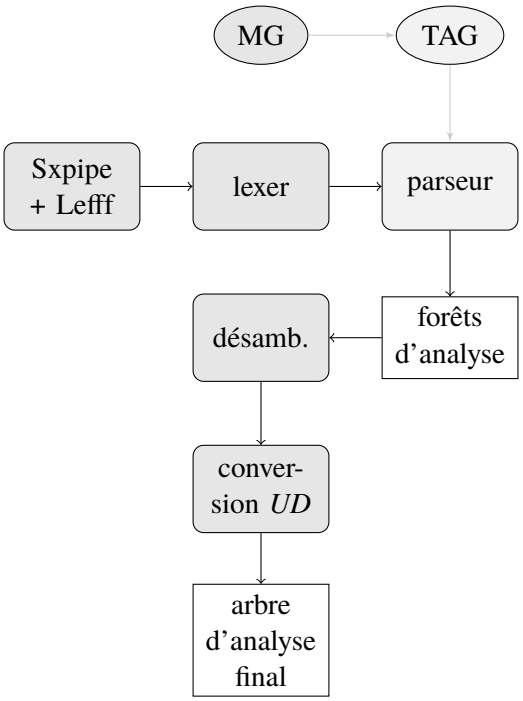


Figure 5.1 – Architecture de la chaîne de traitement développée par Boullier et al. (2005)

Nous procédons tout d'abord à quelques rappels sur la chaîne de traitement installée par *alpi* (cf. figure 5.1). Un texte à analyser est d'abord traité par le segmenteur *Sxpipe* qui procède à une segmentation en phrases et en tokens. Il fournit ainsi des phrases sous forme de 'treillis' de mots à partir des formes des entrées du lexique, afin de conserver les ambiguïtés lexicales pour fournir plus de possibilités au parseur, mais aussi de proposer des corrections (ou des normalisations) de la phrase d'entrée. Le *lexer* utilise ensuite cette sortie et va chercher le contenu des entrées dans le lexique pour les fournir en entrée au parseur. Un sous-ensemble d'arbres élémentaires est sélectionné dans la grammaire en fonction des entrées lexicales liées à la phrase. Il constitue la sous-grammaire utilisée pour faire une analyse syntaxique. Par exemple, si aucun token d'une phrase traitée ne peut être analysé comme un adverbe, les arbres ancrés par des adverbes ne seront pas consultés lors de l'analyse de celle-ci. Cela permet de rendre l'analyse plus efficace. Il est aussi possible de définir des contraintes plus fines sur l'activation d'un arbre, comme la distance entre certains types de mots (cf. fichier *block.db*). Par exemple, un nom commun employé comme adverbe de temps (ex. *ce matin*) ne peut être séparé que de quatre nœuds de sa tête syntaxique dans l'arbre dérivé.

5.1.2 Le formalisme *LTAG*

Dans une grammaire d'arbres adjoints lexicalisée (*LTAG*), les informations lexicales ont une place très importante (Candito 1999). Chaque arbre élémentaire est ancré par un mot du lexique. Le formalisme *TAG* n'a pas besoin d'adaptation pour être lexicalisé, car il l'est "naturellement" (Schabes, A. Abeillé et Joshi 1988). Comme il offre un domaine de localité étendu, il peut faire apparaître un prédicat lexicalisé et ses arguments dans une même structure, et ainsi fournir les mêmes arbres que la grammaire non lexicalisée. Lors de l'analyse syntaxique, les arbres élémentaires sont sélectionnés en fonction de la présence des mots de la phrase dans le lexique. La lexicalisation permet donc de filtrer les analyses possibles pour une phrase en fonction des éléments qui la constituent, ce qui revient à sélectionner une sous-grammaire, réduisant ainsi la complexité de l'analyse (Parmentier 2007).

Depuis les travaux de Vijay-Shanker (1987), les *LTAG* contiennent des structures de traits, qui permettent de faire des requêtes dans le lexique pour l'analyse syntaxique. On y trouve des traits morphologiques, comme le genre et le nombre, qui interviennent lorsque l'accord est nécessaire à l'unification de nœuds. Des traits syntaxiques donnent en outre des informations sur les arguments attendus. On trouve aussi des traits sémantiques, comme la caractère temporel ou inaliénable, qui peuvent également contraindre l'analyse syntaxique.

5.1.3 Les lexiques de la chaîne de traitement

La chaîne de traitement repose en grande partie sur le lexique, non seulement pour la tokenisation et l'analyse lexicale des corpus, mais aussi pour l'analyse syntaxique. Le lexique le plus utilisé est celui du FC, le *Lefff* (Sagot 2010). C'est un lexique morphologique et syntaxique du FC. Il a été constitué en plusieurs étapes, en partie automatiques. Il suit un modèle inflexionnel, qui propose des listes de formes fléchies rattachées à leur lemme. D'autres ressources ont été développées sur son modèle, comme *Leffe* pour l'espagnol (Molinero, Sagot et Nicolas 2009) et *DeLex* pour l'allemand (Sagot 2014). Le travail sur des langues morphologiquement plus riches a motivé la création du formalisme *Alexina* (Sagot 2018 ; Sagot et Walther 2013).

Constitution du *Lefff*

Dans sa première version, il s'agit d'un lexique morphologique qui propose un ensemble de flexions pour chaque lemme (Clément, Lang et Sagot 2004). Un module attribue un lemme à chaque forme des catégories de verbe, d'adjectif et de nom. Ces formes fléchies viennent de grammaires descriptives comme le *Bescherelle* et le *Grevisse*. Dans une deuxième version, des informations syntaxiques sont apportées, comme les cadres de sous-catégorisation et des cadres de contrôle ou d'utilisation de l'impersonnel (Sagot, Clément et al. 2006). Les auteurs préfèrent procéder à ces enrichissements de manière automatique, du fait du nombre important d'entrées et de la complexité de leurs structures.

Les informations syntaxiques du *Lefff* viennent essentiellement de sources externes, notamment des tables du *Lexique-grammaire* de Maurice Gross (Sagot et Danlos 2008). Ces tables (Gross 1975) regroupent des verbes, des adjectifs et des noms prédicatifs, et elles indiquent leur sous-catégorisation. Elles donnent ainsi le nombre d'arguments attendus, leur caractère optionnel, ainsi que la nature de leur réalisation, et si celle-ci est précédée d'une préposition parmi les suivantes : *à, de* (objets seconds), *avec* (constructions symétriques) et *Loc* (ensemble de prépositions locatives). Les fonctions des arguments et leurs réalisations possibles sont renseignées séparément, comme c'est le cas dans *Synlex* (Gardent et al. 2006). Le travail sur le *Lexique-grammaire* a par ailleurs permis d'affiner la représentation des adverbes en *-ment* (Sagot et Fort 2007).

Le *Lefff* bénéficie en outre des informations contenues dans les ressources suivantes : *Multext* (Veronis 1998), *Dicovalence* (Van den Eynde et Mertens 2010), *Synlex*, *DiCo* (Polguère 2003) et *DicoLPL* (Van Rullen et al. 2005). La constitution des entrées d'un lexique est décrite dans la littérature comme étant "difficile", car ces différentes sources présentent parfois des informations contradictoires qu'il faut harmoniser (Danlos et Sagot 2008). La détection des emplois impersonnels est faite grâce à l'outil *ILIMP* (Danlos 2005). Les expressions verbales figées sont décrites grâce aux travaux de Danlos, Sagot et Salmon-Alt (2006).

Le lexique acquiert en couverture et en qualité à mesure qu'il est utilisé pour l'analyse de corpus, grâce à des méthodes de fouille d'erreurs. Le module *Sxpipe* suggère des analyses pour les mots inconnus en les rapprochant d'entrées proches (Sagot, Clément et al. 2006). Cette étape permet d'enrichir le *Lefff* de nouvelles entrées. En cas d'échec de l'analyse syntaxique, un 'suspect' est recherché : il s'agit de la forme qui est le plus probablement la cause de l'absence d'analyse (Sagot et Villemonte de la Clergerie 2008 ; Tolone, Sagot et Villemonte de la Clergerie 2012). La méthode de Nicolas, Farré et Villemonte de la Clergerie (2007) permet d'obtenir des corrections pertinentes, en transformant les mots suspects en mots inconnus et en examinant ce que propose l'analyseur au niveau de ces mots. La fouille d'erreur en sortie de parseur est particulièrement utile pour traiter des corpus spécialisés, dont le vocabulaire spécifique n'a pas été renseigné. Les erreurs peuvent également venir d'entrées existantes. Par exemple, la catégorie morpho-syntaxique ou les traits morphologiques et sémantiques peuvent être erronés. La confrontation de cette ressource avec l'usage linguistique permet en outre de définir des poids pour les entrées en fonction de leur distribution dans les corpus.

Contenu

Le *Lefff* est une ressource morphologique et syntaxique à large couverture, i.e. permettant d'analyser la plupart des formes rencontrées en corpus. Il s'agit d'un lexique extensionnel. Les lemmes recensés sont donc accompagnés d'une liste de formes fléchies possibles et de leurs traits morphologiques, syntaxiques

et sémantiques. Ces formes constituent les entrées du lexique, qui contiennent chacune leur structure de traits. Par exemple, l'entrée "fréquent" en tant qu'adjectif en emploi impersonnel, est écrite sous cette forme :

fréquent 100 adj [pred="fréquent_____1 <Suj:(de-sinf|scompl|sn), Loc:(loc-sn)>", @impers,cat=adj,@ms] fréquent_____1 Default ms %adj_impersonnel adj-4

Voici les informations qu'elle contient :

— poids indicatif (ex. *100*, valeur par défaut)

— partie du discours (ex. *adj*)

— lemme (ex. *fréquent_____1*)

— sous-catégorisation (le cas échéant) selon le cadre théorique des *LTAG*

  — *Arg0* : sujet du verbe en emploi canonique ou gouverneur d'un adjectif (*Suj*)

  — *Arg1* : objet direct ou attribut du sujet du verbe en emploi canonique (absent de cet exemple)

  — *Arg2* : objet indirect ou complément oblique

  — *Loc*, *Dloc, Obl, Obl2* : arguments non-essentiels (locatif, "délocatif" et obliques)

  Pour chacun des arguments attendus, une liste des réalisations syntaxiques est fournie. Le *Lefff* suit l'"Approche Pronominale" (Van den Eynde et Blanche-Benveniste 1978), comme le *Dicovalence*. Il s'agit de "l'étude de la valence à partir des paradigmes de pronoms qu'accepte le verbe" (Van den Eynde et Mertens 2003). Les fonctions sont donc associées à leur paradigme de pronoms (Danlos et Sagot 2008). En cas d'emploi personnel, c'est le pronom personnel sujet (ou clitique nominatif, abrégé *cln*) qui est renseigné. S'il s'agit d'un emploi impersonnel, l'argument est sémantiquement vide, donc aucun pronom clitique n'est attribué à *Arg0*.
  Les réalisations proposées dans l'exemple ci-dessus sont les suivantes :

  — *sn* : syntagme nominal

  — *scompl* : complétive (sujet)

  — *de-sinf* : infinitive précédée de la préposition *de*

  — *loc-sn* : complément oblique précédé d'une préposition locative

— trait morpho-syntaxique sous forme de macro [1] (ex. *@impers* pour l'emploi impersonnel, c'est-à-dire que le gouverneur syntaxique de l'entrée est un pronom impersonnel)

— catégorie grammaticale (ex. *ms* pour masculin singulier)

— classe inflexionnelle qui indique selon quelle table du lexique le lemme doit être transformé pour produire ses formes fléchies (ex. *adj-4*)

Le *Lefff* répartit ses entrées en différents fichiers selon leur partie du discours. Quelques fichiers s'y ajoutent, comme celui des formes contractées, décrites dans un fichier *amlgm* en suivant ce format : *forme = entrée 1 + entrée 2*.

1. Dans le lexique, on trouve un certain nombre de macros. Elles permettent de déterminer un ensemble d'entrées qui recevront un traitement particulier. Par exemple, les participes passés marqués par la macro de la voix active sont analysés avec leur auxiliaire de temps, et non l'auxiliaire du passif.

```
ex. au = à__prep + le__det
```

Les entrées *1* et *2* sont créées spécialement pour ce fichier. Il s'agit de formes fléchies auxquelles on ajoute la partie du discours en suffixe pour qu'elles ne soient appelées que par ce fichier, ces formes spéciales n'étant, jusqu'à présent, jamais observées en corpus. Ce fichier, ainsi que ceux des classes fermées (déterminants, pronoms, interjections, prépositions, conjonctions de coordination et de subordination, ponctuation) sont renseignés manuellement (Sagot et Danlos 2008).

Les schémas d'entités nommées comme les dates ou les adresses sont renseignées dans un fichier séparé. Ces tokens ou chaînes de tokens sont détectés par le segmenteur et renseignés par le lexique.

5.2 OFrLex : premier état et modifications

5.2.1 *OFrLex*, lexique d'ancien français

OFrLex est conçu pour être un lexique d'AF (Sagot 2019). Son modèle est globalement celui du *Lefff* (Guibon et Sagot 2020).

Constitution du lexique

Ce lexique d'AF est constitué à partir des ressources numériques suivantes :

- *FROLEX* (Heiden, Magué et Pincemin 2010), un lexique généré à partir des textes de la *BFM* et du *NCA* et du *Dictionnaire du Moyen Français* (*DMF*, R. Martin, Bazin et Cromer (2012)) et qui utilise le jeu d'étiquettes morpho-syntaxiques *Cattex*
- dictionnaires numérisés :
  - le dictionnaire de Tobler et Lommatzsch, *Altfranzösisches Wörterbuch*, dans des éditions produites et distribuées par Achim Stein [2]
  - le *Lexique de l'ancien français* de Godefroy [3]
  - le *Dictionnaire Electronique de Chrétien de Troyes*, distribué par le CNRTL
- *Wiktionary*
- *Tableaux de conjugaison de l'ancien français* [4] (*TCAF*) de Machio Okada et Hitoshi Ogurisu

Les entrées de ces ressources ont tout d'abord été structurées, puis fusionnées en un ensemble cohérent. Les informations linguistiques ainsi extraites ont ensuite été adaptées au formalisme *Alexina*. Des modules morphologiques créés pour l'AF ont permis de générer les formes fléchies des lemmes rassemblés. Toutes les informations morphologiques attribuées n'ont pas fait l'objet d'un contrôle, faute de ressource adéquate, et les cadres valenciels sont limités à ce qui est disponible dans les dictionnaires, c'est-à-dire la distinction entre verbes transitifs et intransitifs.

Les ressources lexicales pour le FMed sont donc très différentes de celles qui ont permis de constituer le *Lefff*. En nombre plus limité, les dictionnaires disponibles proposent rarement des entrées structurées

2. Le dictionnaire de Tobler et Lommatzsch est disponible à cette adresse : https://www.ling.uni-stuttgart.de/institut/ilr/toblerlommatzsch/downloads.htm.
3. Le *Lexique de l'ancien français* de Godefroy est distribué sur Wikisource : https://fr.wikisource.org/wiki/Lexique_de_l%E2%80%99ancien_fran%C3%A7ais.
4. Les tables du *TCAF* sont consultables en ligne : http://www.micmap.org/dicfro/introduction/tableaux-de-conjugaison.

et ils ne sont pas le produit d'un travail linguistique exhaustif pour décrire les aspects morphologiques, à l'exception de *FROLEX*, dont l'exploitation demande toutefois des adaptations (Sagot 2019).

Premier état du lexique

Le lexique contient un fichier par partie du discours, comme dans le *Lefff*, ainsi qu'un fichier de contractions. Leur syntaxe est généralement identique, à l'exception de la notation des catégories grammaticales. Les textes de FMed que nous voulons annoter ayant une ponctuation modernisée, le fichier de ponctuation reste donc identique à celui du *Lefff*.

Le premier état du lexique couvrait la plupart des mots du corpus *SRCMF* et de la *BFM*, notamment dans les classes dites "ouvertes" (noms, verbes, adjectifs, adverbes). Cependant, ce lexique n'était pas exploitable dans sa forme initiale, car il ne permettait pas l'analyse de phrases par le parseur. D'une part, il manquait un certain nombre d'entrées essentielles dans les classes fermées (pronoms, déterminants, prépositions, conjonctions...), mais c'étaient surtout les informations syntaxiques qui devaient être enrichies et/ou corrigées. Une refonte manuelle des classes fermées du lexique a donc semblé impérative. D'autre part, l'ajout des valences verbales est également très important, car elles sont nécessaires à l'attachement des arguments du verbe. Initialement, seule la distinction entre verbes transitifs et intransitifs était faite. Les verbes n'acceptaient que des sujets nominaux et pronominaux, et les verbes présumés transitifs attendaient un objet direct nominal ou pronominal (Guibon et Sagot 2020). Sans informations plus riches, de nombreuses phrases ne peuvent pas être analysées, comme celles dont le verbe appelle une complétive, une infinitive, ou un attribut du sujet. Le sujet peut aussi avoir d'autres réalisations.

Prévoir l'intégration à une base de textes

Dans l'optique d'annoter le corpus *Profiterole*, dont les textes enrichis linguistiquement seront intégrés à la *BFM*, nous souhaitons intégrer les étiquettes morpho-syntaxiques *Cattex* au lexique. Les étiquettes *Cattex* étant plus informatives pour les classes fermées que celles d'*Universal Dependencies* et, dans une moindre mesure, que celles d'*OFrLex*, nous avons ajouté des traits pour les reconstruire. Ces annotations seront récupérées en sortie de parseur.

Notre objectif premier est de parvenir à une analyse de qualité du corpus *Profiterole*. Nous avons donc commencé par améliorer la couverture du lexique avec les mots annotés de ces textes. Cependant, nous souhaitons aussi rendre possible l'utilisation de la chaîne pour de nouvelles ressources, en particulier l'ensemble de la *BFM*. Pour certaines catégories morpho-syntaxiques, nous nous reportons aussi à l'annotation morpho-syntaxique de cette base pour mieux couvrir son vocabulaire.

5.2.2 Classes fermées

Nous procédons différemment en fonction du statut de la partie du discours, c'est-à-dire si elle fait partie des classes de mots "fermées" ou "ouvertes". Les classes dites "fermées" regroupent un nombre limité d'entrées. Il s'agit de mots "grammaticaux", appartenant aux catégories des pronoms, des déterminants, des prépositions, des interjections, des conjonctions de subordination et de coordination. Le relevé de ces entrées doit être le plus exhaustif possible, car ces catégories sont très fréquentes dans les phrases, et de leur bonne description dépend notamment l'analyse des syntagmes prépositionnels, des subordonnées et de la coordination de constituants. Nous avons aussi inclus les contractions à cette étape de corrections.

Seul un recensement manuel de ces formes peut être assez précis pour être exploité par une grammaire lexicalisée (Butt et al. (1999), p. 168).

Première estimation de l'ambiguïté des classes fermées

Le changement de l'ordre canonique des constituants principaux de *SOV* à *SVO* et, surtout, la perte de la déclinaison à six cas qui existait en latin, réduite à deux cas en AF, ont un impact sur les morphèmes grammaticaux. Le marquage flexionnel par les suffixes laisse peu à peu la place à la fixation de l'ordre des mots et à l'emploi de particules antéposées, notamment les prépositions (Amiot et al. 2020). Par exemple, les compléments de nom sont de plus en plus introduits par une préposition, *de* ou *à*. L'inventaire des prépositions n'est pas fixe comme en FC. Certaines suivent un processus de grammaticalisation qui les prive de leur valeur sémantique et en fait des introducteurs de compléments, comme *de* et *à*, dont la grammaticalisation est achevée en AF. Il s'agit des prépositions fonctionnelles. Leurs emplois varient au cours de la période, car elles ne sont pas toutes encore grammaticalisées, comme *sauve(s), sauf(s)* (12e s.) et *vu(e)(s)* (14e s.) (Amiot et al. (2020), p. 857–858). D'autres prépositions conservent une valeur sémantique, comme *vers*. Leur fréquence d'apparition évolue aussi. Certaines, héritées du latin, sont présentes dès les premiers textes, comme *aprof*, qui disparaîtra par la suite, comme plusieurs autres prépositions. D'autres n'apparaissent qu'en MF, comme *malgré, excepté, pendant,* et *durant*. Certaines formes, très ambiguës, connaissent un pic d'utilisation puis disparaissent, notamment *o* et *estre*. On observe aussi la constitution et l'emploi croissant de locutions propositionnelles de forme *préposition (+ déterminant) + nom + préposition*, comme *à l'encontre de*. N'étant pas immédiatement figées, leur intégration à un lexique statique pose problème. Le *Lefff* fait aussi face à ce type de situation, mais dans une moindre mesure.

Les conjonctions de subordinations suivent des étapes similaires. Leur processus de grammaticalisation couvre toute la période du FMed, ce qui rend difficile leur description dans un lexique. Les formes complexes, formées à partir de *que*, et incluant parfois le démonstratif *ce*, comme *par ce que*, sont décrites en syntaxe (cf. partie 4.6.3), car elles ne sont pas encore figées. Les composants de ces conjonctions peuvent varier, contrairement au FC, pour lequel il est possible de faire un inventaire des mots grammaticaux, même complexes.

Les frontières de certaines parties du discours sont parfois qualifiées de "poreuses" car certains mots en changent progressivement. C'est par exemple le cas de *moult* et *tant*, dont la terminaison est marquée par l'accord et la déclinaison en AF, mais qui tendent à devenir invariables, et deviennent des adverbes en MF.

Les variations graphiques et dialectales, typiques du FMed, contribuent aussi à agrandir l'inventaire des formes des classes fermées. Les entrées de chaque fichier du lexique sont donc vouées à être plus nombreuses qu'en FC, mais aussi plus difficiles à renseigner. Les traits associés dépendent en partie de l'état de langue dans lequel les formes apparaissent (époque, dialecte, domaine). De plus, toutes les formes ne sont pas décrites dans les ouvrages linguistiques. Le recours à des corpus annotés est nécessaire pour mettre en regard les informations contenues dans les grammaires et pour donner les bonnes descriptions.

Correction des classes fermées

La difficulté est d'une part de trouver toutes les formes dont nous avons besoin et de les renseigner adéquatement, et d'autre part de limiter l'ambiguïté autant que possible. Pour chaque partie du discours, nous avons fait une requête dans le *SRCMF* et dans la partie vérifiée de la *BFM* pour obtenir une liste de toutes

les formes qu'il nous faut représenter dans le lexique. Cela nous a permis de détecter les entrées inutiles et d'ajouter celles qui manquaient. Les pronoms et les déterminants ont ainsi obtenu un type (démonstratif, personnel, possessif...) issu de leur étiquette *Cattex*[5], et qui permettra de la reconstruire en sortie de parseur. Il était parfois présent dans la première version, mais seulement pour certaines catégories, et sa notation n'était pas harmonisée.

Pour chaque forme, il a fallu renseigner l'étiquette *UD*, qui est simple à déterminer à partir de l'annotation. Par exemple, les étiquettes *Cattex DET** (*DETdef, DETdem, DETind...*) sont toutes *DET* dans *UD*.

Si l'analyseur rencontre une forme inconnue, un module de correction orthographique est utilisé pour rapprocher le token d'une entrée existante. Les mots grammaticaux les plus fréquents étant très brefs (entre une et trois lettres), la distance d'édition avec la plupart de ces entrées est très faible. De nombreux candidats sont alors suggérés, ce qui augmente artificiellement l'ambiguïté de l'analyse. Par exemple, il y a une distance d'un caractère entre *en* et ces candidats : *e, an, ed, el, em, es, et, ex, ez, én, ·en, end, enn, ens, ent* et *enz*, mais tous n'en sont pas des variants graphiques, notamment *ed, et, ex* et *ez*. Les fichiers des classes fermées doivent donc être aussi exhaustifs que possible.

Une fois l'inventaire des formes établi, celles-ci ont ensuite été décrites selon les grammaires (Amiot et al. 2020 ; Buridant 2000 ; Hasenohr et Raynaud de Lage 1993) et les observations sur corpus. Il était nécessaire que cette démarche soit manuelle, car les informations linguistiques doivent être organisées en structures de traits exhaustives. Une entrée de déterminant ou de pronom doit par exemple contenir ses informations morphologiques (genre, nombre, cas), son statut de défini ou d'indéfini, et ses traits supplémentaires, s'il y en a, comme le trait de "réfléchi" (pour le pronom *se* notamment). Ces informations, lorsqu'elles apparaissaient dans un dictionnaire ou une grammaire, ont souvent demandé une interprétation, ce qui nous a obligé à en faire manuellement l'extraction. À partir des traits de type, nous avons déterminé si les pronoms et les déterminants étaient définis (articles définis, pronoms personnels, articles/pronoms démonstratifs et possessifs) ou non (articles cardinaux, pronoms impersonnels, articles/-pronoms indéfinis). Les structures de traits des entrées existantes ont été harmonisées, et les informations morphologiques ont été corrigées. Comme les formes du féminin au cas sujet et au cas régime sont identiques, ces entrées ont été rassemblées. La grammaire a besoin de catégories plus fines pour les pronoms, car les formes conjointes (appelées "clitiques" en FC) ne peuvent occuper que des places particulières, à proximité immédiate du verbe. Le terme de clitique ne convient pas au FMed, car les formes atones et toniques ne correspondent pas toujours respectivement aux formes conjointes et disjointes. Nous avons cependant gardé la notation du *Lefff* pour désigner ces pronoms conjoints, car nous en faisons le même emploi. Les catégories utilisées sont notées "cl" suivi d'une lettre pour le cas :

— cla : pronom conjoint accusatif (ex. *le*)

— cld : pronom conjoint datif (ex. *li*)

— clr : pronom conjoint réfléchi (ex. *se*)

— clg : pronom conjoint génitif (*en/an*)

— cll : pronom conjoint locatif (*y/i*)

Les requêtes sur corpus ont permis d'en établir l'inventaire, car ces pronoms apparaissent en séquences à proximité du verbe (cf. exemple 1).

5. Les étiquettes de partie du discours *Cattex* sont de forme *PARTIE DU DISCOURS* + *type*, ex. *PROper* pour "pronom personnel"

(1) quant Deus la li tramist
CONsub NOMpro PROper PROper VERcjg
'quand Dieu la lui transmet'
La Vie de Saint Alexis, v. 98 (ca. 1050)

La grammaire utilise une catégorie supplémentaire pour la conjonction de subordination *que*. Ses différentes graphies, par exemple *qu', ke, k'* et *c'*, ont donc été décrites en conséquence. La valence des prépositions, qui comprenait déjà les syntagmes nominaux et adverbiaux, a été étendue pour comprendre les verbes à l'infinitif. Certains pronoms ont gagné un cadre de valence, après un relevé sur corpus. Par exemple, dès l'AF, *chascun* accepte un complément introduit par *de*, comme *aucun* et *un*.

Limiter l'ambiguïté du lexique a été une étape importante, car toutes les entrées des formes rencontrées sont prises en compte lors de l'analyse syntaxique. Il est possible de neutraliser certaines entrées, si elles sont inutiles, mais il existe aussi un mécanisme de contraintes souples, celui des poids. Il intervient lors de la désambiguïsation (cf. partie 3.2.1). Si une forme est peu fréquente, un poids faible lui est attribué, et elle n'est conservée dans l'analyse finale que si c'est la meilleure solution. Cela permet de confirmer l'information de fréquence parfois donnée par les grammaires écrites par des linguistes. Cependant, il est difficile d'attribuer les traits d'époque et de dialecte, car les formes sortent souvent de ces cadres. Par exemple, Hasenohr et Raynaud de Lage (1993) classent le pronom *lié* comme une forme du dialecte de l'ouest, mais on en trouve quatorze occurrences dans *Tristan* de Béroul, identifié comme un écrit franco-picard (cf. exemple 2), et une dans *Cligès* de Chrétien de Troyes, dont le dialecte est le champenois. La forme *lié* peut aussi être le participe passé du verbe *lier* ou l'adjectif qualificatif *lié* (cf. exemple 3). Ajouter un trait de dialecte sur le pronom permettrait de limiter l'ambiguïté sur cette forme à un nombre limité de textes. Cependant, en raison des frontières poreuses entre états de langue et de dialectes, il ne nous est pas possible de documenter la variation graphique dans le lexique.

(2) Mal vos estoit **lié** a fallir
ADVneg PROper VERcjg **PROper** PRE VERinf
'**L'**abandonner vous n'avez pu'
Tristan de Béroul, v. 2 395 (fin 12e s.)

(3) Or sont plus **lié** qu' il ne soloient
CONcoo VERcjg ADVgen **ADJqua** CONsub PROper ADVneg VERcjg
'Ils sont bien plus **gais** que de coutume'
Le Chevalier de la Charrette de Chrétien de Troyes, v. 2 950 (ca. 1177–1181), trad. Catherine Croizy-Naquet

La nouvelle répartition des entrées est donnée dans le tableau 5.1. Le nombre de formes est inférieur au nombre d'entrées, car certaines formes regroupent plusieurs usages, comme *qui*, à la fois pronom relatif et interrogatif (en fonction sujet et objet). Dans la première version, certaines entrées étaient dédoublées, ce que nous avons cherché à corriger dans la deuxième version. Le lexique contient aussi un fichier de formes créées à partir de deux mots, les "contractions". En FC, ce sont des formes comme *du* pour *de le*. Ce fichier a été refait en suivant les mêmes étapes. Il contient désormais 147 formes dont l'analyse est attestée dans la *BFM*. Dans la deuxième version, les fichiers peuvent être plus ou moins volumineux que dans la première version, ce qui est indépendant des étapes réalisées. Nous n'avons cherché qu'à couvrir les formes de nos corpus.

Parties du discours	Version 1		Version 2	
	Nb. entrées	Nb. formes uniques	Nb. entrées	Nb. formes uniques
Pronoms	632	395	812	599
Déterminants	2 035	890	580	492
Prépositions	2 164	707	648	582
Conj. de subordination	116	106	89	89
Conj. de coordination	230	43	48	48
Interjections	186	1 236	57	57
Contractions	91	69	147	107
TOTAL	5 454	3 446	2 381	1 974

Tableau 5.1 – Taille des fichiers de classes fermées après les modifications

Dans cette deuxième version des classes fermées du lexique, le nombre d'entrées diminue de près de la moitié, évinçant ainsi 1 472 formes. Celles-ci pourront être réintégrées si un nouveau texte contient certaines d'entre elles. L'ambiguïté syntaxique se trouve donc diminuée, et les informations ajoutées peuvent être désormais toutes exploitées par le parseur. Toutes les entrées disponibles pour une forme sont considérées lors de l'analyse syntaxique. Il est donc primordial de fournir des descriptions adéquates et précises pour exploiter les mécanismes d'accord et de sélection d'arguments et de traits supplémentaires.

Une ambiguïté toujours forte

L'analyse syntaxique reste parfois très longue : on mesure en moyenne 40,22 secondes par phrase (mais 21,40 par phrase ayant reçu une analyse complète et la moitié des phrases est en fait analysée en moins de 2,6 secondes)[6]. Ceci est en partie dû à la grande souplesse de l'ordre des constituants majeurs en FMed. Les arguments d'un prédicat pouvant apparaître à sa gauche ou à sa droite, le nombre de combinaisons possibles est bien plus élevé qu'en FC. De plus, nous estimons que la graphie variable du FMed ajoute des sources d'ambiguïté.

Certaines formes sont très courantes, mais aussi très ambiguës, ce qui ralentit l'analyse des phrases, car chaque analyse possible est considérée tour à tour par le parseur. Nous donnerons ici quelques exemples, parmi les mots grammaticaux les plus fréquents dans la *BFM*.

Tableau 5.2 – Mots grammaticaux courants

Formes	**Occ. BFM**	**Occ. SRCMF**	**Nb de cat. possibles**	Détail de ces catégories (pourcentage dans la *BFM*)
en	101 416	3 472	4	préposition (58,61 %) clg (31,76 %) **cll** (6,44 %) **cln** (*on*, 3,18%)
Suite du tableau à la page suivante				

6. Cette moyenne est prise sur les expériences du mois d'avril 2022.

Tableau 5.2 – suite du tableau

Formes	Occ. BFM	Occ. SRCMF	Nb de cat. possibles	Détail de ces catégories (pourcentage dans la *BFM*)
la	90 618	3 083	3	article défini fém. sg. (84,08 %) cla fém. sg. (9,16 %) **adverbe** (*là*, 6,76%)
il	83 614	3 400	3	cln masc. sg. (65,82 %) **cln masc. pl.** (22,65 %) *il* impersonnel (11,53 %)
a	82 035	3 369	5	conjugaison du verbe *avoir* (22,57 %) **préposition** (*à*, 77,36 %) **interjection** (*ah*, 0,05 %) **nom commun masc. sg. cas régime** (0,01 %) **nom commun masc. pl. cas sujet** (0 %)
le	75 470	2 528	7	article défini masc. sg. régime (61,96 %) **article défini fém. sg. et pl.** (10,42 %) **cla fém. sg.** (0,12 %) cla masc. sg. (27,49 %) **adj. qual. fém. sg.** (0 %)
l’	57 220	2 150	5	article défini fém. sg. (23,17 %) article défini masc. sg. régime (38,13 %) cla fém. sg. (<38,7 %) cla masc. sg. (<38,7 %) **cla neutre sg.** (<38,7 %)
li	52 581	3 842	6	**cla fém. sg.** (1,94 %) **cld sg. masc. et fém.** (19,99 %) **pronom disjoint masc. sg.** (1,60 %) **article défini masc.** (65,45 %) **article défini fém.** (11,02 %)
les	44 347	1 222	6	article défini fém. pl. (32,02 %) article défini masc. pl. régime (37,30 %) cla pl. fém. et masc. (30,67 %) **adj. qual. fém. pl.** (0 %) **préposition** (0 %)
se	32 855	1 433	6	clr, cla, cld (54,64 %) **conj. de subordination** (*si*, 40,74 %) **adverbe** (*si*, 0,84 %) **article possessif sg. fém. et masc.** (*son*, 1,78 %)
s’	23 872	1 392	7	clr, cla, cld (66,98 %) conj. de subordination (*si*, 23,76 %) **adverbe** (*si*, 2,75 %) **article possessif sg. fém. et masc.** (*son*, 6,51 %) **adj. possessif** (*suen*, 0 %)

Les dix formes représentées dans le tableau 5.2 sont très polysémiques. Elles couvrent, à elles seules, environ 10,3% des mots du *SRCMF* et regroupent cinquante catégories grammaticales (colonne de droite). Les catégories en gras sont celles qui ont disparu en FC. L’analyse du FMed est donc considérablement

plus ambiguë à cause de ces formes. D'après le fichier des verbes, le mot *li* pourrait également être une flexion des verbes *lier, liier, lire, loier* ou *loiier*, ce qui ajouterait quatorze analyses supplémentaires. On relève aussi dix entrées verbales pour *les*, forme conjuguée du verbe *lesser/lessier*. Il est difficile de diminuer l'ambiguïté causée par ces formes sans leur attribuer des traits de dialecte ou d'époque. Cette solution ne concernerait néanmoins que quelques entrées, car la plupart apparaissent dans tout type de textes, avec des fréquences différentes. Nous gardons donc un même inventaire de formes pour tous les textes. On peut cependant attribuer un poids plus faible aux analyses les plus rares, limitant ainsi leurs fréquences d'apparition dans les analyses finales des phrases.

5.2.3 Classes ouvertes

Elles regroupent les classes suivantes : noms communs, noms propres, adverbes, adjectifs et verbes. Leur inventaire est théoriquement fini, car nous traitons d'états de langue anciens. Cependant, en pratique, il n'est pas exhaustif, car tous les corpus n'ont pas été recensés et exploités pour en extraire les mots. Cet inventaire est difficile à fixer à cause de la forte variation en FMed.

Les besoins du parsing

En plus de la liste des formes rencontrées en corpus, le parseur a besoin de certaines informations pour produire l'analyse d'une phrase. Comme pour les pronoms et les déterminants, les informations morphologiques des classes ouvertes permettent d'utiliser les mécanismes d'accord, nécessaires dans une grammaire qui repose en partie sur l'unification des nœuds. Les arbres élémentaires d'une *LTAG* représentent les prédicats avec leur arguments, le cadre valenciel des entrées est donc une information nécessaire.

Lors du traitement de nouveaux textes, il est inévitable de rencontrer des mots inconnus. Il est donc nécessaire d'enrichir le mécanisme de reconnaissance de tels tokens avec des connaissances sur le FMed, en donnant par exemple une liste des suffixes typiques de certaines catégories morpho-syntaxiques. L'objectif d'une couverture totale du lexique n'est pas réalisable, et on peut même avancer qu'il n'est pas souhaitable, à cause de la rareté de certaines variantes graphiques. En revanche, permettre à l'analyseur d'être souple face à de nouvelles formes et renseigner des connaissances linguistiques pour analyser morphologiquement ces formes semble plus pertinent et plus économique.

Modification des entrées

Nous procédons à un filtrage de surface pour réduire le nombre d'entrées (cf. tableau 5.3). Certaines informations ne sont en effet pas exploitées, comme, d'une part, les macros de flexion, qui multiplient parfois le nombre d'entrées pour un seul couple forme – lemme. Posséder de nombreuses entrées pour une même forme, avec des informations identiques, n'empêche pas l'analyse syntaxique en soi. Cependant, cela génère une ambiguïté inutile sur ces formes et gêne ainsi cette analyse, car au-delà d'un certain seuil d'ambiguïté dans l'analyse d'une phrase, le coût de calcul devient trop élevé, et l'analyse risque d'échouer. Par ailleurs, lorsque les formes des cas sujet et régime au féminin étaient identiques, nous les avons rassemblées, ce qui a divisé presque par deux le nombre d'entrées de noms et d'adjectifs ayant ce trait de genre. Le cas n'étant pas utilisé par l'analyse syntaxique, cette économie n'a pas entraîné de confusion pour la grammaire. De manière plus anecdotique, nous avons parfois supprimé manuellement des entrées

qui sont artificiellement rattachées à de multiples lemmes. Certaines formes ont par exemple deux entrées : l'une rattachée à un lemme commun à d'autres variants graphiques d'un même mot, et une autre rattachée à un nouveau lemme, identique à la forme fléchie. Par exemple, l'adverbe *uncore* est rattaché au lemme *uncore* et à celui qui est commun à l'ensemble des dictionnaires, *encore*. Il n'est cependant pas aisé de procéder automatiquement au choix du lemme, car il n'existe pas de standard exhaustif qui couvre toute la période du FMed. On a aussi supprimé les entrées de genre "neutre" et la plupart des entrées composées de plusieurs tokens, notamment dans la catégorie des adverbes. Les séquences concernées, par exemple *al jorn d'ui* (trad. *aujourd'hui*), peuvent être analysées en syntaxe (en l'occurrence, comme syntagme prépositionnel). On ne souhaite pas rapprocher ces locutions de leurs lemmes contemporains. Enfin, certaines formes nous ont parfois semblé être renseignées dans la mauvaise catégorie, comme *kis* dans les adverbes, qui est en réalité une contraction de *ki* (*qui*) et *les*. Nous avons aussi écarté les formes de préfixes des adverbes, comme *r'* et *re-*, que nous préférons ne pas considérer systématiquement comme des tokens distincts.

Parties du discours	Version 1		Version 2	
	Nb. entrées	Nb. formes uniques	Nb. entrées	Nb. formes uniques
Adjectifs	63 256	24 733	46 471	24 356
Adverbes	4 380	3 827	3 903	3 444
Noms communs	161 638	70 730	123 915	69 523
Noms propres	11 255	2 202	2 567	2 195
Verbes	651 648	516 481	632 735	491 928
TOTAL	892 177	617 973	809 591	591 446

Tableau 5.3 – Taille des fichiers de classes ouvertes après le tri des entrées

De nombreuses entrées ont ainsi été supprimées, mais il a fallu en ajouter d'autres. Le nombre des formes manquantes étant élevé, nous avons procédé de manière automatique, contrairement aux formes des classes fermées. Nous avons procédé à un premier ajout de formes du corpus *Profiterole* absentes du lexique grâce aux textes annotés en parties du discours. Nous avons aussi ajouté des formes issues des textes vérifiés de la *BFM*, pour les tokens dépassant les dix occurrences (cf. tableau 5.4).

Parties du discours	Nb. d'ajouts
Adjectifs	1 283
Adverbes	80
Noms communs	3 788
Noms propres	648
Verbes	26 432
Total	32 231

Tableau 5.4 – Ajouts réalisés

Les textes utilisés sont annotés au format *Cattex*, ce qui nous donne des indications riches. Les noms sont divisés en noms propres et noms communs, et les adjectifs sont de type qualificatif, indéfini ou possessif. Les informations morphologiques (genre, nombre) des nouveaux noms communs sont déduites

à partir de leur terminaison, selon les listes proposées dans le tableau 5.5. Par exemple, le mot *acontance*, absent de la première version mais attesté dans *Yvain*, est renseigné comme substantif féminin singulier, car sa terminaison est *-ance*. Le mot *cousins* manquait aussi. Il est renseigné comme substantif masculin (*-in*) au cas sujet singulier et au régime pluriel (*-s*). Nous avons suivi la même démarche que pour la constitution du premier *OFrLex* : les noms masculins sont répartis en nombre et en cas. Il semble que les formes se terminant par *-s, -x* [7] ou *-z* [8] sont traitées comme cas sujet singulier ou cas régime pluriel, et que cette forme sans la lettre finale représente les cas restants (i.e. cas sujet pluriel et cas régime singulier). On trouve cependant quelques exceptions, comme *foiz*, qui est un nom féminin.

Terminaison	Marque du pluriel	Informations morphologiques
-ison, -ïson, -isun, ïsun, -uson, *-te, -tet, -té, -tét, -thét,* *-tie, -tiet, -tié, -tiét,* *-ance, -ence, -once, -unce,* *-esse, esce* *-tion, -cion, -ciun, -sion, -siun, -tiun, -tiün, -siün* *-oise* *-a, -ne, -tu, -re*	aucune	sg.fem
idem	*-s, -z*	pl.fem
-ant, -ent *-esme* *-or, -ort** *-in* par défaut, toutes les terminaisons restantes	aucune	sg.obl.masc pl.nom.masc
idem	*-s, -x, -z*	sg.nom.masc pl.obl.masc

Tableau 5.5 – Règles morphologiques pour les noms communs (* sauf *cort*, nom féminin)

Les entrées verbales ont été complétées en plusieurs étapes, mais elles restent cependant lacunaires. Des tables de flexions ont été ajoutées pour les verbes essentiels, comme *estre, avoir, aidier, aler, croire, devoir*... 573 formes uniques ont été ainsi ajoutées. Ces verbes comptent plus de formes que leurs équivalents contemporains. Le verbe *dire* a 69 formes en FC, contre 209 en FMed (d'après notre relevé actuel). Quelques graphies peuvent être confondues avec des homonymes d'autres catégories morpho-syntaxiques, comme *diz*, qui peut être le verbe dire à diverses conjugaisons (passé simple et présent de l'indicatif, impératif, participe passé), un numéral (*dix*) ou un nom commun (au sens de *parole, propos* [9]). Une description exhaustive des formes connues de ces verbes fréquents permet d'éviter des erreurs d'analyse causées par des mots inconnus ou mal étiquetés.

Les verbes qui acceptent la voix passive ont aussi reçu des entrées de participe passé passif et de participe passé employé comme adjectif, comme il est d'usage dans des lexiques suivant le format du *Lefff*. De nombreuses formes du corpus *Profiterole* manquaient encore dans le lexique. Cependant, le jeu d'étiquette *Cattex* faisant la distinction entre verbes conjugués (*VERcjg*), infinitifs (*VERinf*), participes présents (*VERppa*) et participes passés (VERppe), il a été possible de compléter certaines conjugaisons à

7. La terminaison *-x* est initialement la simple graphie de *-us* (ex. *chevaus* > *chevax*).
8. La terminaison *-z* resulte d'un accident phonétique (dentale + *-s*).
9. Nous nous référons à la définition du *DMF*, disponible à cette adresse : `http://www.atilf.fr/dmf/definition/dit`.

un coût manuel faible. Les entrées de participes présents ainsi ajoutées sont au nombre de 413, et celles de participes passés, au nombre de 4 535 pour la voix active. Par défaut, ces entrées sont décrites comme des transitifs directs, car nous n'avons pas procédé à la lemmatisation de ces entrées et le cadre valenciel en contexte sera renseigné dans une prochaine étape. Certaines ont pu être rapprochées manuellement d'entrées existantes et ont reçu leur cadre valenciel. Les participes passés transitifs ont ensuite été adaptés au passif, avec un changement de l'auxiliaire requis et du cadre valenciel, et en tant que participes passés employés comme adjectifs, pour un total de 13 950 entrées de participes passés. Cette étape a permis de donner la description morphologique des formes (genre et nombre), qui était absente dans la première version du lexique. Nous avons procédé de la même manière pour les 964 infinitifs à ajouter. Les verbes conjugués sont plus difficiles à renseigner, car les changements de radical et les diverses terminaisons possibles selon les temps et les modes peuvent être difficiles à identifier, ce qui rend impossible le traitement automatique de l'ensemble des 7 205 formes restantes. Comme il s'agit essentiellement de formes rares (60,7% n'ont qu'une occurrence et 85,7% en ont moins de cinq), nous comptons sur les suggestions de *Sxpipe* et de son module de correction orthographique pour l'analyse syntaxique des formes les moins fréquentes. Pour permettre d'évaluer la grammaire sur le *SRCMF* et estimer son taux de couverture sur le corpus *Profiterole*, nous avons souhaité intégrer les formes plus courantes en amont. Avec Sophie Prévost, nous avons identifié les tokens les plus susceptibles d'être renseignés, parmi ceux qui sont attestés au moins dix fois dans le corpus, pour les intégrer manuellement au lexique. Cela permet d'ajouter 420 formes conjuguées de verbes déjà renseignés, comme *appeler* et *aimer*, et ainsi de leur attribuer un cadre valenciel plus juste (cf. partie 5.2.4) et une description morphologique (mode, temps, personne et nombre).

Une fois la première étape d'ajouts d'entrée effectuée, nous avons estimé qu'il manquait encore 32 509 formes pour faire l'analyse syntaxique du corpus *Profiterole*. Une deuxième étape a été effectuée en collaboration avec Cristina Holgado, grâce à ses travaux sur la lemmatisation du FMed (Holgado, Lavrentiev et Constant 2021) et son maillage des lemmes d'*OFrLex* avec ceux qui sont utilisés dans la *BFM*. Nous avons ajouté les formes auxquelles un seul lemme a été attribué (et si celui-ci avait un équivalent dans *OFrLex*), car l'ajout non contrôlé de formes identiques sur de multiples lemmes risque d'augmenter considérablement l'ambiguïté. Par exemple, nous n'avons pas souhaité attacher la forme *aingnez* à l'ensemble des lemmes suivants : *aine, aigne, haine, aîné, agne, aine, aim, ange, ane, annes, haigne, haignée*. La lemmatisation a été faite avec *LGeRM*, qui utilise des règles morphologiques. Or, certaines transformations sont souvent proposées, alors qu'elles ne s'appliquent pas à tous les dialectes, comme celle de *g(u)-* en *w-* (et inversement), qui est surtout le fait du picard (ex. 4). Les lemmes proposés sont donc parfois trop éloignés pour procéder à un ajout automatique.

(4) li marchis, qui estoit sires de l' ost, eut l'
DETdef NOMcom PROrel VERcjg NOMcom PRE DETdef NOMcom VERcjg DETdef
arriere **warde**, et **warda** l' ost par deriere
ADVgen NOMcom CONcoo VERcjg DETdef NOMcom PRE ADVgen
'le marquis, qui était le seigneur de l'est, eut l'arrière-**garde** et **garda** l'est par derrière'
Conquête de Constantinople de Robert de Clari, p. 46 (ca. 1300)

Après cette étape (cf. tableau 5.6), il manquait donc encore 24 161 formes pour couvrir totalement le corpus *Profiterole*. Nous estimons que le recours à la lemmatisation est un moyen efficace de renseigner

Parties du discours	Nb. d'ajouts	Nb. formes uniques
Adjectifs	2 467	2 467
Adverbes	202	202
Noms communs	4 807	4 807
Noms propres	874	872
Total	8 350	8 348

Tableau 5.6 – Ajouts réalisés

de nouvelles entrées en utilisant les informations du lexique existant. Certaines de ces formes peuvent aussi être analysées grâce à des mécanismes annexes, présentés ci-après.

Adaptation des mécanismes de reconnaissance des tokens inconnus

Comme nous l'avons indiqué précemment, obtenir un lexique totalement exhaustif est une tâche coûteuse, et, dans le cas d'une langue fortement soumise à la variation graphique, ce n'est pas nécessairement souhaitable. En effet, cette variation génère de nombreux hapax et des cas d'ambiguïté difficiles à renseigner automatiquement. En revanche, nous pouvons nous servir de plusieurs mécanismes pour analyser des formes inconnues.

Tout d'abord, le segmenteur *Sxpipe* peut relier des tokens à des entrées existantes en leur appliquant des préfixes et des suffixes renseignés au préalable pour la langue concernée. Grâce à l'analyse des formes manquantes et le calcul de distances d'édition avec les entrées manquantes effectué par Cristina Holgado, nous pouvons enrichir et corriger l'inventaire initial des affixes.

— Préfixes : *a-, ci-/cy-, entr'-/entr-, in-/im-, mi-/my-, sor-/sur-, sos-/sous/-soz-, très-/tres-...*

— Suffixes : *-ci/-cy, -là*

Enfin, un mécanisme a été ajouté dans l'analyseur lexical des métagrammaires pour opérer des transformations sur les tokens inconnus. Nous nous servons de cet outil pour faire une liste des terminaisons communes et de leurs variations graphiques, par exemple *-tion*, qui peut être écrit *-cion, -ciun, -cïon, -cïun, -sion, -siun, -tiun, -tïun* ou *-sïun*. Ces formes ont été renseignées grâce à des requêtes sur corpus et aux tableaux de Carlier et Combettes (2020) (p. 622–630).

Traits sémantiques

Deux catégories sémantiques de noms communs, les temporels et les inaliénables, peuvent ancrer des modifieurs non introduits par une préposition. La seconde catégorie regroupe habituellement les parties du corps, et nous l'avons étendue à l'équipement, car les comportements syntaxiques de ces deux ensembles sont similaires.

(5) Kar hoi **matin** vos vi plurer des oilz
CONcoo ADVgen **NOMcom** PROper VERcjg VERinf PRE.DETdef NOMcom
'Car, **ce matin** [aujourd'hui matin], je vous ai vu pleurer des yeux.'
Chanson de Roland, v. 3 629 (ca. 1100)

(6) **Lance** levee, l' **escu** pris, A Tristran saut en mié le
NOMcom VERppe DETdef **NOMcom** VERppe PRE NOMpro VERcjg PRE NOMcom DETdef
vis.
NOMcom
'**La lance levée et l'écu pris**, il bondit sur Tristan, l'attaquant de face.'
Tristan de Béroul, v. 4 037 (fin 12e s.)

Il est important de les renseigner dans le lexique pour pouvoir utiliser des arbres spéciaux de la grammaire, ancrés spécifiquement par des tokens porteurs de ces traits. Un premier inventaire de ces formes est effectué dans le *SRCMF-UD 2.7*, à partir des compléments *obl* qui ne sont pas précédés d'une préposition. Nous leur attribuons ces traits :

— <time> : ancres de compléments de temps (ex. *jour, soir, fois*), pour 51 formes (17 lemmes) dans un relevé de 608 occurrences

— <bodypart> : inaliénables, pour 25 formes (14 lemmes) dans un relevé de 32 occurrences.

Le *Lefff* dispose d'autres catégories sémantiques. Nous en renseignons certaines entrées manuellement pour obtenir des exemples d'analyses plus fines :

— <hum> : pour les animés (ex. *Costentin, Damedeu*...)

— <first_name> : pour les prénoms (ex. *Iseult, Yvain, Tristan*...)

— <loc> : pour les noms de lieux (ex. *Abrimacie, Carcassonie*...)

Il est aussi possible d'ajouter de nouvelles catégories sémantiques si nécessaire.

Entités nommées

L'analyse du FMed demande globalement les mêmes ressources que le FC. Si on se réfère aux textes de la *BFM*, les éditions des textes anciens reprennent les mêmes normes que celles des écrits standards contemporains. Les tokens sont séparés par des espaces, sauf en cas d'enclise, et l'usage de la ponctuation est identique. Les manuscrits ne suivent pas ces normes, mais notre outil est uniquement destiné au traitement de textes édités. Quelques adaptations du segmenteur *Sxpipe* sont cependant nécessaires pour traiter quelques cas d'entités nommées.

Chiffres romains Un développement supplémentaire est nécessaire pour les chiffres romains, car leur notation suit des règles différentes en FMed. Cette modification a été faite avec Eric de la Clergerie. La particularité principale des chiffres romains en FMed est qu'ils peuvent être encadrés de points, et parfois n'avoir qu'un point final. À l'aide d'une requête sur l'ensemble de la *BFM*, nous sommes en mesure d'adapter les règles de reconnaissance des chiffres romains dans *Sxpipe*. Voici quelques exemples de formes rencontrées :

— lé *.viii.* kalendes (*Comput*)

— en l'an de l'Incarnation *.M CC XLVIII.* (*Chartes du Hainaut*)

— *.VII.XX.* an (*Erec et Enide*)

— a moins de *C.* lieues (*Mélusine*)

Ces points ne doivent pas être analysés comme des frontières de phrase, sous peine de faire échouer l'analyse. Des espaces peuvent aussi apparaître dans ces séquences, lorsque le chiffre est encadré de points. On observe également une alternance de majuscule et de minuscules dans de nombreux exemples, et l'ordre des lettres ne correspond pas à celui du FC. Par exemple, les formes *.Ixij.* et *.Ixii.* (*Chartes du Hainaut*) ne respectent pas les normes du FC, qui ne prévoit pas la possibilité d'une unité précédant une dizaine, elle-même suivie par des unités. On pourrait en déduire qu'il s'agit de 1+12 ou de 9+2. Il faut aussi noter que *j* peut se substituer à *i*. On cherche à contraindre autant que possible l'analyse des formes à point final, pour éviter de les confondre avec des abréviations. De même, l'analyse des chiffres romains ordinaux (ex. *IIJe*, *Journal* de Nicolas de Baye) est plus contrainte que celle des chiffres cardinaux, afin d'exclure l'analyse d'éléments comme *Ce, ce, Le, le* et *je*.

Une nouvelle description est donc proposée pour prendre en compte ces différents cas de figure. Elle inclut aussi la graphie de *ambedui* (trad. *tous les deux*) contenant des chiffres romains : *ambe.ii.*, *an.ii.*, *en.ii.*.

Chiffres L'inventaire des chiffres écrits en toutes lettres n'est pas fait dans le lexique, mais dans le segmenteur *Sxpipe* afin de les traiter comme des expressions polylexicales. Cela permet de reconnaître des formes comme *cinc cenz* (*Quatre Livres des Rois*) et *cinquante quatre mille* (*Mémoires* de Philippe de Commynes). Ces nombres, tout comme les chiffres romains, peuvent ensuite être analysés comme des déterminants, des pronoms ou des adjectifs lors du traitement de la phrase. Les variations graphiques des chiffres sont nombreuses et doivent être recensées dans le segmenteur pour être détectées. Nous intégrons aussi les ordinaux à ce fichier, pour un total de 200 formes. Les listes des graphies sont obtenues grâce aux annotations *Cattex* de la *BFM*, qui distinguent les déterminants, les adjectifs et les pronoms cardinaux des ordinaux. Nous gardons les règles de combinaison de ces tokens développées pour le FC.

Dates et durées Les chiffres détectés peuvent servir à la détection de dates en amont de l'analyse syntaxique. Nous gardons également les patrons de dates du FC, et nous ajoutons les variations graphiques des noms de mois (63 formes) et des unités de temps comme l'an et le jour. Ces relevés sont également effectués à partir de la *BFM*.

5.2.4 Renseigner les valences verbales

L'information de valence est nécessaire à l'attachement des arguments à leur prédicat. Dans une *LTAG*, le nombre d'arguments attendus est indiqué, ainsi que leur fonction et leur réalisation. On trouve au plus trois arguments : *Arg0*, *Arg1* et *Arg2*. Dans une phrase canonique, *Arg0* est le sujet (*Suj*), *Arg1* l'objet direct (*Obj*) ou l'attribut du sujet (Att), et *Arg2* l'objet second (*Objà/Objde*) ou oblique (*Obl*). Il est possible d'y ajouter un complément locatif (*Loc*) ou délocatif (*Dloc*) s'ils participent du sens du verbe, mais leur analyse n'est pas prioritaire. Nous reprenons la dénomination des principales réalisations du *Lefff* :

— syntagme nominal (*sn*), qui peut être précédée d'une préposition (*à-sn, de-sn*)

— syntagme adjectival (*sa*)

— pronoms conjoints (*cla, cld, clr, cll, clg*)

— pronom conjoint nominatif [10] (*cln*)

— complétive (*scompl*)

— interrogative (*qcompl*)

— infinitive (*sinf*), qui peut être précédée d'une préposition (*à-sinf, de-sinf*)

Travaux existants

Le cadre valenciel d'un verbe contient le nombre de ses arguments essentiels et leurs réalisations syntaxiques. Le statut des pronoms employés, notamment *il* et *se*, doit être précisé (Van den Eynde et Mertens 2003). Le pronom *il* peut être impersonnel, auquel cas il ne peut pas entrer dans un paradigme avec d'autres éléments. D'après ces auteurs, le pronom *se* peut faire partie du paradigme des pronoms (ex. *se laver*), être un "*se* (passif) de formulation" (ex. *cela s'obtient*) ou être un "*se* sans paradigme et sans reformulation", c'est-à-dire qu'il fait partie du prédicateur lui-même (ex. *s'évanouir*). Le lexique doit aussi préciser s'il s'agit d'un verbe plein, d'un verbe support (ex. *forcer la main*), d'une copule ou d'un auxiliaire.

Il existe plusieurs moyens d'obtenir cette information pour un lexique. Le premier consiste à exploiter des sources linguistiques (Gardent et al. 2006), mais nous n'en disposons pas pour le FMed dans un format exploitable automatiquement. Le second consiste à rechercher ces cadres de valence en corpus. Pour les langues qui ne sont pas à ordre libre, il est possible d'utiliser un corpus brut. Cette démarche a été adoptée pour l'anglais (Briscoe et Carroll 1997 ; Carroll et Fang 2004 ; Korhonen 2002). Il est aussi possible d'utiliser un corpus arboré, comme Kupść et A. Abeillé (2008) l'ont fait pour créer *Treelex*. L'annotation des constituants et des fonctions syntaxiques leur a permis, dans un premier temps, de relier les arguments à leur verbe. Dans un deuxième temps, les cadres valenciels sont rendus plus compacts en attachant les pronoms clitiques à leur fonction syntaxique. Enfin, les cadres de la voix passive sont séparés de ceux de la voix active. Cette méthode est réutilisable pour des langues sans ordre fixe des constituants principaux. Cependant l'annotation du *treebank* ne distingue pas les modifieurs optionnels de ceux qui sont nécessaires (ex. *elle va bien*). Les autrices notent aussi l'importance de la taille des données dans cette tâche, en comparant notamment les cadres de valence obtenus avec les verbes des propositions principales seulement (1 362 formes) et ceux obtenus avec tous les verbes (2 006 formes). Nous suivons une procédure similaire, mais avec un corpus annoté en dépendances syntaxiques.

Méthode de constitution des cadres de valences verbales

Nous choisissons de renseigner les cadres valenciels des verbes au moyen d'une extraction automatique sur corpus. Pour chaque verbe conjugué dans la version *Universal Dependencies 2.7* du *SRCMF*, cela permet théoriquement de rassembler les informations syntaxiques dont nous avons besoin pour analyser ce corpus, en faisant remonter les réalisations possibles des arguments au lemme verbal. En l'absence de ressource lexicale fine pour l'ensemble de la période du FMed, nous comptons garder ce premier état du lexique pour la première analyse du corpus *Profiterole*.

10. Ce type de pronom a plus d'autonomie en FMed qu'en FC, mais il ne peut pas occuper certaines positions, comme celle de complément introduit par une préposition.

Interrogation de corpus annotés avec *GREW*

D'après Van den Eynde et Mertens (2003), une grammaire de valence est une forme particulière d'une grammaire de dépendances. Il nous est donc possible de transformer certaines dépendances syntaxiques du *treebank* en nouvelles dépendances de type *Alexina*, le formalisme des lexiques *Lefff* et *OFrLex*. Afin d'extraire les arguments des verbes dans le *SRCMF-UD*, nous utilisons la librairie *Python GREW* (Bonfante, Guillaume et Perrier 2018). Celle-ci permet d'extraire des patrons syntaxiques sous forme de graphes dans un corpus annoté au format *CONLL Universal Dependencies*. Cette représentation en graphes (ex. figure 5.2) permet d'obtenir des dépendances syntaxiques entre des nœuds, mais aussi de représenter d'autres niveaux linguistiques, comme l'ordre linéaire des mots.

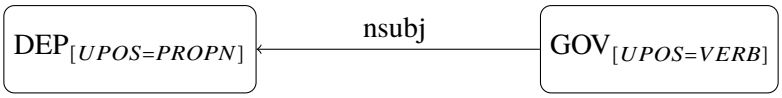


Figure 5.2 – Exemple de graphe cherchant un verbe dont le sujet est un nom propre, dans l'ordre *SV*

Dans un cadre valenciel canonique pour une *TAG*, l'argument *Arg0* est le sujet. En FMed, son expression n'est pas obligatoire, et elle est même minoritaire dans les premiers textes. Nous avons donc écarté la fonction de sujet nominal et pronominal de nos requêtes, la considérant comme une réalisation par défaut (mais optionnelle). Les sujets verbaux et phrastiques ont cependant été recherchés pour en déterminer les réalisations possibles. Le pronom sujet impersonnel ne compte pas comme paradigme dans le lexique. L'annotation du *SRCMF* donnant cette information par le biais de l'étiquette *Cattex PROimp*, il a été aisé de la reporter dans les cadres de valence. Cette requête a été croisée avec celle de la dépendance *expl* (explétif), qui donne les sujets impersonnels, afin de contrôler les informations ajoutées.

Arg1 peut être l'objet direct ou l'attribut du sujet. La fonction d'objet dans le lexique regroupe l'objet nominal (*obj*), la complétive (*ccomp* et *obj* ancré par un verbe conjugué) et l'infinitive (*obj* ancré par un infinitif). Ces différentes réalisations sont renseignées dans le cadre valenciel. Leur pronominalisation est généralement attestée, nous la repérons grâce à la dépendance *obj*, accompagnée par un tri sur les tokens sélectionnés, pour ne conserver que les pronoms conjoints. La fonction d'attribut du sujet a été repérée grâce à la dépendance *cop* (copule), dont l'attribut est la tête. Les différentes réalisations de cette fonction ont aussi été renseignées. Il peut s'agir d'un syntagme nominal, adjectival, adverbial...

L'objet second est symbolisé par *Arg2*. Dans le formalisme *Alexina*, il peut être introduit par la préposition *à* (*Objà*) ou par *de* (*Objde*). Le syntagme introduit par cette préposition peut être nominal ou verbal. Ces réalisations ont été recherchées avec la dépendance *obl* et la dépendance des prépositions et de leurs variantes graphiques. Il n'est cependant pas garanti que le complément extrait soit un argument du verbe. Il peut s'agir d'un modifieur. Nous avons fait l'hypothèse que l'extraction de pronoms conjoints correspondants pour le même verbe informait sur le caractère essentiel du complément. Si un verbe prenait un complément introduit par *à* et un pronom datif (ex. *li, lor*), on pouvait considérer qu'*Objà* était un bon candidat pour remplir *Arg2*. Si un verbe prenait un complément introduit par *de* et un pronom génitif (*en*), on pouvait considérer qu'*Objde* était un bon candidat. Il est possible d'avoir deux réalisations pour *Arg2* (ex. *paroler* de quelque chose à quelqu'un) à condition que *Arg1* soit absent, car la grammaire accepte un maximum de trois arguments. Cependant, les cadres valenciels ainsi obtenus ont parfois généré des conflits. *Objà* a été plusieurs fois confondu avec un complément locatif (*Loc*) et *Objde* avec un complément délocatif (*DLoc*). De plus, la valence des verbes a évolué au cours de la période, provoquant

parfois l'obtention de cadres valenciels trop larges, et on a aussi trouvé une alternance de la préposition introduisant l'objet second. Amiot et al. (2020) (p. 868) résument cela ainsi : "La transitivité verbale a, de manière générale, connu une histoire assez chaotique dans la diachronie du français". Ces cas de figure ont souvent demandé une résolution manuelle. En l'absence d'une ressource de référence pour le FMed, il ne nous était pas possible de rapprocher des cadres automatiquement.

Outre *Loc* et *DLoc*, il est possible d'ajouter des compléments hors du cadre des arguments. *Obl* et *Obl2* peuvent accueillir des compléments supplémentaires introduits par les prépositions *à* et *de*, ce qui peut servir à héberger les dépendances rejetées du cadre argumental, mais aussi des compléments introduits par d'autres prépositions, comme *avec* et *sur*, qu'il est possible de renseigner dans l'entrée.

Les verbes transitifs n'ont pas été systématiquement décrits comme passivables. Cette information a également été extraite du *SRCMF* avec la dépendance *aux:pass*. Le lexique ne transforme pas encore les participes passés de ces entrées automatiquement, les entrées ont donc été générés à part. Recourir à un corpus annoté nous a aussi permis d'extraire automatiquement toutes les flexions des auxiliaires, qui n'étaient pas renseignées.

Le pseudo-paradigme *se* (Van den Eynde et Mertens 2003) est obtenu par une requête de la dépendance *expl* en excluant le pronom impersonnel *il* et ses variants graphiques. Il n'est pas renseigné comme argument, mais comme partie du prédicateur.

Certains verbes présentent des cadres trop complexes pour être renseignés automatiquement et font l'objet d'un traitement manuel. C'est par exemple le cas du verbe *faire* et de ses variantes graphiques, car il nécessite six entrées pour couvrir tous ses cas d'usage en FC, comme dans le *Lefff* (et trois en AF).

(7) ex. entrée canonique

Or **fai** ton mialz et je le mien
CONcoo **VERcjg** DETpso NOMcom CONcoo PROper DETdef PROpos
'Or tu fais de ton mieux, et moi du mien.'
Yvain de Chrétien de Troyes, v. 4 184 (1177–1181)

(8) ex. verbe faible

Fai le **venir**
VERcjg PROper **VERinf**
'Fais-le venir'
Quatre Livres des rois, p. 31 (ca. 1190)

(9) ex. emploi avec un attribut et un pronom réfléchi

Se chaschun ne **se** **fait** **humble** comme che enfant, il
CONsub PROind ADVneg **PROper** **VERcjg** **ADJqua** PRE DETdem NOMcom PROper
n' enterra ja ou royaume des chieulx
ADVneg VERcjg ADVgen PRE.DETdef NOMcom PRE.DETdef NOMcom
'Si chacun ne se fait pas humble comme un enfant, il n'entrera jamais au Royaume des Cieux'
De la erudition de Jean Daudin, p. 394 (1360)

Nous avons décrit les cadres de contrôle (A. Abeillé 1998) en suivant le modèle du *Lefff* (Sagot et Danlos 2008), à partir des dépendances *xcomp*. Il s'agit de verbes qui ont "au moins deux arguments sémantiques", dont un verbe à l'infinitif, et la sélection du sujet de celui-ci est faite conjointement par les deux verbes. Par exemple, le verbe *povoir* et son infinitif partagent le même sujet, ce qui permet d'expliquer l'accord du participe passé *oïe* :

(10) vostre voiz n' i **peut estre oïe** fors que de
DETpos NOMcom ADVneg PROadv **VERcjg VERinf VERppe** ADVgen CONsub PRE
moi tant seulement
PROper ADVgen ADVgen
'votre voix ne peut y être entendue que par moi'
Roman de la rose de Jean de Meun, v. 16 394 (1269–1278)

Le sujet du verbe à contrôle n'est pas toujours identique à l'argument sélectionné par l'infinitif. Il peut s'agir d'un autre dépendant, comme l'objet :

(11) l'en **le accusoit** d' **avoir veu** et **sceu** en quel lieu...
PROper **PROper VERcjg** PRE **VERinf VERcjg** CONcoo **VERcjg** PRE DETint NOMcom
'on **l'**accusait d'**avoir vu** et **su** en quel lieu...'
Registre criminel du Châtelet, p. 211 (1389-1392)

Cependant, l'accord n'est pas toujours suivi, contrairement au FC, mais le lien sémantique demeure :

(12) puis que sa biauté ne son pris ne peut estre
ADVgen CONsub DETpos NOMcom ADVneg DETpos NOMcom ADVneg VERcjg VERinf
d' ome **compris**
PRE NOMcom VERppe
'puisque ni sa beauté, ni son prix, ne peuvent être **compris** par les hommes / puisque sa beauté ne peut être **comprise** par les hommes, ni son prix'
Roman de la rose de Jean de Meun, v. 16 217 (1269–1278)

Ces informations ne peuvent pas être extraites automatiquement, car il faut renseigner entre quels éléments on trouve cette dépendance secondaire, qui n'est pas exprimée dans le format *Universal Dependencies*. En revanche, les parseurs *FRMG* et *MetaMOF* donnent cette dépendance en sortie.

Nous avons ainsi obtenu une première version des cadres de valence pour 1 885 verbes à partir du *SRCMF-UD*, ainsi que quelques cadres décrits manuellement. Dans un deuxième temps, le maillage entre les lemme d'*OFrLex* avec ceux obtenus par la lemmatisation a surtout permis de mieux diffuser les arguments observés en contexte. La répartition des arguments et de leur réalisation est disponible dans le tableau 5.7. Les pronoms conjoints n'y figurent pas, mais ils sont automatiquement proposés avec les arguments adéquats (ex. *cln* pour l'ensemble des réalisations sujet, *cla* pour les réalisations de *Arg1*, *cld* pour *Objà*...). Ces informations sont amenées à être corrigées dans de prochains travaux, notamment avec la fouille d'erreur en sortie de parseur et l'apport de nouvelles données annotées. Cela nous permet cependant de faire des analyses de corpus entiers.

Le lexique du FMed que nous utilisons doit être assez couvrant et contenir des informations morphologiques et syntaxiques pertinentes pour permettre une analyse syntaxique sur corpus. Plus que le manque de ressources lexicales construites pour le *TAL*, la difficulté que présente le FMed par rapport au FC est la variation graphique, qui démultiplie le nombre de formes disponibles pour un lemme. Il ne s'agit donc pas de considérer ce lexique comme un bloc d'entrées, mais comme un réservoir de formes renseignées avec le plus de confiance possible, dans lequel l'analyseur peut puiser pour analyser tous les tokens des phrases.

Arguments	Réalisations	Nouvelle version
Arg0 Suj	sn	22 705
	ilimp	9
	scompl	207
	sinf	92
Arg1 Obj	sn	19 184
	scompl	444
	sinf	71
Arg1 AttSuj	sn	9
	scompl	4
	sinf	5
Arg2 Objà	à-sn	659
	à-scompl	10
	à-sinf	27
Arg2 Objde	de-sn	180
	de-scompl	0
	de-sinf	13
Obl / Obl2	de-sn	178
	par-sn	73
	contre-sn	17
	en-sn	12
	avec-sn	11
Loc	loc-sn	11
Dloc	de-sn	1

Tableau 5.7 – Répartition des arguments principaux

6 Traitement de corpus et évaluation

Le but du développement d'une métagrammaire du FMed à partir de *FRMG* (cf. partie 4) et de l'adaptation du lexique *OFrLex* (cf. partie 5) est d'annoter les textes du corpus *Profiterole* (cf. inventaire 2.6). Les analyses produites par ce système doivent cependant être d'une qualité suffisante pour être intégrées au *treebank*. Pour arriver à ce résultat, des évaluations doivent être conduites dès le développement de la métagrammaire, sur le modèle de la "programmation agile". Au fil du développement, on contrôle les gains de performance apportés par de nouveaux changements, tout en s'assurant de ne pas endommager les analyses précédemment acquises. Faire de tels tests donne aussi parfois des indications sur les prochaines modifications à apporter pour améliorer les analyses.

Une fois qu'on obtient une version exploitable de la métagrammaire et du lexique, nous pouvons procéder à une évaluation classique, avec les scores d'attachement (*UAS*, *Unlabeled Attachment Score*, soit l'attachement des tokens au bon gouverneur) et d'attachement étiqueté (*LAS*, *Labeled Attachment Score*, soit les dépendances et leur étiquette). Cette évaluation est faite sur le *SRCMF* dans un premier temps, puis sur l'ensemble *Gold* du corpus *Profiterole*, qui comprend le *SRCMF* et quelques extraits de textes de MF. Cela permet de déterminer la qualité du parseur et d'en corriger progressivement les défauts grâce aux outils de fouille d'erreur (Sagot et Villemonte de la Clergerie 2008).

Ce parseur symbolique n'est pas le seul parseur utilisé dans le cadre du projet *Profiterole*. On compte aussi *HOPS* (Grobol et Crabbé 2021) et *DyALog-SRNN* (*Shift-Reduce Neural Network*, Villemonte de la Clergerie (2014)), qui sont des parseurs neuronaux, et *Dyalog-SRNN – MetaMOF*, qui est un parseur hybride. Ces deux derniers systèmes sont développés par Eric Villemonte de la Clergerie. En l'absence de données annotées en syntaxe de dépendances pour tous les états de langue, il est difficile d'évaluer ces parseurs et d'en choisir un seul pour l'annotation du corpus *Profiterole*. Pour obtenir la meilleure annotation possible, nous cherchons à combiner les résultats de l'ensemble de ces parseurs grâce à un système de vote pondéré.

Actuellement, les parseurs symboliques font presque figure d'exception. Il convient alors de comparer les analyses produites par les différents systèmes pour déterminer ce qu'une métagrammaire et un lexique peuvent apporter et si leur coût de développement est justifié.

6.1 Evaluation et intégration

Evaluer la version courante d'un programme avant d'accepter sa mise en ligne est une pratique courante sur les plates-formes de versionnage du type *Git*. Cela permet de s'assurer que la version déposée correspond aux attentes et ne constitue pas une régression par rapport aux autres versions, c'est-à-dire une perte de la qualité précédemment acquise. L'utiliser ou continuer le développement à partir de celle-ci est alors considéré comme sûr. Le versionnage d'un programme est non seulement une sécurité pour le déve-

loppeur, mais aussi la méthode actuellement préférée pour permettre à plusieurs personnes de travailler sur le même projet.

Cette pratique a déjà été décrite pour des projets de grammaire (Butt et al. 1999). Nous nous inspirons de leurs méthodes pour ce projet. Avant de mettre en ligne une version de la métagrammaire [1], celle-ci est évaluée sur un ensemble de phrases (ex. tableau 6.1 pour tester l'ordre des constituants majeurs). Comme le FMed ne compte plus de locuteurs, les phrases de tests doivent être extraites de corpus. Pour éviter les erreurs lexicales, nous adaptons ces phrases pour utiliser des mots renseignés dans le lexique, et si c'est nécessaire, nous choisissons des mots non-ambigus. Dans une première version de ce corpus de test, nous avions pris trop de libertés dans ces adaptations, ce qui a conduit à produire des énoncés agrammaticaux (avec par exemple un sujet pronominal postposé séparé de son verbe par un autre constituant majeur). La version actuelle limite les modifications aux ellipses, aux temps des verbes et à la substitution de variations graphiques ou d'entrées partageant strictement les mêmes caractéristiques (cf. tableau 8). Elle compte 89 phrases et un vocabulaire de 371 mots pour évaluer l'ensemble des phénomènes présentés dans le chapitre 4. Nous avons testé ces phrases manuellement, en contrôlant les phénomènes souhaités. Si elles n'ont théoriquement pas d'impact sur l'analyse de la structure en question, les dépendances erronées sont ignorées. C'est la couverture syntaxique de la grammaire que nous souhaitons ainsi évaluer, et non la qualité globale de l'analyse syntaxique. Il est alors possible de repérer la sous-génération (en l'absence de l'analyse souhaitée), mais la sur-génération doit être contrôlée manuellement, en examinant les analyses proposées et en vérifiant qu'aucune n'est causée par une erreur ou un effet de bord qui doit être corrigé.

Ordre des constituants principaux	1	2	3
SOV	Li hermite	Tristan	connut
SVO	Nos	avum	dreit
VOS	perdi	alcun	Fenénne
VSO	ad	ele	colur
OVS	Tutes cestes culurs	ad	ele
OSV	l'imagine	Deus	fist

Tableau 6.1 – Exemple de phrases à tester

6.2 Evaluation de la qualité de l'analyse

6.2.1 Scores

L'annotation syntaxique automatique en dépendances est traditionnellement évaluée avec le score d'attachement *LAS*, c'est-à-dire en déterminant si les dépendances d'une phrase sont bien affectées en label et en gouverneur. Au fil des expériences, il est censé augmenter (cf. tableau 6.2). Cela ne dépend pas uniquement du développement de la métagrammaire, mais aussi celui des modifications apportées au lexique et au segmenteur *Sxpipe*. Ces éléments sont liés, et une mauvaise compréhension de la chaîne risque de supprimer la communication entre certains éléments, annulant ainsi le gain des développements récents.

Les scores actuels dans les meilleurs cas sont autour de 62% de *LAS* (cf. tableau 6.2), ce qui représente globalement une croissance de 32 points, mais reste insuffisant, l'état de l'art pour le FMed étant de

1. Le dépôt *Git* de la métagrammaire est disponible à cette adresse : https://gitlab.inria.fr/mgkit/MetaMOF.

	Janvier 2021		Mars 2021		Octobre 2021		Avril 2022	
	UAS	LAS	UAS	LAS	UAS	LAS	UAS	LAS
alexis	43,20	23,25	66,56	52,96	75,80	61,87	67,45	54,91
aucassin	52,62	33,52	72,37	61,49	78,02	**66,99**	72,21	62,16
beroul	51,54	32,20	72,63	59,78	77,40	66,17	71,51	59,52
graal	50,01	31,30	**75,02**	**63,74**	75,44	64,66	73,74	62,48
lapidaire	47,73	32,27	68,30	57,45	75,44	64,66	66,57	55,27
legier	46,54	24,87	59,69	44,64	71,14	55,05	63,69	50,00
qlr	**54,67**	**35,04**	73,50	61,63	77,84	65,84	**73,75**	**62,66**
roland	53,10	31,72	74,40	62,13	**78,55**	65,84	72,45	60,91

Tableau 6.2 – Evaluations du parseur symbolique de janvier à avril 2022
Légende :
gras : meilleurs scores

90,9% (Grobol, Regnault et al. 2022). La baisse de performances sur la dernière expérience est due à la réactivation d'entrées dans *OFrLex* et à l'ajout de vocabulaire pour traiter les nouveaux textes, entraînant ainsi une augmentation de l'ambiguïté lexicale, qui semble être un défi majeur du parseur. Évalués sur le lexique actuel, les anciens états de la métagrammaire sont légèrement moins performants que la version actuelle. Cependant, ce type de projet s'inscrit dans le temps long, et il tire profit du développement de nouvelles ressources et des nouveaux travaux linguistiques. De plus, nous disposons de nombreuses pistes pour améliorer ce parseur grâce à la fouille d'erreur. Ces résultats restent malgré tout plutôt intermédiaires, car il reste de nombreux ajustements à réaliser, notamment pour adapter la sortie du parseur au format *Universal Dependencies*. Les règles de transformation ne sont pas encore à jour des évolutions du format *UD*, ni des choix spécifiques du *SRCMF-UD*.

6.2.2 Fouille d'erreurs

Tout d'abord, il est très utile de chercher ce qui empêche l'analyse des phrases rejetées par le parseur, car fournir une analyse complète est la première étape à atteindre, même si cela ne garantit pas la qualité de l'analyse. Il peut s'agir d'une de ces erreurs :

— une construction syntaxique non décrite dans la métagrammaire

— une construction syntaxique mal décrite ou trop contrainte dans la métagrammaire

— des entrées lexicales manquantes, erronées, incomplètes dans le lexique ou dans ses *patches* dans la métagrammaire (*missing.lex*, *complete.lex*)

— une règle de segmentation manquante ou erronée

— une règle de conversion au format *Universal Dependencies* manquante ou erronée.

La longueur et la complexité des phrases a tendance à augmenter avec le temps, ce qui accroit la possibilité de faire face à une de ces erreurs. En effet, la complexité d'une *TAG* est polynomiale, et non linéaire, ce qui implique une augmentation considérable de la complexité pour les phrases longues. La couverture des textes plus récents est donc inférieure à celle des textes d'AF.

Lorsque la couverture des corpus est suffisante, c'est-à-dire lorsque la plupart des phrases reçoivent une analyse, l'interface mise en place par Sagot et Villemonte de la Clergerie (2006) est très utile, car elle met en avant les mots les plus fréquents dans les phrases n'ayant pas reçu d'analyse complète (ex.

6.1). Il faut alors interpréter ces erreurs et trouver leur source, car elles peuvent venir des outils, comme la métagrammaire et le segmenteur, ou du lexique, ou de l'interface entre ces composants. Des connaissances en FMed et une bonne compréhension des schémas d'annotation sont primordiales, ainsi que la maîtrise de la chaîne de traitement.

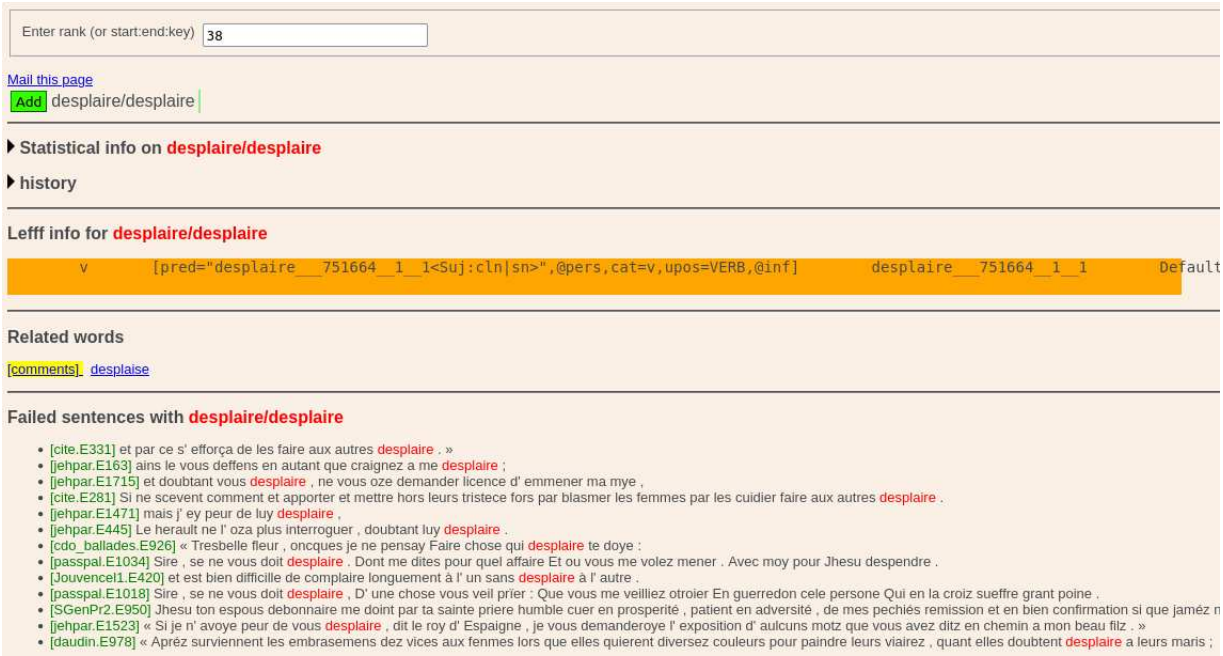


FIGURE 6.1 – Exemple de lexème causant des erreurs d'analyse

Par exemple, le lexème *desplaire* (cf. figure 6.1) empêche l'analyse de nombreuses phrases. La raison principale de ces échecs semble être son cadre de valence, qui ne comprend qu'un argument (le sujet), alors qu'il devrait être bivalenciel. Les phrases du corpus lui attribuent un objet indirect introduit par *à* (ex. *aux autres, à l'autre, a leurs maris*) ou sa forme pronominalisée (*me/te/luy/vous desplaire*). Ce verbe n'est présent que dans les nouveaux textes du corpus, son cadre de valence n'avait donc pas été extrait du *SRCMF*. Dans cette même expérience, le lexème *luy* était aussi identifié comme source d'erreur : il avait en effet été renseigné comme pronom conjoint uniquement (*cld*) et non comme pronom disjoint. Certaines formes étaient simplement absentes du lexique, comme *Guilliam*, une variation graphique de *Guillaume*. La fouille d'erreur se révèle très utile pour identifier le vocabulaire nouveau, en particulier les noms propres.

Evaluation et correction de la couverture lexicale

Les manques dans le lexique d'une part, en termes d'entrées et de traits associés, et l'ambiguïté lexicale d'autre part ont un impact fort sur l'analyse syntaxique, tant pour la qualité d'analyse que pour les temps de calcul (CARROLL et FANG 2004). Un système plus mature pourrait évaluer le lexique dans sa globalité à partir de l'analyse syntaxique seulement (TOLONE, SAGOT et VILLEMONTE DE LA CLERGERIE 2012). Toutefois, la métagrammaire du FMed n'a pas été suffisamment évaluée pour tirer des conclusions sur la couverture lexicale. Mais, pour fournir une évaluation complète et fiable du parseur, nous avons besoin d'un lexique couvrant et possédant l'essentiel des traits morphologiques, syntaxiques et sémantiques dont l'analyse syntaxique a besoin. Autrement dit, pour évaluer correctement les deux ressources, nous devons les améliorer toutes les deux, ce qui est difficile sans évaluation complète ni de l'une, ni de l'autre. Nous

ne cherchons donc pas à faire une réelle évaluation du lexique, mais à déterminer les corrections les plus urgentes à apporter dans l'inventaire des formes et dans les informations qui leur sont associées.

En amont de l'analyse syntaxique, il a été possible d'ajouter des entrées et de compléter une partie des informations manquantes grâce aux textes annotés et à un travail de lemmatisation (cf. chapitre 5). L'amélioration du lexique peut se poursuivre grâce à la fouille d'erreur en sortie de *parsing*. L'interface de Sagot et Villemonte de la Clergerie (2008) fait remonter les tokens dont le taux d'apparition dans les cas d'échec d'analyse est significatif. Ces "coupables" (Nicolas, Farré et Villemonte de la Clergerie 2007) peuvent ainsi être identifiés et corrigés. Par exemple, avant la correction du module de reconnaissance des chiffres romains dans *Sxpipe*, le principal fautif était *.i.*, dont les points signifiaient une étrange frontière de phrase.

Il s'agit ensuite de détecter deux types d'erreur : l'absence de l'entrée avec la bonne étiquette morpho-syntaxique et l'absence des bons traits morphologiques, syntaxiques ou sémantiques. Nous ne cherchons pas, pour l'instant, à réduire le nombre d'entrées en exploitant les sorties du parseur, qui peuvent indiquer si des lexèmes sont superflus ou trop ambigus. Les retours de la fouille d'erreur sont parfois difficiles à interpréter. Lors des expériences de mars 2022, le token "garde" a été identifié comme une source d'erreur fréquente. Or, les traits associés à ses entrées nominales et verbales ont été validées manuellement, car elles correspondent à ce qui est attendu dans le corpus. On peut faire le même constat pour d'autres tokens. Ces éléments apparaitraient donc fréquemment dans des phrases dont l'analyse a échoué, sans en être la cause. En revanche, il est plus aisé d'interpréter certaines erreurs, comme l'absence d'une forme comportant un signe diacritique, qui peut être rapprochée d'une entrée existante (ex. *túit* et *tuit*, pronoms indéfinis). On peut aussi repérer une erreur de conjugaison dans le lexique, comme pour *respondi*, renseigné à la première personne du singulier, mais fréquemment employé à la troisième personne (notamment dans les dialogues). Une entrée correspondante a donc été ajoutée.

Certains échecs d'analyse sont dus aux cadres de valence. Ceux-ci ont été extraits du *SRCMF*, ce qui n'a pas permis de renseigner toutes les entrées dont nous avons besoin. Cependant, nous faisons face à un autre problème : ils ont été renseignés comme des entrées du *Lefff*, qui n'autorise pas toutes les réalisations pour chaque argument. Le verbe *covenir* appelle par exemple un objet direct et un objet indirect. Ce dernier est un syntagme nominal précédé de la préposition *à* ou un pronom conjoint datif. Or, on trouve cette phrase, où la position d'objet est occupée par une infinitive, et celle d'objet indirect, par un pronom accusatif :

(1) et les covenoit departir li .i. et li autre
CONcoo PROper VERcjg VERinf DETdef PROind CONcoo DETdef PROind
'et il leur convenait de partir, aux uns et aux autres'
Queste del saint Graal (1225-1230)

Cette démarche de correction reste néanmoins manuelle. Une campagne de validation (Fort et Guillaume 2008) serait trop coûteuse à ce stade. Il serait plus rapide de procéder à une évaluation automatique du lexique, mais nous ne disposons pas de lexique de FMed dans un format équivalent (Falk, Francopoulo et Gardent 2007).

Comme l'évaluation du lexique et la qualification des erreurs est difficile, il est souvent compliqué de trouver les corrections à apporter. Il faut notamment déterminer si les erreurs viennent du lexique ou de la grammaire, et parfois décider de déplacer le traitement d'un phénomène dans l'une ou l'autre des ces composantes.

Evaluation de la couverture syntaxique

La couverture syntaxique est un premier indice de la qualité de la grammaire. Avec la fouille d'erreur, nous nous intéressons en priorité aux phrases dont l'analyse a échoué, c'est-à-dire aux phrases pour lesquelles le parseur ne propose pas d'annotation. Nous mettons de côté, dans un premier temps, la qualité de cette annotation, car elle est en partie déterminée par la désambiguïsation. Cette première étape consiste à s'assurer que les phénomènes les plus courants du FMed peuvent être traités.

Tableau 6.3 – Couverture par texte
Légende :
gras : taux de couverture supérieur à 90%
rouge : taux d'échec supérieur à 20%

Textes	nb. de ph.	Analyse complète	Echec d'analyse	Analyse robuste
AlexisRaM	529	89,41%	2,08%	8,51%
DescrEngl	171	**92,98%**	2,34%	4,68%
DialGreg1	1 050	76,86%	9,05%	14,09%
DialGreg2	1 188	74,66%	9,34%	16,00%
GuillMachFortuneH	1 748	76,14%	16,30%	7,56%
Jouvencel1	1 079	74,42%	13,35%	12,23%
Lapidfp	509	89,59%	4,13%	6,28%
QJoyesKa	2 679	80,85%	11,98%	7,17%
SBath1	571	72,33%	23,82%	3,85%
SBath2	638	76,96%	16,14%	6,90%
SBath3	88	87,50%	2,27%	10,23%
SEustPr1	638	88,40%	7,99%	3,61%
SGenPr2	962	83,16%	10,40%	6,44%
YvainKu	3 881	85,52%	9,53%	4,95%
adgar	2 112	88,59%	6,53%	4,88%
amiamil	2 852	89,55%	3,12%	7,33%
anglure	1 487	73,71%	16,61%	9,68%
aucassin	1 032	89,44%	4,07%	6,49%
baye1	767	58,93%	29,73%	11,34%
beauma1	947	63,57%	30,41%	6,02%
becket	1 989	87,23%	5,58%	7,19%
beroul	3 321	**92,05%**	3,76%	4,19%
brut2	1 952	**91,60%**	3,89%	4,51%
cambps	502	**91,04%**	1,00%	7,96%
cdo ballades	1 869	79,67%	8,13%	12,2%
cite	1 253	73,66%	15,64%	10,7%
clari	2 347	82,06%	13,08%	4,86%
coinci nca	409	79,71%	13,20%	7,09%
commyn1	1 269	74,15%	12,84%	13,01%

Suite du tableau à la page suivante

Tableau 6.3 – suite du tableau

Textes	nb. de ph.	Analyse complète	Echec d'analyse	Analyse robuste
comput	1 523	88,71%	5,91%	5,38%
daudin	1 328	76,28%	16,27%	7,45%
dictier	291	68,38%	18,56%	13,06%
eneas1	3 714	**90,44%**	4,74%	4,82%
eulaliBfm	20	75,00%	0%	25,00%
fauvel	1 575	83,24%	10,73%	6,03%
fouke	2 372	85,50%	4,60%	9,90%
gcoin1	1 442	81,28%	13,73%	4,99%
grchron9	736	73,23%	19,97%	6,80%
grchron j2c5	514	59,73%	29,18%	11,09%
hlanc	932	63,63%	17,70%	18,67%
jehpar	1 751	79,44%	10,39%	10,17%
maniere1396	1 398	80,26%	5,29%	14,45%
maniere1399	659	85,43%	2,12%	12,45%
maniere1415	273	79,49%	5,13%	15,38%
melusine	1 787	82,04%	10,74%	7,22%
menreims	1 868	87,21%	6,64%	6,15%
monstre	1 411	68,11%	18,99%	12,90%
moree	1 124	75,09%	19,04%	5,87%
passion	343	80,17%	1,75%	18,08%
passpal	1 345	87,43%	4,54%	8,03%
qgraal cm	3 117	81,10%	14,37%	4,53%
qlr	3 934	88,18%	4,55%	7,27%
quadrilogue	826	61,38%	26,03%	12,59%
quenouilles1	539	77,18%	13,36%	9,46%
roland	3 915	**95,48%**	1,25%	3,27%
rosem2	1 481	79,95%	13,77%	6,28%
sarrasin	138	76,81%	20,29%	2,90%
slethgier	189	88,89%	1,06%	10,05%
strasbBfm	3	0%	33,33%	66,67%
tdechamp	1 980	82,63%	9,80%	7,57%
tyolet	430	86,51%	7,67%	5,82%
ursins	105	74,29%	13,33%	12,38%
vergy	424	77,83%	17,45%	4,72%
Total	83 326	82,75%	9,73%	7,52%

Les taux de couverture du tableau 6.3 ont été obtenus sur l'ensemble du corpus *Profiterole*. Les textes peuvent recevoir une analyse complète, qui forme un arbre syntaxique (avec une seule tête). L'analyse robuste ne parvient pas à assembler toutes les dépendances syntaxiques en un arbre. On y trouve alors

des arbres partiels. Ce mode d'analyse peut être désactivé, mais il offre un retour sur les capacités du parseur. Lorsque aucune analyse, même partielle, ne peut être apportée, la phrase n'est pas traitée et elle est comptée dans les "échecs d'analyse".

La longueur moyenne des phrases ainsi rejetées est de 26 tokens. Dans certains cas, l'absence d'analyse est normale, notamment s'il y a une erreur de segmentation. Pour cette expérience (cf. tableau 6.3), nous avons utilisé la segmentation réalisée par *HOPS* (Grobol et Crabbé 2021), qui atteint l'état de l'art pour le FMed, mais qui produit parfois quelques erreurs. Ce parseur propose, par exemple, 149 "phrases" qui ne contiennent en réalité qu'un signe de ponctuation. D'autres ne contiennent qu'une conjonction de coordination (quatre occurrences) ou sont visiblement tronquées, comme celle-ci : "car la cité de" (*Lettre à Nicolas Arrode* de Jean Sarrasin). Les parseurs statistiques et neuronaux produisent des analyses pour toutes sortes de phrases, même tronquées, mais on peut vouloir éviter ce type de résultat.

Dans une expérience faite sur le SRCMF-UD (version 2.9), la longueur moyenne des phrases rejetées est de 25 tokens. Dans les phrases courtes, on observe la présence régulière de signes de ponctuation, qui posent effectivement un problème à l'analyse syntaxique, malgré une description exhaustive dans le lexique et la présence des arbres nécessaires dans la grammaire. Des mécanismes hérités de *FRMG* doivent permettre d'analyser des phrases contenant une ponctuation englobante sur deux (parenthèse, guillemet). Cependant, la ponctuation a été récemment réintroduite dans le *SRCMF*, et *MetaMOF* n'a pas encore été testé sur son annotation de manière intensive.

Parmi les phénomènes syntaxiques qui semblent poser un problème, on trouve les verbes introduisant une infinitive, avec une montée de la séquence de pronoms conjoints, mais sans mécanisme de contrôle. Dans la métagrammaire, nous avons ajouté une description pour la montée de cette séquence devant les modaux. Or, il semble que ce comportement ne se limite pas à ces quelques verbes, comme en témoigne cet exemple :

(2) et li compaignon **la** corent **tenir**
CONcoo DETdef NOMcom **PROper** VERcjg **VERinf**
'et les compagnons courent **la tenir**'
Queste del saint Graal (1225-1230)

Les phrases longues semblent plus difficiles à traiter, à cause de l'ambiguïté qu'elles entrainent. En effet, le nombre de gouverneurs potentiels pour chaque token augmente, allongeant le temps de calcul jusqu'à dépasser le temps accordé à chaque analyse, qui est de 300 secondes (tandis que la plupart des phrases sont analysées en moins de trois secondes). Les phrases contenant plusieurs modifieurs de phrases sont particulièrement représentées dans l'ensemble des énoncés rejetés. Des expériences sont en cours pour exploiter des modèles de relation de *left-corner* probabiliste, qui fournit un guide pour la phrase entière à partir des premiers mots.

6.2.3 Annotation morpho-syntaxique

La qualité de l'annotation morpho-syntaxique reste faible, à hauteur de 69,95% sur le *SRCMF* (cf. tableau 6.4) et de 62,88% sur les expériences intermédiaires pour l'ensemble du corpus *Profiterole* annoté par *HOPS* (cf. tableau 6.5, et version complète en annexe, cf. tableau 9). On peut expliquer une partie de ces erreurs par l'ambiguïté lexicale du FMed. Les tables de confusion contiennent, pour chaque étiquette de la colonne de gauche, les taux d'annotation par l'ensemble des étiquettes disponibles. Par

exemple, 78,4% des tokens annotés *VERB* dans le corpus de référence ont reçu la bonne étiquette lors de l'annotation, mais 6,7% ont reçu l'étiquette *NOUN* à la place.

	VERB	PUNCT	NOUN	PRON	ADV	DET	ADP	CCONJ	AUX	ADJ	SCONJ	PROPN	NUM	INTJ	Total
VERB	**78,4**	-	6,7	0,1	0,3	0,1	0,9	-	7,8	4,5	-	0,9	-	0,2	25 115
PUNCT	-	**100,0**	-	-	-	-	-	-	-	-	-	-	-	-	23 510
NOUN	3,0	-	**93,8**	0,4	0,7	0,2	0,1	-	0,2	1,0	-	0,5	-	0,1	22 167
PRON	1,2	-	5,5	**78,9**	3,1	5,1	0,6	2,1	0,5	0,4	1,4	0,7	0,4	-	20 496
ADV	2,3	-	5,8	3,4	**75,9**	1,6	3,7	3,9	0,5	0,6	1,4	0,7	0,1	-	17 250
DET	0,3	-	1,0	5,3	2,9	**87,3**	0,7	0,1	0,1	0,1	0,1	0,3	1,8	-	13 888
ADP	1,6	-	1,9	0,7	0,7	0,4	**91,3**	0,1	0,5	0,2	0,3	1,4	0,2	0,7	13 497
CCONJ	0,3	-	1,6	0,7	1,4	0,4	2,4	**88,7**	-	0,2	1,2	1,1	-	1,4	8 448
AUX	13,6	-	5,1	0,1	0,3	-	2,1	0,1	**76,8**	0,8	-	0,3	-	0,5	6 317
ADJ	5,7	-	14,0	1,6	2,4	2,1	0,3	-	0,4	**73,1**	-	0,4	0,1	-	5 362
SCONJ	0,5	-	1,4	10,8	11,6	3,2	4,3	14,2	2,2	-	**49,3**	2,3	-	-	4 506
PROPN	2,4	-	32,6	-	-	-	0,1	-	1,3	0,5	-	**62,9**	-	0,1	4 744
NUM	5,3	1,3	23,3	0,5	-	3,9	0,3	0,5	3,4	6,0	-	2,1	**53,5**	-	619
INTJ	14,5	-	-	0,8	-	-	1,6	-	-	-	-	2,4	-	**79,0**	124

TABLEAU 6.4 – Evaluation de l'analyse morpho-syntaxique sur le *SRCMF*
Légende :
gras : taux d'exactitude de l'annotation
gris : score d'exactitude inférieur à 50%
orange : taux d'erreur supérieur à 5%
rouge : taux d'erreur supérieur à 10 %

Comme en FC, *que* pose un cas d'ambiguïté entre son utilisation en tant que pronom et en tant que conjonction de subordination (URIELI et TANGUY 2013). Il en va de même pour *si*, qui peut être conjonction de subordination et adverbe intensifieur (ainsi que nom commun). On observe cependant bien plus d'ambiguïté en FMed, avec des formes qui peuvent être à la fois conjonctions de coordination et de subordination (*car*, même si c'est un cas rare), adverbe et conjonction de coordination (*mais*). Le mot *si* a non seulement les valeurs du FC, mais il peut aussi être déterminant possessif, et son variant graphique, *se*, peut être confondu avec le pronom réfléchi.

Le FMed présente de nombreux autres cas d'ambiguïté lexicale. On note, par exemple, des confusions entre nom commun et verbe. Certains lexèmes peuvent en effet appartenir aux deux catégories, comme *estre, voie* et *saut*. Faute de tri suffisant dans le lexique, des noms propres ont été confondus avec des noms communs, car ils ont été renseignés dans les deux fichiers, comme *Tristan* et *Iseut*. Les mots inconnus ont tendance à être analysés comme un verbe ou un nom, car ce sont des catégories les plus représentées dans le corpus. Lorsqu'un token inconnu commence par une majuscule (ce qui est courant en FMed à cause de la versification), un biais oriente l'analyse vers le nom propre, ce qui entraîne des erreurs d'étiquetage.

	NOUN	VERB	PUNCT	PRON	ADV	CCONJ	ADJ	AUX	PROPN	Total
NOUN		5,0								124 872
VERB	5,2							9,2		131 697
DET				6,1						78 179
ADV	5,3									82 078
ADJ	14,7	6,7								31 054
AUX	5,5	21,6								25 377
SCONJ				20,1	11,0	12,1				22 237
PROPN	25,5									23 717
NUM	17,4	7,4	10,6				9,6			2 912
INTJ		8,4							7,0	654

TABLEAU 6.5 – Confusions par partie du discours (au delà de 5%) dans le corpus *Profiterole*
Légende :
rouge : taux d'erreur supérieur à 10%

Certains noms communs font partie d'entités nommées (ex. *Table Reonde*) et ils sont annotés comme noms propres, mais nous n'en rendons pas compte avec notre parseur, car nous avons peu enrichi le module de reconnaissance d'entités nommées pour le FMed dans *Sxpipe*. On trouve aussi des confusions entre noms communs et adjectifs qualificatifs, car ces catégories partagent certains lexèmes, qui peuvent être utilisés dans des contextes similaires.

(3) *Li* **fel** jaianz, cui Dex confonde
DETdef **ADJqua** NOMcom PROrel NOMpro VERcjg
'Le traître géant, que Dieu le confonde'
Yvain de Chrétien de Troyes, v. 3 848 (1177–1181)

(4) *Li* **fel** de coi nos nos pleignons S' en alast come desconfiz
DETdef **NOMcom** PRE PROrel PROper PROper VERcjg PROper PROadv VERcjg CONsub VERppe
'Le traître dont nous nous pleignons se retirerait en déroute'
Yvain de Chrétien de Troyes, v. 3 234 (1177–1181), trad. Geneviève Joly

Nous remarquons également qu'il reste des corrections à apporter dans le segmenteur *Sxpipe* pour la description des chiffres romains, en particulier pour ceux qui contiennent des tirets bas (ex. *._xvi_.*) et ceux dont le suffixe ordinal *-e* est placé avant le point (ex. *.VIJe.*).

Les verbes et les auxiliaires sont parfois confondus, notamment les verbes qui ont le plus souvent un statut d'auxiliaire modal, mais peuvent aussi fonctionner comme de simples verbes (ex. *devoir, vouloir*). Ce choix peut être aussi difficile en FC, car les modaux ont généralement plusieurs entrées, pour les cadres de valence classiques et pour les cadres avec un contrôle, qui sont des auxiliaires dans le format *Universal Dependencies*. En FC, les participes passés ont une forme reconnaissable, même s'ils sont parfois annotés à tort en tant qu'adjectifs. En FMed, ces formes peuvent être confondues avec d'autres parties du discours, notamment les interjections et les noms communs, même si la fréquence de cette dernière confusion est plus faible. On trouve donc de nombreuses erreurs d'annotation pour les auxiliaires de temps, *estre* et *avoir*.

La faible qualité de l'annotation morpho-syntaxique s'explique donc en grande partie par des erreurs de description et de désambiguïsation, mais c'est aussi la nature ambiguë du lexique du FMed qui rend cette tâche difficile.

6.2.4 Annotation syntaxique

Les scores de l'analyse syntaxique sont calculés sur les phrases ayant reçu une analyse. Comme une seule phrase des *Serments de Strasbourg* (sur trois) a été analysée, l'écart entre les évaluations des analyses morpho-syntaxique et syntaxique est considérable. Tous les scores sont tous améliorables, ce qui demande un travail sur les différents composants de la chaîne. Certaines dépendances semblent plus faciles à obtenir que d'autres, notamment les dépendances courtes comme *det* (déterminant) et *case* (l'attachement de la préposition). La reconnaissance des arguments du verbe est, en revanche, plus difficile à obtenir. Le sujet et l'objet sont fréquemment confondus, mais ils peuvent être analysés comme d'autres éléments de la phrase, notamment des modifieurs. La reconnaissance totale des arguments du verbe reste encore très faible (cf. tableau 6.7). Pour les phrases du *SRCMF* comprenant à la fois un sujet et un objet direct pour le verbe, quelle que soit leur réalisation (y compris les pronoms conjoints), nous avons cherché la

Textes	POS	UAS	LAS
alexis	68,10	67,45	54,91
aucassin	71,99	72,21	**62,16**
beroul	73,03	71,51	59,52
graal	67,34	**73,74**	**62,48**
lapidaire	68,20	66,57	55,27
legier	61,62	63,69	50,00
qlr	72,74	**73,65**	**62,66**
roland	**74,79**	72,45	60,91
strasbourg	12,98	66,67	61,11
yvain	66,68	68,40	56,21

Tableau 6.6 – Résultats d'analyse sur le *SRCMF*
Légende :
gras : meilleurs scores (différence inférieure à 0,5 points)

Ordre	SRCMF	Parsing	
	Occ.	Occ.	%
SOV	8 443	626	7,41
SVO	6 037	190	3,15
OSV	2 476	68	2,75
OVS	1 882	79	4,20
VSO	720	36	5,00
VOS	223	18	8,07

Tableau 6.7 – Analyse complète des différents ordres

proportion de phrases pour lesquelles les deux dépendants était correctement annotée, car cela fixe des repères importants pour la phrase. Parmi les erreurs, on trouve des échecs d'analyse et une confusion avec d'autres arguments du verbe ou modifieurs.

La reconnaissance de la tête de phrase peut également être améliorée. On remarque que le parseur a des difficultés à différencier les propositions principales des subordonnées (cf. confusions entre *xcomp* et *root* d'une part et *root* et *acl* d'autre part).

6.3 Comparaison avec les autres parseurs

6.3.1 Parseurs neuronaux

FRMG n'est pas le seul outil développé pour le FC qui peut être adapté au FMed. Il est possible d'utiliser des modèles neuronaux du FC et de les "post-entraîner" sur des données brutes du FMed (Grobol, Regnault et al. 2022), c'est-à-dire qu'on adapte légèrement les poids du modèle avec des plongements lexicaux appris sur corpus, en faisant l'hypothèse que le modèle initial de FC est un bon guide pour les données de FMed. On obtient ainsi jusqu'à 90,90% de LAS sur le *SRCMF* avec le modèle *Camembert* (L. Martin et al. 2020).

Le parseur utilisé dans cette première expérience, *HOPS*, "combine à la fois un étiqueteur morpho-syntaxique, un analyseur basé sur l'algorithme de Dozat et Manning (2017) et l'utilisation de riches représentations lexicales (Bojanowski et al. 2017 ; Devlin et al. 2019)" (Grobol et Crabbé 2021). D'après

Grobol, Prévost et Crabbé (2022), les différences du FC et du FMed sont peu marquées. Ce parseur est moins performant sur l'analyse des constituants principaux en FMed, notamment à cause de la variation de l'ordre des mots et des différences de graphie, mais il est meilleur sur d'autres attachements, comme les modifieurs.

	nsubj	obl	root	obj	iobj	xcomp	cop	ccomp	expl	csubj	Total
nsubj	**66,4**	0,4	1,3	1,9	0,2	0,2	-	-	0,1	-	52 781
obl	1,0	**45,3**	1,5	1,7	0,2	0,3	-	0,1	-	-	51 545
root	2,3	0,8	**76,8**	1,3	-	2,6	0,3	0,6	-	0,2	61 446
obj	1,9	1,2	1,4	**54,7**	0,3	0,3	-	0,1	-	-	45 281
iobj	0,4	1,7	0,5	1,8	**54,4**	0,1	-	-	-	-	9 991
xcomp	0,9	0,8	7,0	1,0	-	**40,0**	-	0,2	-	0,5	8 561
cop	0,6	0,7	22,8	0,9	-	1,1	**37,3**	2,2	-	0,1	6 935
ccomp	1,5	0,4	5,7	1,5	-	4,2	0,3	**32,5**	-	0,3	4 974
expl	4,2	0,1	1,1	1,5	0,3	-	-	0,1	**0,3**	-	4 580
csubj	1,0	0,8	19,7	0,8	-	6,8	0,1	3,7	-	**2,0**	732

Tableau 6.8 – Evaluation de *MetaMOF* en prenant les sorties d'*HOPS* comme étalon : le verbe et de ses arguments
Légende :
gras : taux d'exactitude de l'annotation
gris : score d'exactitude inférieur à 50%
orange : taux d'erreur supérieur à 5%
rouge : taux d'erreur supérieur à 10 %

Nous prenons les sorties du parseur *HOPS* comme référence (cf. tableau 6.8) pour avoir une première estimation de la qualité d'analyse de *MetaMOF* sur les nouveaux textes et pour comparer ces deux parseurs. Les scores sont semblables à ceux obtenus sur le *SRCMF*. On retrouve la difficulté de notre parseur à identifier les arguments du verbe, et les confusions entre propositions principales et subordonnées. Cependant, il n'est pas toujours souhaitable de chercher à aligner sur celles d'*HOPS*, car il n'a pas de guidage linguistique, ce qui l'amène parfois à attribuer deux sujets au verbe, notamment un *nsubj* et un *csubj* (129 occurrences) alors qu'il s'agit d'un autre argument.

Dyalog-SRNN est un parseur avec une couche statistique par transitions, guidé par une couche neuronale de type Dozat (biaffine + *MST*) (Villemonte de la Clergerie 2014) qui peut prendre en entrée des treillis de mots, comme *FRMG*, mais cette option n'est pas utilisée actuellement. Cet outil peut donc exploiter les informations d'une première analyse morpho-syntaxique, voire celles d'un lexique comme le *Lefff* pour le FC. Le parseur peut aussi prendre en entrée des traits de *clustering* appris sur de larges corpus.

Dans son traitement du FMed, *Dyalog-SRNN* atteint 73,90% de *LAS* (version d'octobre 2021). Comme pour *HOPS*, la difficulté principale semble être d'identifier les arguments du verbe. Les différents arguments sont souvent confondus, comme le sujet nominal (*nsubj*) et l'objet nominal (*obj*), notamment à cause de la souplesse de l'ordre des constituants majeurs. On trouve aussi des erreurs attendues en FC, comme la confusion entre les pronoms personnels (*nsubj, obj*) et impersonnels (*expl*), de forme parfois identique, mais sémantiquement vides.

6.3.2 Parseur hybride

Pour le FC, le parseur hybride *Dyalog-SRNN – FRMG* tire profit des analyses de *FRMG*, ce qui permet de pallier certaines faiblesses du parseur neuronal, comme la reconnaissance de dépendances à longue

distance. Le parseur hybride est même meilleur que la métagrammaire sur l'annotation de certaines dépendances, en particulier celles qui induisent de la profondeur (Villemonte de la Clergerie 2014).

Pour le FMed, les sorties de *MetaMOF*, converties en *Universal Dependencies*, sont fournies comme trait à *Dyalog-SRNN*, qui apprend à les utiliser, ou non, pour prendre ses décisions. Ce parseur atteint 74,78% de *LAS* dans les expériences d'octobre 2021. Cela représente un gain de performance par rapport à *Dyalog-SRNN* seul. L'analyse syntaxique reste cependant à améliorer, notamment avec les mises à jour de la chaîne du parseur symbolique (*Sxpipe, OFrLex, MetaMOF* et interface entre le lexique et la métagrammaire).

Une métagrammaire et son entourage (le lexique, le segmenteur, le module d'interface entre grammaire et lexique et les scripts d'adaptation du format de sortie) se développent sur le temps long, ce qui handicape un projet d'annotation et d'exploitation d'un vaste corpus comme *Profiterole*. Les méthodes neuronales sont plus rapides à adapter et fournissent des résultats comparables à l'état de l'art sur des langues plus étudiées et moins soumises à la variation, comme le FC (et pour lesquelles on dispose de corpus arborés). En revanche, une métagrammaire peut servir de guide, car ses analyses syntaxiques et morpho-syntaxiques sont motivées linguistiquement. Malgré ses lacunes, *MetaMOF* semble aider le parseur *Dyalog-SRNN*. Ce rôle est amené à se développer au fil des mises à jour. Dans un système de vote entre parseurs, un parseur symbolique apporte une voix complémentaire, qui peut approuver l'annotation des autres parseurs (ex. tableau 6.8) ou, au contraire, prévenir des éventuelles lacunes, comme une mauvaise segmentation ou des conflits linguistiques, comme la double reconnaissance de sujets. Dans un premier temps, *MetaMOF* peut donc servir d'auxiliaire aux parseurs existants, mais son ambition reste de continuer à améliorer ses analyses, en exploitant les retours sur corpus, une fois que l'adaptation de base a été réalisée.

7 Discussion et perspectives

7.1 Réutilisabilité des outils

Dans une optique de sciences ouvertes, les outils et les ressources développés dans le cadre du projet *ANR Profiterole* seront disponibles et libres de droit. Il est important que la métagrammaire, le lexique et le corpus arboré puissent être facilement utilisables.

7.1.1 Partage du parseur

Déploiement

L'installation de la chaîne de traitement doit être simple et documentée. Un conteneur *Docker* [1] a été développé à cette fin. Il fournit l'environnement nécessaire au fonctionnement de la chaîne, la rendant suffisamment étanche pour éviter les conflits de versions de logiciels. Une autre solution consiste à créer une machine virtuelle. Cependant, un *Docker* consomme moins de ressources et reste fiable et simple à utiliser.

La métagrammaire et les autres composants de la chaîne sont aussi disponibles sur *Gitlab*. Ces dépôts doivent être maintenus et permettre aux utilisateurs de faire des retours. L'outil de versionnage *git* est intéressant pour cela, car il permet d'ouvrir des branches de développement et de faire remonter des problèmes et des questions de manière publique. Un dialogue entre développeurs peut alors avoir lieu via le système de tickets, et ainsi être archivé. La fonctionnalité *wiki* permet d'héberger une documentation complète.

Documentations

Un guide à l'usage des développeurs de métagrammaires au format SMG est en cours de rédaction pour enrichir la documentation existante. Il contient la description de la chaîne de traitement, des scripts et des fichiers qui la composent, ainsi que des instructions pour les modifier. Un descriptif de la syntaxe du langage *SMG* est également présent, et complète la documentation de *FRMG* [2].

La documentation doit aussi s'adresser aux linguistes. Le formalisme des *TAG* est efficace pour traiter des langues naturelles, mais il impose des représentations parfois éloignées des descriptions linguistiques. La documentation doit justifier ces choix. Il est aussi nécessaire de dresser un inventaire des phénomènes traités par la métagrammaire pour expliquer le taux de couverture syntaxique.

1. Ce Docker est disponible au téléchargement à cette adresse : https://gitlab.inria.fr/almanach/docker_webservices.
2. La documentation des classes de FRMG est disponible à cette adresse : http://alpage.inria.fr/frmgwiki/wiki/pr%C3%A9sentation-des-m%C3%A9ta-grammaires#ref2.

Communauté d'utilisateurs

Il est important de rejoindre une communauté d'utilisateurs, pour communiquer sur les formalismes à base de métagrammaires et développer de bonnes pratiques, mais aussi pour faire utiliser son outil. D'une part, il y a un intérêt scientifique à partager ses travaux, a fortiori s'ils concernent des méthodes qui se développent sur le temps long, afin d'avoir des retours sur ses derniers développements. D'autre part, parvenir à faire utiliser un outil, c'est le faire perdurer. Un ensemble d'éléments peuvent contribuer à faciliter la prise de contact, comme une documentation accompagnée de tutoriels, un système facilitant l'installation ou un script de coloration syntaxique pour un éditeur populaire, comme *Visual Studio Code* [3]. Le système de tickets et de commentaires disponible sur *Git* a été peu utilisé pour ce projet, mais il permet à tous les utilisateurs de faire des retours sur le système et de suggérer des changements, qu'ils maintiennent eux-mêmes le programme ou non.

Dans le cadre du groupe de recherche *Déméter* [4], soutenu par le GDR *LIFT*, des discussions ont eu lieu à ce sujet. Face à la variété des systèmes à base de métagrammaire et aux besoins de la communauté des linguistes de terrain, une entreprise de mutualisation des ressources a été mise à l'étude. Un projet de plate-forme permettant la constitution, le stockage et l'exploitation de grammaires et de ressources lexicales a été ébauché, et il fait partie des pistes pour poursuivre le travail sur *MetaMOF*.

7.1.2 Partage du lexique

Outre les fichiers utilisés pour l'analyse syntaxique, il est utile de partager le lexique dans un format tabulaire, avec toutes les formes fléchies et l'ensemble des informations récoltées. Pour cela, un travail a été réalisé avec Cristina Holgado, afin de rassembler les variations graphiques sous des lemmes communs. Ces fichiers seront disponibles à partir de juin 2022 sur un dépôt *Git* indépendant.

7.2 Traitement automatique de corpus hétérogènes

La première intuition de cette thèse était d'adapter la métagrammaire *FRMG* en un système diachronique pour le FMed. Cela permet de limiter certaines analyses à des sous-ensembles de textes d'un corpus, en se basant sur leurs métadonnées. C'est ce que propose le projet *XLE* pour les textes journalistiques d'une part, et les textes techniques d'autre part (Kaplan, King et Maxwell III 2002). *FRMG* a déjà été adapté pour traiter, par exemple, des écrits de botaniques (Role, Gavilanes et Villemonte de la Clergerie 2007) et des archives industrielles des 17e et 18e siècles (Chagué et al. 2019). Nous avons envisagé de limiter certaines analyses à des états de langue précis, comme l'ordre *datif – accusatif* pour les pronoms conjoints en MF (cf. partie 4.3.3). Cependant, parmi les tendances indiquées dans les grammaires du FMed, aucune n'était adaptée à ce système de filtre : toutes décrivent des frontières floues pour les phénomènes syntaxiques, ce qui rend ce type de contrainte difficile à implémenter. Cette approche conduit en fait à considérer un corpus hétérogène comme un ensemble regroupant différents sous-ensembles homogènes. Cela revient à faire l'hypothèse d'un "ordre" sous-jacent en FMed, constitué d'états langagiers identifiables aux caractéristiques définies et régulières. Or, il semble qu'une telle organisation n'existe pas, du moins pour la syntaxe. La variation dialectale a peu d'impact sur cette dernière (Heiden et Lavrentiev

3. Ceci a été mis en place en collaboration avec Yoann Dupont. Le script est disponible à cette adresse : https://gitlab.inria.fr/mgkit/syntax_hightlighting_smg/-/tree/dev
4. Le site Internet de présentation est disponible à cette adresse : https://demeter.inria.fr/fr/.

2004) et les phénomènes syntaxiques ne sont pas exclusifs de certaines époques, de certains domaines ou de certaines formes de textes. Au contraire, la syntaxe du FMed forme un *continuum* où des structures très diverses cohabitent dans des proportions variables. Les constructions apparaissent à des fréquences différentes selon les textes, mais elles ne peuvent pas être limitées à un type de texte particulier. Sans l'admettre, l'ambition d'une métagrammaire diachronique était de "diviser pour régner", mais l'hétérogénéité du FMed a résisté à cette simplification, reléguant le choix de l'analyse finale au module de désambiguïsation. En revanche, la métagrammaire permet assez facilement de couvrir l'ensemble du spectre des variations possibles sur une construction syntaxique, même si cela peut conduire à une surgénération d'analyses.

On peut en revanche envisager d'ajouter des traits dialectaux et diachroniques aux entrées du lexique. Cela permettrait d'enrichir les requêtes sur corpus, mais aussi de limiter la consultation de certaines entrées lors de l'analyse syntaxique. Nous souhaitons aussi explorer une deuxième piste d'amélioration du côté de la désambiguïsation. Actuellement, un seul modèle de désambiguïsation est utilisé pour l'ensemble des textes, mais nous n'avons pas fait d'expériences pour déterminer s'il est possible d'en entraîner plusieurs, ni s'ils apporteraient un gain en qualité, ou si un tel module ne suffit pas à gérer l'hétérogénéité d'un tel corpus. Il serait alors préférable d'exploiter la métagrammaire au sein d'un système hybride comme *Dyalog-SRNN – MetaMOF* pour produire des analyses pertinentes. Il est aussi possible d'ajouter des traits de désambiguïsation liés à des métadonnées comme l'époque, le dialecte et le domaine, mais cela suppose un volume minimum de données annotées pour apprendre un modèle. Actuellement, ce modèle n'est entraîné que sur les textes du *SRCMF* et sur quelques textes de MF annotés manuellement.

7.3 Utilisation des systèmes symboliques

Les performances actuelles des parseurs neuronaux sont souvent remarquables, y compris sur des langues peu dotées, notamment grâce aux techniques par transfert (Scrivner et Kübler 2012) et à l'utilisation de plongements contextuels. L'annotation du FMed n'échappe pas à cette tendance (Grobol, Prévost et Crabbé 2022). Néanmoins, il semble que les méthodes neuronales gagnent à intégrer des ressources linguistiques, comme le parseur hybride *Dyalog SRN – MetaMOF*, qui consulte les prédictions de *MetaMOF*, guidées par des descriptions linguistiques. D'un point de vue général, il est utile de laisser de la place à différentes directions de recherche (Church 2011). Les systèmes hybrides sont encore peu utilisés, mais ils permettent d'informer un modèle performant et très optimisé avec des contraintes linguistiques. Cela s'avère utile pour traiter des états de langue fortement soumis à la variation et présentant de nombreux cas d'ambiguïté, ou encore en cas d'erreur de segmentation.

De plus, malgré son temps de développement long, l'utilisation d'un système symbolique semble plus économe en électricité. Cette dimension est désormais prise en compte dans l'entrainement de modèles, et elle fait l'objet de recommandations par les communautés de chercheurs [5]. L'utilisation de parseurs symboliques, seuls ou en tant que ressource pour d'autres parseurs, peut être une piste pour certaines langues.

5. Les recommandations de la communauté *ACL* sont disponibles à cette adresse : https://www.aclweb.org/portal/sites/default/files/Efficient%20NLP%20policy%20document%20full%20document.pdf?tpcc=nleyeonai.

7.3.1 Études linguistiques

Cependant, l'annotation de vastes corpus n'est pas la seule finalité des parseurs symboliques. Ils permettent en effet d'accéder à différents niveaux d'analyse et d'explorer des hypothèses sur des ensembles de phrases, ce qui les rend utiles à plusieurs égards.

Premièrement, une métagrammaire permet de développer et de modifier tout ou partie d'une construction syntaxique, puis d'analyser des phrases à partir de la grammaire générée. La première étape met à l'épreuve les connaissances du linguiste sur une langue, car il faut fournir une description dans la métagrammaire qui permette de générer les arbres désirés. Ceci ne concerne pas seulement des phénomènes syntaxiques isolés, mais tout un système qui décrit des sous-arbres dont la structure et les propriétés doivent leur permettre de s'assembler en arbres *TAG*. Les choix de description doivent donc rester cohérents et minimaux pour faciliter la tâche de développement dans le temps. En outre, les contraintes doivent être suffisamment précises pour permettre d'insérer les éléments voulus à des emplacements adéquats dans les arbres. Pour *TAG*, il faut prendre en compte l'ordre linéaire et la cooccurrence prédicat-argument. Dans les travaux de linguistique, les descriptions ne sont pas faites selon ces contraintes. Il faut donc les adapter. La métagrammaire devient alors une représentation de la compétence linguistique, et elle peut servir à illustrer des grammaires traditionnelles. Elle peut aussi aider le linguiste à préciser ces dernières, grâce à l'examen des arbres générés. Ceux-ci ne doivent pas contenir de structures agrammaticales. S'en assurer n'est pas entièrement possible pour les états de langue anciens comme le FMed. Cependant, avec l'évaluation sur corpus arboré, on peut identifier les arbres qui provoquent fréquemment des erreurs d'analyse.

Deuxièmement, les systèmes symboliques donnent plus d'informations sur l'analyse des phrases que l'annotation finale pour constituer le corpus arboré. La forêt de dérivation nous informe, par exemple, sur les arbres (ou les règles) utilisés, la nature des opérations et les entrées lexicales sélectionnées. Cela permet de justifier les annotations fournies dans un corpus arboré (Bladier et al. 2022). Avoir accès à ces compléments aide à l'examen de phénomènes syntaxiques et enrichit les possibilités de requêtes.

Troisièmement, développer une métagrammaire impose une nouvelle façon de raisonner sur des états de langue. En se plaçant du point de vue d'un formalisme, on décrit la langue d'une manière nécessairement "rigide", car on cherche ce qui est régulier dans la langue. Cela nous conduit à poser des questions différentes, comme lorsque nous avons voulu déterminer l'ordre exact d'apparition des pronoms conjoints, et si le changement d'ordre des pronoms clitiques accusatifs et datifs était assez significatif pour être implémenté. Nous avons aussi cherché à attribuer certaines descriptions syntaxiques à des états de langue précis (sans succès), ce qui a permis de pousser à l'extrême l'idée que la syntaxe varie selon les états de langue. Les fréquences d'attestation des phénomènes syntaxiques selon des sous-ensembles de textes restent à observer précisément, mais notre expérience semble prouver que des frontières étanches ne peuvent pas être établies dans ce *continuum* de langue.

Plus généralement, le fait d'adapter *FRMG*, et non de développer un système indépendant, nous a amenés à réfléchir à l'évolution de la langue. Chaque phénomène du français a été étudié pour déterminer les changements à effectuer dans la métagrammaire. Certaines évolutions sont très bien identifiées, comme celle de l'ordre des constituants majeurs. D'autres ont demandé un examen minutieux du corpus, comme l'ordre des pronoms conjoints et la montée de ces pronoms devant un modal. Certains éléments grammaticalisés en FC ne le sont pas en FMed, comme les conjonctions de subordination composées de plusieurs tokens. En FC, elles sont répertoriées dans le lexique (ex. *parce que, bien que*). En FMed, ces structures ne

sont pas figées et elles sont analysées syntaxiquement (cf. partie 4.6.3). Les locutions de forme *préposition + (ce) + conjonction* reçoivent une nouvelle analyse, et celles qui sont de forme *adverbe + conjonction* sont analysées par un arbre hérité de *FRMG*, qui traite les modifieurs précédant les conjonctions de subordination (ex. *C'est vraiment parce que tu y tiens*). Parfois, on trouve des phénomènes similaires en FMed et en FC, bien qu'ils répondent à des motivations différentes. C'est le cas de la complétive non-introduite par une conjonction de subordination, attestée en AF, et qui s'observe parfois aussi en FC, en particulier à l'oral. Certaines constructions présentent des similitudes, mais elles reçoivent des traitements différents. Par exemple, le complément déterminatif en FMed n'est pas nécessairement introduit, et il peut se placer à gauche et à droite du nom qu'il modifie. En FC, cet ordre libre laisse place à un complément nécessairement postposé. Il peut apparaître sans préposition dans certains contextes (ex. *un département recherche, un développeur logiciel embarqué*). Ces constructions semblent proches, mais elles sont décrites différemment (cf. partie 4.4.1). L'adaptation d'un système existant implique donc d'évaluer les changements entre les états de langue, pour déterminer s'il faut réutiliser une structure du système initial, ou s'il faut en décrire une nouvelle, ce qui mène à s'interroger différemment sur l'évolution de la langue.

7.3.2 Étudier l'hétérogénéité des données

Le but de la métagrammaire est de générer une grammaire qui peut analyser tous les états de langue du français médiéval. Elle ne prend pas en compte, à ce jour, la diversité des textes. C'est à partir des sorties d'analyse syntaxique qu'on peut étudier l'hétérogénéité des données. Il est en effet connu que l'évolution de la langue, les dialectes et les genres ont un impact sur la syntaxe, et cela peut être en partie quantifié par des études sur corpus (annotés en syntaxe ou non), mais l'utilisation d'un système symbolique peut apporter des pistes supplémentaires. En effet, un corpus arboré dont l'annotation a été vérifiée permet de faire des requêtes et d'extraire des graphes, ce qui fournit au linguiste un corpus d'exemples, en vérifiant toutefois que les analyses correspondent à ce qui était attendu. Cependant, un système symbolique permet de faire des requêtes directement dans l'arbre de dérivation, en cherchant par exemple des arbres particuliers de la grammaire ou des phénomènes syntaxiques tels qu'ils sont décrits dans une classe. Par exemple, on peut chercher la distribution des différents types de comparatives à travers le corpus *Profiterole*. Nous gardons, dans le tableau 7.1, les cas où la partie gauche de la comparative (*plus/tant/si...*) est modifiée par la deuxième partie qui contient le comparant (*que...*) [6].

A priori, la syntaxe des phrases devient plus complexe avec le temps, en particulier en MF (cf. parties 1.2.4 et 1.2.5). On s'attend donc à voir le nombre de comparatives augmenter, et cette tendance semble se confirmer (cf. figure 7.1). D'après les analyses de MetaMOF, cette construction fait figure d'exception en TAF, et peu de subordonnées n'ont pas de verbe conjugué. Certains textes montrent des différences par rapport au reste du corpus :

— *Saint Alexis* est le seul texte de TAF dans lequel on trouve cette construction.

— Certains textes semblent contenir des innovations :

  — *Yvain* : *plus ADJ que ADJ* :

6. Nous avons recherché, dans les arbres de dérivation, le nom des classes qui décrivent ce phénomène : la première partie est décrite dans des classes dont le nom commence par "comparative as" puis donne la réalisation (adjectif, groupe verbal...), et la deuxième partie, dans des classes qui commencent par "mod on comparative" et donnent la réalisation.

plus... \que...		adj	adv	PP	N2	S
adj	taf	0	0	0	0	1
	af	1	0	1	4	87
	mf	1	0	1	14	103
adv	taf	0	0	0	0	0
	af	0	0	1	4	56
	mf	0	1	3	4	70
N2	taf	0	0	0	0	0
	af	0	0	0	0	18
	mf	0	0	0	0	16
PP	taf	0	0	0	0	0
	af	0	0	1	0	36
	mf	0	0	1	2	20
anteadj	taf	0	0	0	0	0
	af	0	0	1	1	28
	mf	0	0	1	1	50
vmod	taf	0	0	0	1	1
	af	0	0	12	31	344
	mf	0	1	10	16	201
v	taf	0	0	0	0	0
	af	0	0	0	1	24
	mf	0	0	0	4	14

Tableau 7.1 – Structures en deux partie : *plus/tant...* (horizontal) *que...* (vertical)

(1) Je l' apel plus **malvés** que **preu**
PROper PROper VERcjg ADVgen **ADJqua** CONsub **ADJqua**
'Je l'appelle plus mauvais que preux.'
Yvain, v. 1 324 (ca. 1177-1181)

— *Coutumes de Beauvaisis* : *plus ADJ que PP*

(2) Et meismement avarice hebergiee en cuer de baillif est
CONcoo ADVgen NOMcom VERppe PRE NOMcom PRE NOMcom VERcjg
plus **mauvese** et plus perilleuse qu' **en autre gent**
ADVgen **ADJqua** CONcoo ADVgen ADJqua CONsub **PRE ADJqua NOMcom**
'Et, de même, l'avarice qui réside dans le coeur du bailli est plus mauvaise et plus périlleuse que chez d'autres gens.'
Coutumes de Beauvaisis de Philippe de Beaumanoir, p. 21 (1283)

— *Comput* : *plus PP que PP*

(3) Plus **el mais de feverer** Quë **el**
ADVgen **PRE.DETdef NOMcom PRE NOMcom** CONsub **PRE.DETdef**
mais de jenver
NOMcom PRE NOMcom
'Plus au mois de février qu'au mois de janvier'
Comput de Philippe de Thaon, v. 2 127 (1113)

— *Quinze Joyes de mariage* : ce texte, parmi les plus récents du corpus, présente une plus grande variété que les autres (adv/adv, vmod/adv, PP/PP), par exemple :

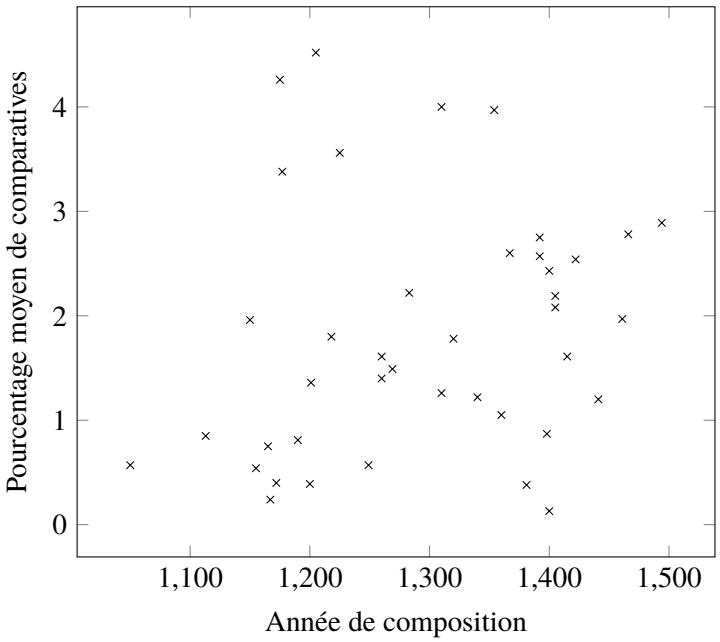


FIGURE 7.1 – Évolution globale de l'utilisation des comparatives (chaque point représente un texte). Coefficient de Pearson : 0,2.

(4) si le croit plus **fermement** que **davant**
ADVgen PROper VERcjg ADVgen **ADVgen** CONsub **ADVgen**
'[Il] le croit plus fermement qu'avant.'
Quinze Joyes de mariage, p. 118v (ca. 1400)

Les types de comparatives les plus fréquentes ne sont pas, pour autant, réparties uniformément dans le corpus, ni même dans des sous-ensembles. Par exemple, il n'y a pas de corrélation stricte entre l'époque et le taux d'apparition de ce type de subordonnée, bien que la syntaxe devienne globalement plus complexe. De même, le genre littéraire semble privilégié (cf. tableaux 7.2 et 7.3), mais cela ne suffit pas à garantir une forte fréquence de ces constructions. La longueur médiane des phrases n'offre pas non plus d'indication sur le nombre d'occurrences à anticiper. Par exemple, *Yvain* est le texte d'AF qui présente le plus de comparatives corrélatives, mais ses phrases sont généralement moins longues que celles de *Vie de Sainte Bathilde* ou du *Roman de la rose*.

A posteriori de l'analyse syntaxique, il semble possible de proposer des critères pour qualifier l'hétérogénéité d'un corpus, notamment un inventaire de constructions particulières. Elles placent certains textes aux marges, en dehors de la compréhension habituelle des états de langue. Ces phénomènes plus rares sont parfois difficiles à obtenir lors d'une analyse automatique, soit, pour un système statistique, parce qu'ils sont absents du corpus d'apprentissage, soit, pour un système symbolique, parce qu'ils ont échappé à la vigilance des développeurs. Pour les étudier, il est possible de les extraire dans des corpus arborés, mais cela ne donne pas accès aux arbres de dérivation, qui rendent plus claires les analyses en indiquant quelles classes de la métagrammaire ont été utilisées.

On peut aussi intégrer la morphologie aux recherches sur l'hétérogénéité du corpus, notamment en cherchant à déterminer l'empreinte dialectale des textes à partir des entrées lexicales sélectionnées. Par exemple, le pronom conjoint *le* au féminin (cf. tableau 5.2) est utilisé dans le dialecte picard, mais tous

les textes écrits dans ce dialecte ne présentent pas nécessairement cette particularité. La graphie des mots est aussi un indice, et on peut quantifier les marques d'un dialecte dans un texte.

Développer une métagrammaire pour une langue permet de représenter la compétence qu'on en a acquise, de mettre celle-ci à l'épreuve de l'analyse syntaxique, d'annoter des corpus et d'y faire des recherches précises. Même si la métagrammaire est encore en développement, comme c'est le cas pour *MetaMOF*, les résultats de ses analyses peuvent être exploités pour étudier des états de langue.

Siècle	Textes	Nb. de comp.	Nb. mots	Nb. de ph.	Moyenne dans ph.	Longueur médiane ph.	Genre
12	*Yvain*	131	47 995	3 881	3,38 %	11	litt
13	*Queste Graal*	111	44 829	3 117	3,56 %	12	litt
13	*Clari*	106	38 088	2 347	4,52 %	13	hist
13	*Aucassin*	44	11 679	1 032	4,26 %	9	litt
13	*Quatre Livres des Rois*	32	45 131	3 934	0,81 %	8	rel
13	*Récit ménestrel de Reims*	30	22 543	1 868	1,61 %	10	hist
13	*Chansons* Th. Champ.	27	25 393	1 980	1,36 %	11	litt
13	*Miracles* de Coinci	26	19 794	1 442	1,80 %	11	rel
12	*Tristan*, Béroul	25	32 800	3 321	0,75 %	8	litt
13	*Roman de la rose*	22	22 518	1 481	1,49 %	12	did
13	*Coutumes Beauvaisis*	21	22 637	947	2,22 %	21	jur
12	*Eneas*	20	40 547	3 714	0,54 %	8	litt
12	*Comput*	13	16 711	1 523	0,85 %	9	litt
13	*Ami et Amile*	11	29 946	2 852	0,39 %	8	litt
12	*Lapidaire*	10	5 516	509	1,96 %	9	did
13	*Bathilde 1*	8	11 175	571	1,40 %	17	rel
12	*Becket*	8	22 556	1 989	0,40 %	10	rel
13	*Coll. miracles* Adgar	5	24 132	2 112	0,24 %	9	rel
13	*Lettre* J. Sarrasin	3	2 619	138	2,17 %	16	hist
11	*Saint Alexis*	3	5 536	529	0,57 %	8	rel

Tableau 7.2 – Comparatives en ancien français

Siècle	Textes	Nb. de comp.	Nb. mots	Nb. de ph.	Moyenne dans ph.	Longueur médiane ph.	Genre
15	*Quinze Joyes de mariage*	65	39 448	2 679	2,43 %	12	litt
14	fauvel	63	21 775	1 575	4,00 %	12	did
15	jehpar	53	28 967	1 836	2,89 %	14	litt
14	*Mélusine*	46	26 661	1 787	2,57 %	11	litt
14	*Livre seyntz medicines*	37	21 861	932	3,97 %	18	rel
14	*Fouke le Fitz Waryn*	30	29 427	2 372	1,26 %	10	litt
15	*Ballades* Ch. d'Orléans	30	26 136	1 869	1,61 %	12	litt
15	*Le Jouvencel 1*	28	22 756	1 423	1,97 %	15	did
15	*Cité Dames*	26	22 838	1 253	2,08 %	14	litt
14	*Geneviève 2*	25	13 838	962	2,60 %	11	rel
15	*Quadrilogue*	21	21 041	826	2,54 %	21	did
14	*Chronique de Morée*	20	21 759	1 124	1,78 %	15	hist
15	*Chronique* Monstrelet	17	32 757	1 411	1,20 %	18	hist
15	*Evangiles Quenouilles* 1	15	9 727	539	2,78 %	16	did
15	*Bathilde* 2	14	12v517	638	2,19 %	16	rel
14	*De la erudition* J. Daudin	14	22v794	1 328	1,05 %	14	did
14	*Récit...* O. d'Anglure	13	27 923	1 487	0,87 %	16	litt
14	*Gdes chron. de France* 9	9	15 377	736	1,22 %	17	hist
14	*L'Art de dictier*	8	6 421	291	2,75 %	15	did
14	*Chroniques Jean II et Ch. V*	2	14 281	528	0,38 %	19	hist
15	*Journal 1* N. de Baye	1	24 111	767	0,13 %	22	jur

Tableau 7.3 – Comparatives en moyen français

Bibliographie

Abeille, Anne, Kathleen Bishop, Sharon Cote, et Yves Schabes (1990). « A Lexicalized Tree Adjoining Grammar for English ». *Technical Reports (CIS)*.

Abeillé, Anne (1993). *Les Nouvelles Syntaxes*. Armand Colin, Paris.

– (1998). « Verbes «à montée» et auxiliaires dans une grammaire d'arbres adjoints ». *Linx. Revue des linguistes de l'université Paris X Nanterre* 39, p. 119-158.

– (2002). *Une grammaire électronique du français*. CNRS Editions.

Abeillé, Anne, Lionel Clément, et François Toussenel (2003). « Building a treebank for French ». In : *Treebanks : Building and Using Parsed Corpora*. Anne Abeillé, Kluwer Academic Publishers.

Abeillé et Danièle Godard (1999). « French Word Order And Lexical Weight ». In : *The Nature and Function of Syntactic Categories*, p. 325-360.

Agić, *Željko*, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, et Anders Søgaard (2016). « Multilingual Projection for Parsing Truly Low-Resource Languages ». *Transactions of the Association for Computational Linguistics*.

Alonso, Miguel A., David Cabrero, Eric Villemonte de la Clergerie, et Manuel Vilares (1999). « Tabular Algorithms for TAG Parsing ». In : *Ninth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Bergen, Norway, p. 150-157.

Amiot, Dany, Claire Badiou-Monferran, Bernard Combettes, Benjamin Fagard, Christiane Marchello-Nizia, et Maj-Britt Mosegaard Hansen (2020). « Catégories invariables ». In : *Grande Grammaire historique du français*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. de Gruyter, p. 856-961.

Aubin, Sophie, Adeline Nazarenko, et Claire Nédellec (2005). « Adapting a general parser to a sublanguage ». In : *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005)*. Sous la dir. de G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, et N. Nikolov. INCOMA Ltd., Borovets, Bulgaria, p. 89-93.

Baldwin, Timothy, John Beavers, Emily M. Bender, Dan Flickinger, Ara Kim, et Stephan Oepen (2008). « Beauty and the Beast : What Running a Broad-Coverage Precision Grammar over the BNC Taught Us about the Grammar — and the Corpus ». In : *Linguistic Evidence : Empirical, Theoretical and Computational Perspectives*. Sous la dir. de Stephan Kepser et Marga Reis. De Gruyter Mouton, p. 49-70.

Bamman, David et Gregory Crane (2011). « The Ancient Greek and Latin Dependency Treebanks ». In : *Language Technology for Cultural Heritage*. Sous la dir. de Caroline Sporleder, Antal van den Bosch, et Kalliopi Zervanou. Springer, Berlin, Heidelberg, p. 79-98. isbn : 978-3-642-20227-8.

Bamman, David, Marco Passarotti, Roberto Busa, et Gregory Crane (2008). « The Annotation Guidelines of the Latin Dependency Treebank and Index Thomisticus Treebank : the Treatment of some specific Syntactic Constructions in Latin ». In : *Proceedings of the Sixth International Conference on*

Language Resources and Evaluation (LREC'08). European Language Resources Association (ELRA), Marrakech, Morocco.

Baron, Alistair et Paul Rayson (2008). « VARD2 : A tool for dealing with spelling variation in historical corpora ». In : *Postgraduate conference in corpus linguistics*.

Barteld, Fabian (2017). « Detecting spelling variants in non-standard texts ». In : *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics (EACL)*, p. 11-22.

Barteld, Fabian, Chris Biemann, et Heike Zinsmeister (2018). « Variations on the theme of variation : Dealing with spelling variation for fine-grained POS tagging of historical texts ». In : *Proceedings of the 14th Conference on Natural Language Processing KONVENS 2018*. Sous la dir. d'Adrien Barbaresi, Hanno Biber, Friedrich Neubarth, et Rainer Osswald. Vienna, Austria.

Barteld, Fabian, Ingrid Schröder, et Heike Zinsmeister (2015). « Unsupervised regularization of historical texts for POS tagging ». In : *Proceedings of the workshop on corpus-based research in the humanities (CRH)*. Polish Academy of Sciences : Institute of Computer Science, p. 3-12.

Ben Fraj, Fériel (2011). « Construction d'une grammaire d'arbres adjoints pour la langue arabe (Construction of a tree adjoining grammar for the Arabic language) ». In : *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*. ATALA, Montpellier, France, p. 109-115.

Ben-David, Shai, John Blitzer, Koby Crammer, et Fernando Pereira (2006). « Analysis of Representations for Domain Adaptation ». *Advances in neural information processing systems (NIPS)* 19, p. 137-144.

Bender, Emily M., Dan Flickinger, et Stephan Oepen (2002). « The Grammar Matrix : An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars ». In : *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Sous la dir. de John Carroll, Nelleke Oostdijk, et Richard Sutcliffe. Taipei, Taiwan, p. 8-14.

Berdicevskis, Aleksandrs et Hanne Eckhoff (2020). « A Diachronic Treebank of Russian Spanning More Than a Thousand Years ». In : *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France, p. 5251-5256.

Bladier, Tatiana, Kilian Evang, Valeria Generalova, Zahra Ghane, Laura Kallmeyer, Robin Möllemann, Natalia Moors, Rainer Osswald, et Simon Petitjean (2022). « RRGparbank : A Parallel Role and Reference Grammar Treebank ». In : *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*. Marseille, France.

Blanche-Benveniste, Claire et André Chervel (1969). *L'orthographe*. FeniXX.

Blitzer, John, Ryan McDonald, et Fernando Pereira (2006). « Domain Adaptation with Structural Correspondence Learning ». In : *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, p. 120-128.

Bohnet, Bernd (2010). « Very High Accuracy and Fast Dependency Parsing is not a Contradiction ». *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. Sous la dir. d'Association for Computational Linguistics, p. 89-97.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter, et Jan Hajič (2013). « Joint Morphological and Syntactic Analysis for Richly Inflected Languages ». *Transactions of the Association for Computational Linguistics* 1, p. 415-428.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, et Tomas Mikolov (2017). « Enriching Word Vectors with Subword Information ». *Transactions of the Association for Computational Linguistics* 5, p. 135-146.

Bollmann, Marcel (2018). « Normalization of historical texts with neural network models ». Thèse de doctorat. Ruhr-Universität Bochum.

Bollmann, Marcel et Anders Søgaard (2016). « Improving historical spelling normalization with bi-directional LSTMs and multi-task learning ». In : *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, p. 131-139.

Bonfante, Guillaume, Bruno Guillaume, et Guy Perrier (2018). *Application of Graph Rewriting to Natural Language Processing*. Wiley-ISTE.

Boullier, Pierre, Lionel Clément, Benoît Sagot, et Eric Villemonte de la Clergerie (2005). « Chaînes de traitement syntaxique ». In : *Actes de TALN 2005*. Dourdan, France, p. 103-112.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, et George Smith (2002). « The TIGER treebank ». *Proceedings of the workshop on treebanks and linguistic theories*.

Briscoe, Ted et John Carroll (1997). « Automatic Extraction of Subcategorization from Corpora ». *arXiv preprint*.

Buridant, Claude (2000). *Nouvelle Grammaire de l'ancien français*. Sedes, Paris.

Butt, Myriam, Tracy Holloway King, Maria-Eugenia Nino, et Frédérique Segond (1999). *A Grammar Writer's Cookbook*. CSLI Publications, United States.

Camps, Jean-Baptiste, Elena Albarran, Alice Cochet, et Lucence Ing (2016). « Geste : un corpus de chansons de geste ».

Camps, Jean-Baptiste, Thibault Clérice, Ariane Pinche, Lucence Ing, Frédéric Duval, et Naomi Kanaoka (2020). *Deucalion, Modèle Ancien Francais*. Version 0.3.0.

Candito, Marie (1996). « A principle-based hierarchical representation of LTAGs ». *Proceedings of the 16th Conference on Computational Linguistics (Coling 1996)*.

– (1999). « Organisation modulaire et paramétrable de grammaires électroniques léxicalisees application au francais et à l'italien ». Thèse de doctorat. Université Paris 7.

Candito, Marie, Benoît Crabbé, et Pascal Denis (2010). « Statistical French dependency parsing : treebank conversion and first results ». In : *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), p. 1840-1847.

Carlier, Anne et Bernard Combettes (2020). « Morphologie dérivationnelle vs. flexionnelle ». In : *Grande Grammaire Historique du Français (GGHF)*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. De Gruyter Mouton, p. 622-631.

Carlier, Anne, Bernard Combettes, Céline Guillot-Barbance, Christiane Marchello-Nizia, Evelyne Oppermann-Marsaux, Sophie Prévost, et Catherine Schnedecker (2020). « Syntaxe interne des groupes de mots et morphèmes ». In : *Grande Grammaire historique du français*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. de Gruyter, p. 971-1054.

Carlier, Anne, Céline Guillot-Barbance, Christiane Marchello-Nizia, et Lene Schøsler (2020). « Catégories variables : noms, adjectifs, pronoms et déterminants ». In : *Grande Grammaire Histo-*

rique du Français (GGHF). Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. De Gruyter Mouton, p. 632-744.

Carroll, John et Alex C. Fang (2004). « The Automatic Acquisition of Verb Subcategorisations and their Impact on the Performance of an HPSG Parser ». In : *International Conference on Natural Language Processing (IJCNLP 2004)*. Sous la dir. de Keh-Yih Su, Jun'ichi Tsujii, Jong-Hyeok Lee, et Oi Yee Kwong. Springer. Springer, Berlin, Heidelberg, p. 646-654.

Carroll, John, Nicolas Nicolov, Olga Shaumyan, Martine Smets, et David Weir (1999). « Parsing with an Extended Domain of Locality ». In : *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*. EACL '99. Association for Computational Linguistics, Bergen, Norway, p. 217-224.

Cazal, Yvonne, Gabriella Parussa, et Elena Llamas-Pombo (2020). « Graphies : des usages à la norme ». In : *Grande Grammaire historique du français*. de Gruyter, p. 501-549.

Cecchini, Flavio Massimiliano, Timo Korkiakangas, et Marco Passarotti (2020). « A new latin treebank for universal dependencies : Charters between ancient latin and romance languages ». In : *Proceedings of The 12th Language Resources and Evaluation Conference (LREC'12)*, p. 933-942.

Cecchini, Flavio Massimiliano, Rachele Sprugnoli, Giovanni Moretti, et Marco Passarotti (2020). « UDante : First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works ». In : *Seventh Italian Conference on Computational Linguistics*. CEUR-WS. org, p. 1-7.

Chagué, Alix, Victoria Le Fourner, Manuela Martini, et Eric Villemonte de la Clergerie (2019). « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? » In : *Colloque DHNord 2019 "Corpus et archives numériques"*.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, et Yoshua Bengio (2014). « Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation ». In : *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, p. 1724-1734.

Chomsky, Noam (1959). « On certain formal properties of grammars ». *Information and control* 2 :2, p. 137-167.

– (1993). *Lectures on government and binding : The Pisa lectures*. 9. Walter de Gruyter.

Church, Kenneth Ward (2011). « A Pendulum Swung too Far ». *Linguistic Issues in Language Technology* 6.

Clément, Lionel et Alexandra Kinyon (2003). « Generating Parallel Multilingual LFG-TAG Grammars from a MetaGrammar ». In : *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Sapporo, Japan, p. 184-191.

Clément, Lionel, Bernard Lang, et Benoît Sagot (2004). « Morphology based automatic acquisition of large-coverage lexica ». In : *LREC 04*. Lisbonne, Portugal, p. 1841-1844.

Cohen, Raphael, Yoav Goldberg, et Michael Elhadad (2012). « Domain adaptation of a dependency parser with a class-class selectional preference model ». In : *Proceedings of ACL 2012 Student Research Workshop*, p. 43-48.

Combettes, Bernard (2020). « Niveau informationnel ». In : *Grande Grammaire historique du français*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. de Gruyter, p. 1739-1777.

Combettes, Bernard et Julie Glikman (2020). « Syntaxe de la phrase complexe ». In : *Grande Grammaire Historique du Français (GGHF)*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. De Gruyter Mouton, p. 1338-1465.

Combettes, Bernard, Christiane Marchello-Nizia, et Sophie Prévost (2020). « Syntaxe de la phrase simple ». In : *Grande Grammaire historique du français*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. de Gruyter, p. 1220-1337.

Combettes, Bernard et Sophie Prévost (2011). « La disparition du schéma V2 en français : le rôle de l'opposition 'marqué' / 'non marqué' dans le domaine syntaxique ». In : sous la dir. de Thomas Verjans and Claire Badiou-Monferran. Disparitions. Contributions à l'étude du changement linguistique. Champion, Dijon, France, p. 283-301.

Constant, Mathieu, Isabelle Tellier, Denys Duchier, Yoann Dupont, Anthony Sigogne, et Sylvie Billot (2011). « Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français ». In : *TALN*. T. 1. Montpellier, France, p. 321.

Crabbé, Benoît (2005). « Représentation informatique de grammaires fortement lexicalisées, Application à la grammaire d'arbres adjoints ». Thèse de doctorat. Université Nancy 2.

Danlos, Laurence (2005). « ILIMP : Outil pour repérer les occurrences du pronom impersonnel il ». In : *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*. ATALA, Dourdan, France, p. 121-130.

Danlos, Laurence et Benoît Sagot (2008). « Méthodologie lexicographique de constitution d'un lexique syntaxique de référence pour le français ». In : *Proceedings of the workshop"Lexicographie et informatique : bilan et perspectives"*. Nancy, France.

Danlos, Laurence, Benoît Sagot, et Susanne Salmon-Alt (2006). « French frozen verbal expressions : from lexicon-grammar to NLP applications ». In : *Proceedings of the 25th Lexis and Grammar Conference, Palerme, Italie*.

de Kok, Daniël (2015). « A poor man's morphology for German transition-based dependency parsing ». In : *International Workshop on Treebanks and Linguistic Theories (TLT14)*, p. 50.

Demske, Ulrike, Nicola Frank, Stefanie Laufer, et Hendrik Stiemer (2004). « Syntactic Interpretation of an Early New High German Corpus ». In : *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*, p. 175-182.

Dereza, Oksana (2019). « Lemmatisation for under-resourced languages with sequence-to-sequence learning : A case of Early Irish ». In : *Proceedings of Third Workshop "Computational linguistics and language science"*. Sous la dir. de Gerhard Wohlgenannt, Ruprecht von Waldenfels, Svetlana Toldova, Ekaterina Rakhilina, Denis Paperno, Olga Lyashevskaya, Natalia Loukachevitch, Sergei O. Kuznetsov, Olga Kultepina, Dmitry Ilvovsky, Boris Galitsky, Ekaterina Artemova, et Elena Bolshakova. T. 4, p. 113-124.

Destemberg, Antoine (2017). *Atlas de la France médiévale*. Autrement, Paris.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova (2019). « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding ». In : *NAACL*.

Dozat, Timothy et Christopher D. Manning (2017). « Deep Biaffine Attention for Neural Dependency Parsing ». *ArXiv* abs/1611.01734.

Dredze, Mark, John Blitzer, Partha Pratim Talukdar, Kuzman Ganchev, João Graça, et Fernando Pereira (2007). « Frustratingly Hard Domain Adaptation for Dependency Parsing ». In : *Proceedings*

of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Association for Computational Linguistics, Prague, Czech Republic, p. 1051-1055.

Dukes, Kais et Tim Buckwalter (2010). « A dependency treebank of the Quran using traditional Arabic grammar ». In : *2010 the 7th International Conference on Informatics and Systems (INFOS)*. IEEE, p. 1-7.

Dupont, Yoann (2017). « La structuration dans les entités nommées ». Thèse de doctorat. Université Sorbonne Paris Cité.

Earley, Jay (1970). « An Efficient Context-Free Parsing Algorithm ». *Commun. ACM* 13 :2, p. 94-102.

Eckhoff, Hanne, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag T. T. Haug, Odd Einar Haugen, et Marius Jøhndal (2018). « The PROIEL treebank family : a standard for early attestations of Indo-European languages ». *Language Resources and Evaluation* 52 :1, p. 29-65.

Estarrona, Ainara, Izaskun Etxeberria, Ricardo Etxepare, Manuel Padilla-Moyano, et Ander Soraluze (2020). « Dealing with dialectal variation in the construction of the Basque historical corpus ». In : *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, p. 79-89.

Falk, Ingrid, Gil Francopoulo, et Claire Gardent (2007). « Evaluer synlex ». In : *Traitement Automatique de la Langue Naturelle - TALN 2007*. Toulouse, France.

Ferreira, Auphélie (2019). « Influence du contexte communicationnel sur l'emploi des constructions syntaxiques de type [V.ØP.] avec les verbes croire et penser ». In : *Rencontres des Jeunes Chercheurs en Sciences du Langage 2019*. Paris, France.

Fort, Karën (2012). « Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus ». Thèse de doctorat. Université Paris-Nord - Paris XIII.

– (2016). *Grammaires formelles*. Cours de Licence Langue et informatique à l'Université Paris-Sorbonne.

Fort, Karën et Bruno Guillaume (2008). « Sylva : plate-forme de validation multi-niveaux de lexiques ». In : *Traitement Automatique des Langues Naturelles*. Avignon, France.

Frank, Anette (2001). « Treebank Conversion. Converting the NEGRA treebank to an LTAG grammar ». In : *Proceedings of the Workshop on Multi-layer Corpus-based Analysis*. EUROLAN 2002 Summer Institute, Iasi et Romania, p. 29-43.

Gabay, Simon, Thibault Clérice, Jean-Baptiste Camps, Jean-Baptiste Tanguy, et Matthias Gille-Levenson (2020). « Standardizing linguistic data : method and tools for annotating (pre-orthographic) French ». In : *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress (DTUC '20)*. Hammamet, Tunisia, p. 1-7.

Gardent, Claire, Bruno Guillaume, Guy Perrier, et Ingrid Falk (2006). « Extraction d'information de sous-catégorisationa partir du lexique-grammaire de Maurice Gross ». *TALN 2006*.

Gerdes, Kim, Sylvain Kahane, Rachel Bawden, Julie Beliao, Eric Villemonte de la Clergerie, et Ilaine Wang (2019). « Annotation tools for syntax ». In : *Rhapsodie : a Prosodic-Syntactic Treebank for Spoken French*. T. 89. John Benjamins Publishing Company.

Gingras, Francis (2004). « La mauvaise langue et les lettres : statuts de la rumeur et de l'écrit à la naissance du roman (1150-1230) ». *Protée* 32 :3, p. 87-99.

Glessgen, Martin-Dietrich (2003). « L'élaboration philologique et l'étude lexicologique des Plus anciens documents linguistiques de la France à l'aide de l'informatique ». *Mémoires et documents de l'École des chartes* 71, p. 371-386.

Gómez-Rodríguez, Carlos, Miguel A. Alonso, et Manuel Vilares (2006). « On Theoretical and Practical Complexity of TAG parsers ». In : *Proceedings of FG 2006 : The 11th conference on Formal Grammar, volume of FG Online Proceedings, pages–. CSLI publications, Stanford, CA, USA.(Cited on page.)* Sous la dir. de Shuly Wintner. Malaga, Spain, p. 87-101.

Goux, Mathieu et Pierre Larrivée (2020). « Expression et position du sujet en ancien français : le rôle de la personne pronominale ». In : *SHS Web of Conferences, 7e Congrès Mondial de Linguistique Française (CMLF)*. T. 78. EDP Sciences, p. 03002.

Grobol, Loïc et Benoît Crabbé (2021). « Analyse en dépendances du français avec des plongements contextualisés ». In : *28e Conférence sur le Traitement Automatique des Langues Naturelles*. Lille (virtuel), France.

Grobol, Loïc, Sophie Prévost, et Benoît Crabbé (2022). « Is Old French tougher to parse ? » In : *20th International Workshop on Treebanks and Linguistic Theories*. Sofia, Bulgaria.

Grobol, Loïc, Mathilde Regnault, Pedro Ortiz Suarez, Benoit Sagot, Laurent Romary, et Benoît Crabbé (2022). « BERTrade : Using Contextual Embeddings to Parse Old French ». In : *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22)*. Marseille, France.

Gross, Maurice (1975). *Méthodes en syntaxe*. Hermann, Paris.

Guibon, Gaël et Benoît Sagot (2020). « OFrLex : A Computational Morphological and Syntactic Lexicon for Old French ». In : *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*. Marseille, France, 3217-3225 (updated version).

Guibon, Gaël, Isabelle Tellier, Mathieu Constant, Sophie Prévost, et Kim Gerdes (2014). « Parsing Poorly Standardized Language Dependency on Old French ». In : *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*. V. Henrich and E. Hinrichs and D. de Kok and P. Osenova and A. Przepiórkowski, Tübingen, Germany, p. 51-61.

Guibon, Gaël, Isabelle Tellier, Sophie Prévost, Mathieu Constant, et Kim Gerdes (2015). « Searching for Discriminative Metadata of Heterogenous Corpora ». In : *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*. Dickinson, Markus and Hinrichs, Erhard and Patejuk, Agnieszka and Przepiórkowski, Adam, Varsovie, Poland, p. 72-82.

Guillot, Céline, Serge Heiden, et Alexei Lavrentiev (2018). « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique ». *Diachroniques. Revue de Linguistique française diachronique*. Les états anciens des langues à l'heure du numérique 7, p. 168-184.

Guillot, Céline, Sophie Prévost, et Alexei Lavrentiev (2013). « Manuel de référence du jeu Cattex09 ».

Hajic, Jan (1998). « Building a syntactically annotated corpus : The prague dependency treebank ». *Issues of valency and meaning*, p. 106-132.

Han, Chung-hye, Juntae Yoon, Nari Kim, et Martha Palmer (2000). « A Feature-based Lexicalized Tree Adjoining Grammar for Korean ». *IRCS Technical Reports Series*.

Hara, Tadayoshi, Yusuke Miyao, et Jun'ichi Tsujii (2005). « Adapting a Probabilistic Disambiguation Model of an HPSG Parser to a New Domain ». In : *Second International Joint Conference on Natural Language Processing : Full Papers*.

Hasenohr, Geneviève et Guy Raynaud de Lage (1993). *Introduction à l'ancien français de Guy Raynaud de Lage*. Sedes.

Haug, Dag T. T. (2012). « From dependency structures to LFG representations ». In : *Proceedings of the LFG12 Conference*, p. 271-291.

Haug, Dag T. T. et Marius Jøhndal (2008). « Creating a parallel treebank of the old Indo-European Bible translations ». In : *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, p. 27-34.

Heiden, Serge et Alexei Lavrentiev (2004). « Ressources électroniques pour l'étude des textes médiévaux : approches et outils ». *Revue française de linguistique appliquée* 9 :1.

Heiden, Serge, Jean-Philippe Magué, et Bénédicte Pincemin (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement ». In : *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Sous la dir. de Sergio Bolasco, Isabella Chiari, et Luca Giuliano. T. 2. 3. Edizioni Universitarie di Lettere Economia Diritto, Rome, Italy, p. 1021-1032.

Heiden, Serge et Sophie Prévost (2002). « Etiquetage d'un corpus hétérogène de français médiéval : enjeux et modalités ». In : *Romance Corpus Linguistics - Corpora and Spoken Language, Tübingen, Gunter Narr Verlag Tübingen*. Sous la dir. de C.D. Pusch et W. Raible, p. 127-136.

Holgado, Cristina, Alexei Lavrentiev, et Mathieu Constant (2021). « Évaluation de méthodes et d'outils pour la lemmatisation automatique du français médiéval ». In : *Traitement Automatique des Langues Naturelles*. Sous la dir. de Pascal Denis, Natalia Grabar, Amel Fraisse, Rémi Cardon, Bernard Jacquemin, Eric Kergosien, et Antonio Balvet. ATALA. Lille, France, p. 153-161.

Joshi, Aravind K. (1985). « Tree adjoining grammars : How much context-sensitivity is required to provide reasonable structural descriptions ? » In : *Natural Language Parsing : Psychological, Computational, and Theoretical Perspectives*. Sous la dir. de David R. Dowty, Lauri Karttunen, et Arnold Zwicky. Studies in Natural Language Processing. Cambridge University Press, Cambridge, Royaume-Uni, p. 206-250.

– (2004). « Tree-Adjoining Grammars ». In : *The Oxford handbook of computational linguistics*. Sous la dir. de Ruslan Mitkov. Oxford University Press.

Joshi, Aravind K., Leon S. Levy, et Masako Takahashi (1975). « Tree adjunct grammars ». *Journal of Computer and System Sciences*.

Joshi, Aravind K. et Yves Schabes (1991). « Tree-adjoining grammars and lexicalized grammars ». *Technical Reports (CIS)*, p. 445.

– (1997). « Tree-Adjoining Grammars ». In : *Handbook of Formal Languages : Volume 3 Beyond Words*. Sous la dir. de Grzegorz Rozenberg et Arto Salomaa. Springer Berlin Heidelberg, p. 69-123.

Joshi, Aravind K., Kenneth Vijay-Shanker, et David Weir (1990). « The Convergence of Mildly Context-Sensitive Grammar Formalisms ». *University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-01*.

Kallmeyer, Laura (2010). *Parsing beyond context-free grammars*. Springer.

Kallmeyer, Laura, Benjamin Burkhardt, Timm Lichte, et Simon Petitjean (2017). *Grammatikimplementierung mit Tree Adjoining Grammar (TAG)*. Cours à l'Université Heinrich Heine de Düsseldorf.

Kaplan, Ronald M., Tracy Holloway King, et John T. Maxwell III (2002). « Adapting existing grammars : the XLE experience ». *COLING-02 on Grammar engineering and evaluation*.

Kaplan, Ronald M. et John T. Maxwell III (1994). « Grammar writer's workbench ». *Xerox Corporation, Version* 2.

Kestemont, Mike, Guy de Pauw, Renske van Nie, et Walter Daelemans (2016). « Lemmatization for variation-rich languages using deep learning ». *Digital Scholarship in the Humanities* 32 :4, p. 797-815.

Kinyon, Alexandra (2000). « HYPERTAGS : Beyond POS Tagging ». In : *Natural Language Processing — NLP 2000*. Sous la dir. de Dimitris N. Christodoulakis. Springer, Berlin, Heidelberg, p. 81-90.

Kogkitsidou, Eleni et Philippe Gambette (2020). « Normalisation of 16th and 17th century texts in French and geographical named entity recognition ». In : *Proceedings of the 4th ACM SIGSPATIAL Workshop on Geospatial Humanities*, p. 28-34.

Korhonen, Anna (2002). *Subcategorization acquisition*. Rapp. tech. University of Cambridge, Computer Laboratory.

Kroch, Anthony S. et Aravind K. Joshi (1985). « The Linguistic Relevance of Tree Adjoining Grammar ». *University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-85-16*.

Kroch, Anthony S., Ann Taylor, et Beatrice Santorini (2000). « The Penn-Helsinki Parsed Corpus of Middle English (PPCME2) ». *CD-ROM, second edition, release 4*. Sous la dir. d'University of Pennsylvania Department of Linguistics.

Kupść, Anna et Anne Abeillé (2008). « Growing treelex ». In : *International Conference on Intelligent Text Processing and Computational Linguistics*. Sous la dir. d'Alexander Gelbukh. Springer, p. 28-39.

Lafferty, John D., Andrew McCallum, et Fernando Pereira (2001). « Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data ». In : *ICML*, p. 282-289.

Lavergne, Thomas, Olivier Cappé, et François Yvon (2010). « Practical Very Large Scale CRFs ». In : *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL '10. Association for Computational Linguistics, Uppsala, Sweden, p. 504-513.

Lecointe, Jean (1997). « Le style en-ant au XVIe siècle en France : conscience syntaxique et options stylistiques ». *L'Information grammaticale* 75 :1, p. 10-14.

Lee, John et Yin Hei Kong (2012). « A dependency treebank of classical Chinese poems ». In : *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 191-199.

– (2014). « A dependency treebank of Chinese Buddhist texts ». *Digital Scholarship in the Humanities* 31 :1, p. 140-151. eprint : https://academic.oup.com/dsh/article-pdf/31/1/140/21517990/fqu048.pdf.

Lezius, Wolfgang (2002). « Ein Suchwerkzeug für syntaktisch annotierte Textkorpora ».

Lim, KyungTae, Cheoneum Park, Changki Lee, et Thierry Poibeau (2018). « SEx BiST : A Multi-Source Trainable Parser with Deep Contextualized Lexical Representations ». In : *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics. Bruxelles, Belgique, p. 143-152.

Mairey, Aude et Mourad Aouini (2021). « PALM : Un modèle neuronal pour l'étiquetage morphosyntaxique des textes médiévaux ». In : *JADT 2020 : 15èmes Journées Internationales d'Analyse statistique des Données Textuelles*.

Manjavacas, Enrique, Ákos Kádár, et Mike Kestemont (2019). « Improving Lemmatization of Non-Standard Languages with Joint Learning ». In : *NAACL-HLT (1)*, p. 1493-1503.

Marchello-Nizia, Christiane (1979). *Histoire de la langue français aux XIV^e^ et XV^e^ siècles*. Bordas.
– (1995). *L'évolution du français : ordre des mots, démonstratifs, accent tonique*. Armand Colin, Paris.
– (1999). *Le français en diachronie : douze siècles d'évolution*. Editions Ophrys.
– (2008). « L'évolution de l'ordre des mots en français : Chronologie, périodisation, et réorganisation du système ». *Congrès Mondial de Linguistique Française*.

Marchello-Nizia, Christiane et Sophie Prévost (2020). « Expression et position des constituants majeurs dans les divers types de propositions ». In : *Grande Grammaire historique du français*. Sous la dir. de Christiane Marchello-Nizia, Bernard Combettes, Sophie Prévost, et Tobias Scheer. de Gruyter, p. 1055-1219.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, et Britta Schasberger (1994). « The Penn Treebank : Annotating Predicate Argument Structure ». In : *Human Language Technology : Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Martin, Louis, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djamé Seddah, et Benoît Sagot (2020). « CamemBERT : a Tasty French Language Model ». *arXiv preprint arXiv:1911.03894*.

Martin, Robert, Sylvie Bazin, et Pierre Cromer (2012). *Dictionnaire du moyen français*.

Martineau, France (1990). « La construction «accusatif avec infinitif» avec les verbes causatifs et de perception en moyen français ». *Revue québécoise de linguistique* 19 :1, p. 77-100.

McClosky, David, Eugene Charniak, et Mark Johnson (2010). « Automatic Domain Adaptation for Parsing ». In : *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Los Angeles, California, p. 28-36.

McDonald, Ryan, Slav Petrov, et Keith Hall (2011). « Multi-Source Transfer of Delexicalized Dependency Parsers ». In : *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Edinburgh, Scotland, UK., p. 62-72.

Millour, Alice et Karën Fort (2019). « Unsupervised Data Augmentation for Less-Resourced Languages with no Standardized Spelling ». In : *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, p. 776-784.

Molinero, Miguel, Benoît Sagot, et Lionel Nicolas (2009). « A morphological and syntactic wide-coverage lexicon for Spanish : The Leffe ». In : *RANLP 2009 - Recent Advances in Natural Language Processing*.

Müller, Stefan, Anne Abeillé, Robert D. Borsley, et Jean-Pierre Koenig, éd. (2021). *Head Driven Phrase Structure Grammar*. Empirically Oriented Theoretical Morphology and Syntax 9. Language Science Press, Berlin.

Nicolas, Lionel, Jacques Farré, et Eric Villemonte de la Clergerie (2007). « Confondre le coupable : Corrections d'un lexique suggérées par une grammaire ». In : *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*. Sous la dir. de Farah Benamara, Nabil Hathout, Philippe Muller, et Sylwia Ozdowska. IRIT, Toulouse, France, p. 295-304.

Nivre, Joakim, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira et al. (2016).

« Universal dependencies v1 : A multilingual treebank collection ». In : *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 1659-1666.

Nivre, Joakim, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, et Deniz Yuret (2007). « The CoNLL 2007 shared task on dependency parsing ». In : *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 915-932.

Parmentier, Yannick (2007). « SemTAG : une plate-forme pour le calcul sémantique à partir de Grammaires d'Arbres Adjoints ». Thèse de doctorat. Université Henri Poincaré - Nancy I.

Parussa, Gabriella (2010). « Editer les textes de théâtre en langue française : aperçu historique et nouvelles perspectives ». *Médiévales. Langues, Textes, Histoire* 59 :59, p. 41-61.

Pedrazzini, Nilo (2020). « Exploiting cross-dialectal gold syntax for low-resource historical languages : Towards a generic parser for pre-modern Slavic ». *arXiv preprint arXiv:2011.06467*.

Pedrazzini, Nilo et Hanne Martine Eckhoff (2021). « OldSlavNet : A scalable Early Slavic dependency parser trained on modern language data ». *Software Impacts* 8. issn : 2665-9638.

Plank, Barbara et Gertjan van Noord (2011). « Effective Measures of Domain Similarity for Parsing ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, p. 1566-1576.

Polguère, Alain (2003). « Étiquetage sémantique des lexies dans la base de données DiCo ». *Traitement automatique des langues* 44 :2, p. 39-68.

Prévost, Sophie (2005). « Constitution et exploitation d'un corpus de français médiéval : enjeux, spécificités et apports ». In : *Sémantique et corpus*. Sous la dir. d'A. Condamines (éd). Série " Traité IC2 " : Cognition et traitement de l'information. Hermès/Lavoisier, p. 147-176.

– (2008). « Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports ». *Corpus* 7, p. 35-64.

– (2011). « Français médiéval en diachronie : du corpus à la langue (mémoire de synthèse) ». Habilitation à diriger des recherches. Ecole normale supérieure de lyon (ENS Lyon).

– (2015). « Diachronie du français et linguistique de corpus : une approche quantitative renouvelée ». *Langages*. La fréquence textuelle : bilan et perspectives 197 :1, p. 23-45.

– (2020). « Une grammaire fondée sur un corpus numérique ». In : *Grande Grammaire historique du français*. de Gruyter, p. 37-53.

Rainsford, Thomas, Céline Guillot, Alexei Lavrentiev, et Sophie Prévost (2012). « La zone préverbale en ancien français : apport des corpus annotés ». In : *3e Congrès Mondial de Linguistique Française*. Sous la dir. de Franck Neveu ; Valelia Muni Toke ; Peter Blumenthal ; Thomas Klingler ; Pierluigi Ligas ; Sophie Prévost ; Sandra Teston-Bonnard. T. 1. Lyon, France, p. 159-176.

Riegel, Martin, Jean-Christophe Pellat, et René Rioul (1994). *Grammaire méthodique du français*. Presses universitaires de France, Paris.

Rimell, Laura et Stephen Clark (2008). « Adapting a Lexicalized-Grammar Parser to Contrasting Domains ». In : *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, p. 475-484.

Rocio, Vitor, Mário Amado Alves, Gabriel Pereira Lopes, Maria Francisca Xavier, et Graça Vicente (2003). « Automated creation of a medieval portuguese partial treebank ». In : *Treebanks : Building and Using Parsed Corpora*. Anne Abeillé, Kluwer Academic Publishers.

Rögnvaldsson, Eiríkur, Anton Karl Ingason, Einar Freyr Sigurðsson, et Joel Wallenberg (2012). « The Icelandic Parsed Historical Corpus (IcePaHC) ». In : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Sous la dir. de Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, et Stelios Piperidis. European Language Resources Association (ELRA), Istanbul, Turkey, p. 1977-1984.

Role, François, Milagros Fernandez Gavilanes, et Eric Villemonte de la Clergerie (2007). « Large-scale Knowledge acquisition from botanical texts ». In : *International Conference on Application of Natural Language to Information Systems*. Springer, p. 395-400.

Rothwell, William et David Trotter (2005). « Anglo-Norman Dictionary 2 ».

Rouquier, Magali (2014). *L'Émergence des constructions clivées, pseudo-clivées et liées en français*. Classiques Garnier.

Sagot, Benoît (2010). « The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French ». In : *7th international conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.

– (2014). « DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German ». In : *Language Resources and Evaluation Conference*.

– (2018). « Informatiser le lexique ». Habilitation à diriger des recherches. Sorbonne Université.

– (2019). « Développement d'un lexique morphologique et syntaxique de l'ancien français ». In : *26ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*. Toulouse, France.

Sagot, Benoît et Pierre Boullier (2008). « SxPipe 2 : architecture pour le traitement pré-syntaxique de corpus bruts ». *Traitement Automatique des Langues* 49 :2, p. 155-188.

Sagot, Benoît, Lionel Clément, Eric Villemonte de la Clergerie, et Pierre Boullier (2006). « The Lefff 2 syntactic lexicon for French : architecture, acquisition, use ». In : *LREC 06*. Gênes, Italy, p. 1-4.

Sagot, Benoît et Laurence Danlos (2008). « Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire — Constructions impersonnelles et expressions verbales figées ». *Cahiers du CENTAL*. Description linguistique pour le traitement automatique du français 5, p. 107-126.

Sagot, Benoît et Karën Fort (2007). « Améliorer un lexique syntaxique à l'aide des tables du Lexique-Grammaire : adverbes en-ment ». In : *26e Colloque International sur le Lexique et la grammaire 2007*.

Sagot, Benoît et Eric Villemonte de la Clergerie (2006). « Error mining in parsing results ». *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*.

– (2008). « Fouille d'erreurs sur des sorties d'analyseurs syntaxiques ». *Revue TAL (Traitement Automatique des Langues)* 49 :1, p. 41-60.

Sagot, Benoît et Géraldine Walther (2013). « Implementing a Formal Model of Inflectional Morphology ». In : *Systems and Frameworks for Computational Morphology*. Cerstin Mahlow et Michael Piotrowski, Berlin, Heidelberg : Springer Berlin Heidelberg, p. 115-134.

Schabes, Yves (1991). « The Valid Prefix Property and Left to Right Parsing of Tree-Adjoining Grammar ». In : *Proceedings of the Second International Workshop on Parsing Technologies*. Association for Computational Linguistics, Cancun, Mexico, p. 21-30.

Schabes, Yves, Anne Abeillé, et Aravind K. Joshi (1988). « Parsing strategies with "lexicalized" grammars : Application to tree adjoining grammars ». *Technical Reports (CIS)*, p. 691.

Schabes, Yves et Aravind K. Joshi (1991). « Parsing with Lexicalized Tree Adjoining Grammar ». In : *Current Issues in Parsing Technology*. Sous la dir. de Masaru Tomita. The Springer International Series in Engineering and Computer Science. Springer US, p. 25-47.

Schabes, Yves et Richard C. Waters (1995). « Tree Insertion Grammar : A Cubic-Time, Parsable Formalism that Lexicalizes Context-Free Grammar without Changing the Trees Produced ». *Computational Linguistics* 21 :4, p. 479-513.

Schauwecker, Yela et Achim Stein (2018). « Automatic Morphosyntactic and Dependency Annotation of the Anglo-Norman Text Database ». In : *Grammar and Corpora 2016*. Konopka, Marek and Trawinski, Beata and Waßner, Ullrich and Fuß, Eric, Heidelberg, Germany, p. 357-376.

Schmid, Helmut (1994). « TreeTagger ». *arXiv preprint arXiv:1911.03894*.

– (2019). « Deep learning-based morphological taggers and lemmatizers for annotating historical texts ». In : *Proceedings of the 3rd international conference on digital access to textual cultural heritage*, p. 133-137.

Schøsler, Lene (1984). *La déclinaison bicasuelle de l'ancien français : son rôle dans la syntaxe de la phrase, les causes de sa disparition*. Odense University Press.

Scrivner, Olga et Sandra Kübler (2012). « Building an old Occitan corpus via cross-Language transfer ». In : *Proceedings of KONVENS 2012*. Sous la dir. de Jeremy Jancsary. LThist 2012 workshop. ÖGAI, p. 392-400.

Shieber, Stuart M. (1985). « Evidence against the context-freeness of natural language ». In : *Philosophy, language, and artificial intelligence*. Springer, p. 79-89.

Simonenko, Alexandra, Benoît Crabbé, et Sophie Prévost (2020). « Text form and grammatical changes in Medieval French : A treebank-based diachronic study ». In : *Diachronic Treebanks for Historical Linguistics*. Sous la dir. d'Hanne Martine Eckoff, Silvia Luraghi, et Marco Passarotti. John Benjamins Publishing Company, p. 95-128.

Siouffi, Gilles (2020a). « Les données historiques, géographiques et démographiques ». In : *Grande Grammaire historique du français*. de Gruyter, p. 91-109.

– (2020b). « Les genres textuels ». In : *Grande Grammaire historique du français*. de Gruyter, p. 121-134.

Sleator, Daniel D. et Davy Temperley (1993). « Parsing English with a Link Grammar ». In : *Proceedings of the Third International Workshop on Parsing Technologies*. Association for Computational Linguistics, Tilburg, Netherlands et Durbuy, Belgium, p. 277-292.

Souvay, Gilles et Jean-Marie Pierrel (2009). « LGeRM Lemmatisation des mots en moyen français ». *Traitement Automatique des Langues* 50 :2, p. 21.

Stein, Achim (2014). « Parsing Heterogeneous Corpora with a Rich Dependency Grammar ». In : *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, p. 2879-2886.

Stein, Achim (2016). « Old French Dependency Parsing : Results of Two Parsers Analysed from a Linguistic Point of View ». In : *Proceedings of the Tenth International Conference on Language Resources*

and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovenia, p. 707-713.

– (2018). « Diachronic syntax based on constituency and dependency annotated corpora : Theoretical and methodological issues ». *Linguistic Variation* 18 :1, p. 74-99.

Stein, Achim et Pierre Kunstmann (2003). « Etiquetage morphologique et lemmatisation de textes d'ancien français ». In : *Ancien et moyen français sur le Web : enjeux méthodologiques et analyse du discours*. Les Editions David, Ottawa, Canada.

Stein, Achim, Pierre Kunstmann, et Martin-Dietrich Glessgen (2011). *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-Dietrich Gleßgen, version 3*. Stuttgart, Allemagne.

Stein, Achim et Sophie Prévost (2013). « Syntactic annotation of medieval texts : the Syntactic Reference Corpus of Medieval French (SRCMF) ». en. In : *New Methods in Historical Corpora* (Manchester, Royaume-Uni, 29 avr. 2011). Sous la dir. de Paul Bennett, Martin Durrell, Silke Scheible, et Richard J. Whitt. Corpus Linguistics and International Perspectives on Language. Gunter Narr Verlag, p. 275-282. isbn : 978-3-8233-6760-4.

Straka, Milan (2018). « UDPipe 2.0 prototype at CoNLL 2018 UD shared task ». In : *Proceedings of the CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 197-207.

Straka, Milan, Jana Straková, et Jan Hajič (2019). *Evaluating Contextualized Embeddings on 54 Languages in POS Tagging, Lemmatization and Dependency Parsing*. arXiv : 1908.07448 [cs.CL].

Taylor, Ann, Mitchell Marcus, et Beatrice Santorini (2003). « The Penn Treebank : an Overview ». In : sous la dir. d'Anne Abeillé. Springer, Dordrecht, Pays-Bas, p. 5-22.

The York-Helsinki parsed corpus of Old English poetry (YCOEP) (2001). Oxford Text Archive.

Thomasset, François et Eric Villemonte de la Clergerie (2005). « Comment obtenir plus des métagrammaires ». In : *Actes de la 12ème conférence sur le Traitement Automatique des Langues Naturelles. Articles longs*. ATALA, Dourdan, France, p. 1-10.

Tilander, Gunnar (1963). « De sa femme ne voit mie, construction syntaxique d'origine cynégétique ». *Romania* 84 :335, p. 289-306.

Tolone, Elsa, Benoît Sagot, et Eric Villemonte de la Clergerie (2012). « Evaluating and improving syntactic lexica by plugging them within a parser ». In : *LREC 2012 - 8th International Conference on Language Resources and Evaluation*. Istambul, Turkey, electronic-version.

Traugott, Elizabeth Closs et Susan Pintzuk (2008). « Coding the York-Toronto-Helsinki Parsed Corpus of Old English Prose to investigate the syntax-pragmatics interface ». *Studies in the History of the English Language IV. Empirical and Analytical Advances in the Study of English Language Change. Berlin/New York : Mouton de Gruyter*, p. 61-80.

Trips, Carola et Michael Percillier (2020). « Lemmatising Verbs in Middle English Corpora : The Benefit of Enriching the Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2), the Parsed Corpus of Middle English Poetry (PCMEP), and A Parsed Linguistic Atlas of Early Middle English (PLAEME) ». In : *Proceedings of The 12th Language Resources and Evaluation Conference*, p. 7170-7178.

Urieli, Assaf et Ludovic Tanguy (2013). « L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane ». In : *20e conférence du Traitement Automatique du Langage Naturel (TALN)*. Sables d'Olonne, France, (publication en ligne).

Van den Eynde, Karel et Claire Blanche-Benveniste (1978). « Syntaxe et mécanismes descriptifs : présentation de l'approche pronominale ». *Cahiers de lexicologie* 32 :1, p. 3-27.

Van den Eynde, Karel et Piet Mertens (2003). « La valence : l'approche pronominale et son application au lexique verbal ». *Journal of French language studies* 13 :1, p. 63-104.

– (2010). « Le dictionnaire de valence Dicovalence : manuel d'utilisation (version 2.0) ».

van Reenen, Pieter, Evert Wattel, et Margôt van Mulken (2006). « Champagne 1270-1300, Chartes en langue française conservées aux Archives de l'Aube ».

Van Rullen, Tristan, Philippe Blache, Cristel Portes, Stéphane Rauzy, Jean-François Maeyhieux, Marie-Laure Guénot, M-L Balfourier, et J-M Bellengier (2005). « Une plateforme pour l'acquisition, la maintenance et la validation de ressources lexicales ». In : *Actes, Traitement Automatique des Langues Naturelles (TALN)*. TALN. TALN, p. 41-48.

Veronis, Jean (1998). « Multext-lexicons, a set of electronic lexicons for european languages ». *CD-ROM distributed by ELRA/ELDA* 47.

Vijay-Shanker, Kenneth (1987). « A Study of Tree Adjoining Grammars ». Thèse de doctorat. University of Pennsylvania.

Vijay-Shanker, Kenneth et Yves Schabes (1992). « Structure Sharing in Lexicalized Tree-Adjoining Grammars ». In : *COLING 1992 Volume 1 : The 14th International Conference on Computational Linguistics*.

Villemonte de la Clergerie, Eric (2001). « Refining Tabular Parsers for TAGs ». In : *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

– (2002). « Parsing Mildly Context-Sensitive Languages with Thread Automata ». In : *COLING 2002 : The 19th International Conference on Computational Linguistics*.

– (2005). « From Metagrammars to Factorized TAG/TIG Parsers ». In : *Proceedings of the Ninth International Workshop on Parsing Technology (IWPT)*. Association for Computational Linguistics, Vancouver, British Columbia, Canada, p. 190-191.

– (2010). « Building factorized TAGs with meta-grammars ». In : *The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10*. New Haven, CO, United States, p. 111-118.

– (2012). « Etude du traitement de certains compléments de phrase dans le cadre d'une méta-grammaire ». In : *LGC'12 - 31ème 30ème Colloque international sur le Lexique et la Grammaire*. Sous la dir. de Jan Radimsky. Intitut de langues Romanes, University of South Bohemia. Nové Hrady, République Tchèque.

– (2013). « Improving a symbolic parser through partially supervised learning ». In : *The 13th International Conference on Parsing Technologies (IWPT)*. Naria, Japan.

– (2014). « Jouer avec des analyseurs syntaxiques ». In : *TALN 2014*. ATALA. Marseilles, France.

Villemonte de La Clergerie, Eric, Benoît Sagot, Rosa Stern, Pascal Denis, Gaëlle Recourcé, et Victor Mignot (2009). « Extracting and Visualizing Quotations from News Wires ». In : *LTC 2009 - 4th Language and Technology Conference*. Sous la dir. de Zygmunt Vetulani. T. 6562. Lecture Notes in Artificial Intelligence. Springer, Poznań, Poland, p. 522-532.

Villemonte de la Clergerie, Eric, Benoît Sagot, Lionel Nicolas, et Marie-Laure Guénot (2009). « FRMG : évolutions d'un analyseur syntaxique TAG du français ». In : *Journée de l'ATALA sur : Quels analyseurs syntaxiques pour le français ?*

Wrisley, David (2018). « The Open Medieval French Initiative (OpenMedFr) ».

XTAG Research Group (2001). *A Lexicalized Tree Adjoining Grammar for English*. Rapp. tech. IRCS-01-03. IRCS, University of Pennsylvania.

Yu, Juntao (2018). « Semi-Supervised Methods for Out-of-Domain Dependency Parsing ». Thèse de doctorat.

Annexe

Jeux d'étiquettes

Cattex

TABLEAU 1 – Catégories morpho-syntaxiques *Cattex*

Catégorie	Signification	Exemples
VERcjg	verbe conjugué	Quant il la **voit** venir
VERinf	verbe à l'infinitif	Quant il la voit **venir**
VERppe	participe passé	ce que j'ai toz jorz **celé**
VERppa	participe présent	Et aussi fist il en **veillant**
NOMcom	nom commun	en ceste **nuit**
NOMpro	nom propre	**Boort**
ADJqua	adjectif qualificatif	en pechié **mortel**
ADJind	adjectif indéfini	une **aute** nef
ADJcar	adjectif cardinal	apres ces **.ii.** vertuz
ADJord	adjectif ordinal	li **quarz** jorz
ADJpos	adjectif possessif	contre .i. **suen** voisin
PROper	pronom personnel	Et aussi fist **il** en veillant
PROimp	pronom impersonnel	**il** me semble
PROadv	pronom adverbial	il n'**i** dormist ja mes
PROpos	pronom possessif	je i lesseré le **mien**
PROdem	pronom démonstratif	Si estrange leu come **cist** est
PROind	pronom indéfini	si resgardent li **uns** l'**autre**
PROcar	pronom cardinal	si pria a **.ii.** de ses neveuz
PROord	pronom ordinal	Et quant il estoit venuz au **nuevieme**
PROrel	pronom relatif	Ce fu li premiers rois crestiens **qui** maintint le roiaume d'Escoce
PROint	pronom interrogatif	**Qui** estes vos ?
PROcom	pronom composé	**ledit** de Clerieux avoit creü
DETdef	déterminant défini	Ce fu **li** premiers rois
DETndf	déterminant non défini	et mistrent **une** bele tombe sus lui
DETdem	déterminant démonstratif	en **ceste** nuit
DETpos	déterminant possessif	e i firent **son** non escrire
DETind	déterminant inéfini	il sont parfet de **toutes** vertuz
DETcar	déterminant cardinal	et quant il i a esté **.x.** ou **.xx.** ans
DETrel	déterminant relatif	Il ne set **quel** chose puisse avenir
DETint	déterminant interrogatif	**Quele** aventure vos a ça amenez ?
DETcom	déterminant composé	**ledit** roy
ADVgen	adverbe générique	il set **bien**
ADVneg	adverbe négatif	**ne, pas, mie, point**
ADVint	adverbe interrogatif	**Quant** venistes vos ci ?
ADVsub	adverbe subordonnant	il ne set **coment** il i puisse estre venuz
PRE	préposition	il se mist ou grant chemin **de** la forest
CONcoo	conjonction de coordination	**et, mes…**
CONsub	conjonction de subordination	Et de cel serpent est tele la vertu **que se** nus hons tient nule de ses costes
INJ	interjection	**Ha** fet la damoisele
PONfbl	ponctuation faible	, : ; -
PONfrt	ponctuation forte	. ! ?

Suite du tableau à la page suivante

Tableau 1 – suite du tableau

Catégorie	Signification	Exemples
PONpga	ponctuation gauche (ouvrante)	« (
PONpdr	ponctuation droite (fermante)	»)
PONpxx	guillemets droits	” ’

Ces catégories et leurs exemples sont issus du document de référence disponible à l'adresse suivante : http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf.

Alexina et métagrammaires

Catégories morphosyntaxiques

TABLEAU 2 – Catégories morpho-syntaxiques dans les lexiques *Lefff* et *OFrLex* (SAGOT et DANLOS 2008)

Catégorie	Signification
v	verbe
auxAvoir	auxiliaire avoir
auxEtre	auxiliaire être
cln	clitique nominatif
cla	clitique accusatif
cld	clitique datif
clg	clitique génitif
cll	clitique locatif
clr	clitique réfléchi
clneg	clitique négatif
y	pronom y
en	pronom en
pro	pronom (non clitique)
pri	pronom interrogatif
prel	pronom relatif
nc	nom commun
np	nom propre
adv	adverbe
advneg	adverbe négatif
advPref	préfixe adverbial
adj	adjectif qualificatif
adjPref	préfixe adjectival
coo	conjonction de coordination
csu	conjonction de subordination
que	conjonction de subordination *que*
det	déterminant
prep	préposition
pres	interjection
caimp	*ça* impersonnel
ilimp	*il* impersonnel
epsilon	forme à ignorer
etr	séquence en langue étrangère
parentf	ponctuation fermante
parento	ponctuation ouvrante
poncts	ponctuation forte
ponctw	ponctuation faible
sbound	frontière de phrase non lexicalisée
sa	syntagme adjectival
sn	syntagme nominal
sinf	syntagme ancré par un verbe à l'infinitif
scompl	complétive

Suite du tableau à la page suivante

Tableau 2 – suite du tableau

Catégorie	Signification
qcompl	complétive interrogative

Fonctions syntaxiques

TABLEAU 3 – Catégories morpho-syntaxiques dans les lexiques *Lefff* et *OFrLex*

Catégorie	Signification	Argument *TAG*
Suj	sujet	Arg0
Obj	objet	Arg1
Att	attribut du sujet	Arg1
Objà	objet indirect introduit par à	Arg2
Objde	objet indirect introduit par de	Arg2
Obl	complément oblique	non essentiel
Obl2	2e complément oblique	non essentiel
Loc	complément locatif	non essentiel
Dloc	complément “dé-locatif”	non essentiel

Pronoms conjoints

TABLEAU 4 – Pronoms personnels conjoints *le* + *me*

Référence	Contexte gauche	Pivot	Contexte droit
AlexisRa	Pechét	le m’	at tolut.
roland	Deus	le me	doinst venger !
roland	Ben	le me	guarde, si cume tel felon !
roland	Ferut vos ai, car	le me	pardunez ! Rollant respunt...
roland	Dunc	la me	ceinst li gentilz reis, li magnes.
guill1	Tue merci, ben	le m’	as adubé.
guill1	Tue merci, avant her	le m’	adubas.
guill1	Jol vus dirrai quant tu	le m’	as demandé
guill1	Ore est le terme qu’ele	le me	soleit offrir
guill1	Guiburc ma dame	le me	prestad de sun gré.
guill1	Fiz a putein, avez	le me	vus emblez ?
guill1	Sainte Marie	le m’	ad amené.
guill1	Beneit seit l’alme qui	le me	ceinst al lé !
guill1	Qui	le me	irreit hucher
thebes1	Se	la me	rens em pes sanz guerre
thebes1	Tes deus biax eux, or	les me	euvres !
thebes2	Se vous	le me	loez ainsi
thebes2	ferai ma gent apareillier se	le me	voulez conseillier.
ChrSMichel	Se l’en souffrir	le me	voleit.
brut2	E tute rien	le me	destine Que vus encor hui les ventreiz
eneas1	tolir	la me	velt par ravine, mes il lo comparra
eneas2	se gel trespas [...] que de fornir	la me	defaille
eneas2	Alt uns de vos, se	la m’	aport.
eneas2	ne quit noiant [...] que ele ja	le me	mandast.
eneas2	que la terre m’otroieroit, o sa fille	la me	donroit.
floire_jl	quant vos querre	le m’	envoiastes
fresne	De vostre leit	le m’	Alaitiez !
fresne	L’abeesse kil me bailla, A garder	le me	comanda.
lanval	Asez	le m’	ad hum dit sovent
chievrefoil	Plusur	le m’	unt cunté e dit
eliduc	Mil feiz	le me	saluërez.

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
CommPs	car Dominus [...]	le me	dist.
CommPs	Et tu	le me	dunras in multitudine misericordie tue
ProvSerlo	Bele	la me	fai, bele la te ferai.
ErecKu	Sivre	le me	covient adés
ErecKu	Conbatre t'an covient a moi se tu ne	le me	clainmes quite.
ErecKu	Uns chevaliers [...]	le m'	a doné.
ErecKu	Si	le m'	a il mout desfandu...
ErecKu	Qui que il soit, si	le me	dites, puis s'an ira seürs et quites
ErecKu	por ce que je ne vos conui, pardoner bien	le me	devez.
ErecKu	Amis, savroiz	le me	vos dire [...] comant cist chastiax ci a
ErecKu	mes des que feire	le m'	estuet, et c'est chose qu'an feire puet
becket	ja	le me	jurerez
becket	Quant vus	le me	loez, sa volenté otrei.
becket	Li plus privé de lui	le m'	unt mustré en fei.
becket	Uns des convers as monies (ne	le m'	unt pas nummé) Out mult esté grevé
becket	E mi ami de France	le m'	unt fait bien nuncier
BenDuc1b	repairom Senz mal [...] Qu'eisi	le m'	offreiz vos a faire
BenDuc1b	Moct ai poi Deu servi encore, Servir	le me	convient des ore
Fantosme	E s'il ço ne volt faire e tut	le me	desdie,
Fantosme	E a lui frai cuntraire, si Deu	le me	cunsent.
Fantosme	Celui qui	la me	porta gueredun li ert rendu !
CligesKu	Ne ja ne	le me	porloigniez Se otroier le me devez.
CligesKu	Se otroier	le me	devez.
CligesKu	Qui	le me	chalonge ? Ce que je cuit dire mançonge.
CligesKu	Autremant dire	le m'	estuet.
CligesKu	Amors, qui me done a lui tote, Espoir	le me	ra doné tot
CligesKu	Si com mes pansez	le m'	aporte.
CligesKu	Car oï dire	le m'	avez,
CligesKu	Si me crut tant qu'il	le me	dist
CharretteK	Or	la me	bailliez, Et si n'an dotez ja de rien
CharretteK	Se vuel que tu	le me	plevisses, Que tu ne fuies ne ganchisses
CharretteK	Car refuser ne la te doi Des que demandee	la m'	as
CharretteK	Et si m'apele de covant Et mout vilmant	le me	reproche.
CharretteK	et pri Que vos plus ne	le me	celez.
CharretteK	Direiez	le me	vos ? - Je, non, Fet li chevalier
CharretteK	Mes se vos	le me	diseiez, Grant corteisie fereiez
CharretteK	Por ce que tant fole boche as Que vilmant	la me	reprochas.
CharretteK	si	le me	done.
CharretteK	Por ce	le me	doiz bien doner Que jel te cuit guerredoner
CharretteK	Mes ami verai me clamast Qant por	li me	sanbloit enors A feire
CharretteK	Et se vos ja	le me	devez Pardoner
CharretteK	Se vostre congiez	le m'	otroie, Tote m'est delivre la voie
CharretteK	Puis qu'a dire	le me	besoigne.
CharretteK	Se je onques	le me	pansai.
CharretteK	Riens nule retenir nel puet, Que il	le me	jura sor sainz Qu'il vanroit
CharretteK	Et maintenant	le me	mandez.
CharretteK	Mes se il	le me	vialt noier, Ja n'i loierai soldoier
CharretteK	Malemant	la m'	as bestornee, Car g'iere el mont
CharretteK	don, et vos	le me	donastes Mout volantiers quant jel vos quis...
YvainKu	Quant vos tant	le m'	avez celé.
YvainKu	Autrui que toi n'en doi blasmer, Que tu	le m'	anbles a veüe.
YvainKu	Que nes veoir ne	le me	lez, Celui qui est si pres de moi.
YvainKu	Or	le me	dites, Si soiez de l'amande quites
YvainKu	se je ne le comant Et mes consauz ne	le m'	aporte, Ne vos iert overte ma porte.
beroul	Car	le me	faites delivrer...
beroul	Et il m'ont dit Qu'il	le m'	ont dit.
eracle	car li angles	le m'	encarja et cose u Dius ait rien a faire
eracle	Por amor Diu	le me	dona que onques plus mot n'i sonna

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
eracle	Or	le me	fait besoigne vendre
eracle	Boinairement	le me	desis et acroire le me fesis
eracle	Boinairement le me desis et acroire	le me	fesis que il afoleroit por courre
eracle	Or amerai si serai large, car Amors fine	le m’	encarge que je le soie
eracle	Je vos tienç molt a deceü que vos	le m’	avés tant teü.
eracle	Biaus sire, Amors	le me	fist faire.
PercevalKu	je vos salu, si con ma mere	le m’	aprist
PercevalKu	cui qu’il soit grief, que ma mere	le m’	anseigna
PercevalKu	S’il	les me	done, bel m’an iert
PercevalKu	se il ne vialt tenir de moi sa terre, que il	la me	rande, ou il anvoit qui la desfande vers moi
PercevalKu	Li rois, fet il,	les me	dona.
PercevalKu	si	le m’	anseigna a savoir
PercevalKu	se tu	le me	deïs por mal
PercevalKu	et il	la me	dona.
PercevalKu	mes tant fist que il la beisa par force, si	le me	conut.
PercevalKu	ce que vos voldroiz [...] que ensi	le me	comanda li vaslez
PercevalKu	Par ton gabois tolu	le m’	as, si que ja mes nel cuit veoir.
PercevalKu	sire Kex, plus belemant [...]	le me	poïssiez dire.
PercevalKu	Or an pansez, que je m’an vois, que il	le m’	estuet sivre el bois.
PercevalKu	Mes sire li rois qui est ci	le me	comande et ge le di
PercevalKu	l’an	le me	devroit atorner a trop leide recreantise
PercevalKu	puis que tu	le m’	as demandé.
PercevalKu	Et neporquant [...] panre et amener	le m’	alez, si seroiz quites de mon fié.
PercevalKu	Amis, quant tu	le me	consoilles, a ton consoil me voel tenir
PercevalKu	Einz	le m’	avroiz acreanté.
PercevalKu	Or	les me	nomez.
PercevalKu	puis que Dex veoir	le me	lesse.
SBernAn	et ci ne	la me	proichet mies solement apermenmes
SBernAn	la cause por kai il ceu facet encerche et om	la me	dist.
SBernAn	Ancor ne	la me	doignes tu mies, !
SBernAn	selonc ceu qu’il doneir	la me	uuelt uos couf iu, !
SBernAn	si dist.	le me	repent de ceu que ie ai fait l’omme.
SermMadn	Sire, puis k’il ne	le me	loit veir, donez le moi sentir par vo grace.
adgar	E beneit seit [...] Cil ke comencier	le me	fist E par ki jol faz en avant !
adgar	Laissez ester ! Bonitus	la me	deit chanter.
adgar	Faites	le me	aveir sanz delai
SBernCant	ke ge l’oüsse rendut a ceaz ki	le m’	avoient comandeit
BlondNesle	La bele qui bien	les me	puet merir.
BlondNesle	A morir sui livrez, Se trop	le me	delaie.
BlondNesle	Quant sa biautez	la me	fist acointier !
BlondNesle	Et la bele	le me	desfent
BlondNesle	Tantes biautez i remir, Quant	la me	loist reguarder
qlr	Il	le me	dunad á sun plaisir
qlr	E si tu	le me	céiles, ícel mal vienge sur tei
qlr	El lit	le me	portez
qlr	El lit le me portez é tost	le m’	Ociez !
qlr	É savereíez	le me	vus mustrer ?
qlr	vifs les pernez é tut liëz	les me	Menez !
qlr	é Deu	le m’	ad celed é nient nel m’ad mustred.
qlr	Nostre Sire	le m’	a moustre ke tu seras rois de Syrie.
tydorel	Demandai li qui il estoit, dit moi qu’il	le me	mostreroit.
tydorel	Il savoit bien certainement e bien	le me	disoit sovent
amiamil	Tout orendroit	le m’	a il fiancié
amiamil	Par Deu, bien	le me	samble.
amiamil	Voz la panréz, que li rois	le m’	a dit.
amiamil	Miex voz venist que	le m’	envoïssiéz, Que voz folie ne mal en feïssiéz.
desire	se vus ja mes un liu veez ke vus rendre	le me	pussez, Sire, ne me oblïez mie.
villehard.	et vostre gent	le m’	ont tolu

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
villehard.	et vos m'aviez convent que vos	le m'	aideriez a conquerre
belinc	Por coi je vois en cest afaire, Dius	le me	doinst a bon cief traire !
belinc	Volentiers vos prendrai a fame, Se Artus	le me	velt loer
belinc	Tant	le me	convenra soufrir Con il vos venra a plaissir.
belinc	Irai je, u je remanrai ? Ma dame	le m'	a desfendu
belinc	grant merchis vos en renc De ço que vos	le m'	envoiastes
dole	li cuers	le m'	a toz jors bien dit.
dole	un tel cop m'i dona [...] que tot	le m'	enbarra
dole	Ha ! ha ! Juglet, or i parra, com vos	le me	saluerez.
dole	Or	le me	pardonez.
dole	Ge sai bien qu'il	le me	donront
dole	et si tost com il	le m'	avront creanté deboneremen t
dole	Dame, de ce sui ge dolenz, mes il	le m'	estuet a soufrir.
dole	Trop	le m'	a vendu son joël
dole	Une pucele	le me	dit qui en porta mout les messages.
dole	vos dites ci qu'il	le me	nie, qu'il onqes n'ot mon pucelage
aucassin	Avés	le me	vos tolue ne enblee ?
aucassin	se vos ne	le m'	afiés, se je ne vous fac ja cele teste voler.
aucassin	quant si belement [...]	le m'	as ore dit.
aucassin	et Dix	le me	laist trover !
renart10	si con il	le m'	a en convent...
renart11	s'on	le me	loe.
renart11	Vos me mantez la vostre foi, or	la m'	avez.
renart11	tenez vos [...] que orandroit	le m'	envoia mesire Huon le doien
gcoin1	A traitier si bien	la m'	apregne Que boen essample
gcoin1	Si com mes cuers	le me	devine
gcoin1	Que diras tu ? Car	le me	di ! Chaitive, adonques que diras ?
gcoin1	Que diras tu ? Car	le me	di ! Lasse, se tu parler peüsses
tdechamp	Meus aim que [...] don Un regart, quant	le me	lance.
tdechamp	Or	la m'	estuet servir
tdechamp	Et qant plus avrai cheance, Plus	la me	couvient douter.
tdechamp	La ou Amors	la m'	amena veoir Oi je adès un tres douz atochier
tdechamp	Grant	la me	fist, quant le cuer a de moi.
tdechamp	Baudoÿn [...] quant	le me	fera, La bele que je n'os nonmer...
tdechamp	A terre lez	li m'	assis.
tdechamp	Morir m'estuet, s'Amors	le me	consent
qgraal	sire, fait ele, vos	le me	celez, por quoi faites vos ce ?
qgraal	vos le conoissiez si certainement vos	le me	poez bien dire.
qgraal	A non Dieu, fet ele, puis que vos nel	le me	volez dire je le vos dirai.
qgraal	Mes la grant amor que j'ai [...]	le me	rueve dire
qgraal	Diex, fait ele, mes cuers ne	le me	dit pas qui me met en toutes les mesaises
qgraal	devant que aventure	le m'	ameint.
qgraal	et por ce	le m'	a il enchargié que je le vos baillasse.
qgraal	et il le nos contera, car autresi	le m'	a il promis.
qgraal	Si vos pri que vos	le me	donez
qgraal	s'est [...] si durement corrociez... qu'il	le m'	a bien mostré puis ersoir.
qgraal	Mes puis que einsint est avenu a soffrir	le me	covient
qgraal	et por ce vos pri ge que vos	le me	dioiz.
qgraal	tant com il soit en ma garde, par force	le me	poez vos tolir.
qgraal	Por ce, fet cil, qu'il	le m'	a tolu a force si m'en a mort et maubailli
qgraal	donc vos soiez a conseillier se vos	le me	dites que je ne vos en conseil
qgraal	ne conoistroie je mie la senefiance se vos ne	la me	disiez.
qgraal	Vos	les m'	avez si bien devisees
qgraal	qu'il	le me	covenoit fere.
qgraal	car einsi	le m'	a mandé li haulz sires.
qgraal	Si vos pri por Dieu que vos	le m'	otroiez.
rosel	Or veil cel songe rimeer [...] qu'Amors	le me	prie et comande.
rosel	ne voie ne leu [...] ne hom nez qui	le me	mostrast n'iert ilec, que j'estoie seus.

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
rosel	Veoir	la m’	estuet
rosel	Sanz faille, Amors	le me	fist fere, dont je ne puis mon cuer retrere
rosel	et si	le m’	a il pardoné en la fin
rosel	Lors	le m’	a Franchise envoié.
rosel	Jes cuidoie avoir achetez ; or	les me	vent tot de rechief
fournival	Mais paour m’est prise Qi	le m’	a tolu
fournival	Bien vueil qu’autres	le m’	ost bien reprochier
fournival	Quar on ne	le me	doit mie Torner a si grant folie
ChirAlb	et il	le me	dist.
atrper	Quant ele	le m’	a conmandé, Des qu’il li plaist et vient a gré.
atrper	Qui ne counoist pas le païs ? S’or	le m’	avoient leu ocis, U aucune beste sauvage
atrper	Amis, fait il, ce est plus bel Que	la me	bailliés par amor
atrper	Sire, fait il, quant isi va Que rendre	le m’	estuet en fin
atrper	Aprés none m’i conbatrai, Si com vous	le m’	avés loé.
atrper	Si com le conte	le m’	afice, Ki plus ert blanc que nule flor
atrper	Car se vous	le me	poés rendre, Dont m’arés vous toute garie
atrper	Des que vous	le m’	avés promis
atrper	Si n’estoit pas seant ne bel Que plus	le me	contretenist.
atrper	dire [...] que je l’ai perdu, Si ne sai qui	le m’	a tolu.
atrper	Or	le me	couvient aler querre
atrper	Or	le m’	a ci cest chevalier Por la soie amor ramené.
atrper	Mais or voi qu’il	le m’	estuet fare
atrper	Si vous pri que	le me	dounois.
atrper	Et se Ragidiax	le m’	otroie J’en menrai la soie et la moie
atrper	bien savoir [...] n’en doutés ja, Si conme cil	le me	conta.
atrper	Covenant l’oi en la bataille, Et Gavains	le m’	a creanté.
bestam	Certes, douce dame, je croy, Ja monstré ne	le m’	eüssiés, Se courroucie ne feussiés
vergy	onques mes ne	le m’	osa dire.
vergy	ne tant ne la tenisse a voire, se ce ne	le me	feïst croire et me meïst en grant doutance
menreims	Biaux fillues, faites penre une corde, et si	la me	faites metre ou col
menreims	li cuers	le me	dist.
menreims	si comme mes cuers	le me	devine
menreims	En non Dieu [...] ainsi n’ira il pas, vous	les me	rendrez, et en sera chascuns de vous
menreims	je ne vous en saurais gré, et se vous	le me	rendez, je vous pardonrais mon mautalant.
MirNDCh.	Ce qu’il veult ne puis contredire ; Donques	le m’	estuet il a dire...
MirNDCh.	Si com li livres	le me	chante.
RutebZ	Bien	le m’	ot griesche en couvent Quanque me livre...
RutebZ	soul en mon ostei. Je cui li vens	les m’	at ostei, L’amours est morte...
rosem1	Dame Oyseuse	le me	fist faire...
rosem1	Bien	le m’	avoit Reson noté
rosem1	ou, s’il li plest, qu’il	le m’	amande, ou j’en prendré par moi l’amande
rosem1	je la vodroie bien entendre, s’ous	la me	voliez aprendre.
rosem1	Povreté m’a vaé le pas, a l’issir	le me	deffendi.
rosem1	Fortune ainsinc	les me	toli par Povreté, qui vint o li.
rosem2	m’an sai de tant c’onques	le me	pansai et qu’audience li doné
rosem2	des larrons qui	le me	renoient quant il ont fet ce qu il queroient.
rosem2	qui coutoit plus de mil besanz [...] si	le me	metoit l’en asseure
rosem2	Se je puis riche home baillier, vos	le me	verrez si taillier qu’il n’avra ja tant mars
rosem2	c’est chose fete si tost con	la m’	avrez retrete...
rosem2	qu’en	le me	dist, et jou redis, et que cil la rose besa
rosem2	Toutevois, se	le me	demande, que puis je dire a sa demande ?
rosem2	Car il	le me	convient repondre
rosem2	Pieça que bien	le me	disoient li ribaut
rosem2	il ne me prisoit un pois, et bien	le me	disoit.
rosem2	Mes puis qu’il	le m’	a presanté, et receü son presant
rosem2	qu’Amors par confort	le m’	anvoie.
rosem2	pri vos que	le me	pardoignez, et de par moi leur respoignez
rosem2	bien me trova fol debonere, deable	le me	firent fere.

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
rosem2	tant vos fiaiz an moi que vos	le me	diaiz.
rosem2	Por ce pri que vos	le me	dites par guerredons et par merites
rosem3	je li conteré sa cheance devant Dieu, qui	le me	bailla quant a s'ymage le tailla
rosem3	puis que saluez	les m'	avroiz si con saluer les savroiz
rosem3	Touz	les me	dist, onc puis ne sis
rosem3	et pri que tu	le me	pardoignes
rosem3	se trovasse qui	le m'	offrist, ou...
rosem3	ou, san plus, qui	le me	soffrist.
rosem3	Nature, qui	la me	bailla, des lors que primes la tailla
SGenPr1	Or	le me	bailliéz et elles li baillierent
SGenPr1	or me dites que vous portéz et ne	le me	celéz mie !
SGenPr1	Marcheanz	la m'	ont vendue.
grchron1	tout en la maniere que tu	le m'	orras raconter.
JantRect	Por coi ne	le me	donez vos ?
beauma1	et il	la me	corront et fet l'anieus
beauma1	se je ne di pour quoi et de quoi il	les me	doit.
beauma1	si comme se j'ai un cheval et.III. homme	le me	demandent, et dit chascuns qu'il est siens
beauma1	je li puis redemander arrieres et	les me	doit rendre s'il ne prueve
beauma1	Sire, tel cheval qu'il me demande il	le me	vendi tel nombre d'argent
beauma1	car vous ne	les me	baillastes pas a ferme
beauma1	ce que coustume suefre a donner [...] il	le me	convient soufrir.
beauma1	Je doi poursuir celui qui	la m'	osta par action de larrecin
beauma1	se je puis savoir qu'il	la m'	ostast par courage d'embler
beauma1	li sires ne	le me	puet mes demander
beauma1	Et s'il	le me	renvoie dedens les.
beauma1	tant comme il le tient et il	le me	renvoie
beauma1	lb. ou qu'il	le me	lesse pour les.
beauma1	tout ne	le m'	eust il pas mandé
beauma1	j'i met aucuns cous resnables [...] il	les me	doit bien rendre
beauma1	et on	la m'	oste de ma main
beauma1	tout soit ce que cil qui	le m'	empeeche n'en port pas la chose
beauma1	Mes, moi resaisi, se cil qui	la m'	osta prueve la chose a sieue, il la ravra.
fauvel	Force d'amour	le me	fait faire.
passpal	Ainz veil que vous	le me	menés A Herode, puis li direz
passpal	Mais ton maitre	la me	sena Pour ce que eschaper cuida.
passpal	Or	le me	tien bien a cest post
passpal	Si	le m'	amenez devant moy
passpal	Don es tu roys ? Or	le me	di.
passpal	Bien ait li roys qui	les m'	envoie ! Bien set grant mestier en avoie.
passpal	Quar tu meïmes	le me	deïs Que grant pieça leur as promis.
moree	Et nulle personne par droit ne	le me	puet reputer pour mauvestié
GMFort.	Mais Amours	le me	firent faire Qui m'i donnerent ligement
GMFort.	Et l'esperance doublée Qui de	li me	vient.
GMFort.	riens n'en sceüst Qu'elle fait faire	le m'	eüst.
GMFort.	Et si tost qu'elle dit	le m'	ot, Je n'eüsse dit un seul mot
GMFort.	Puis que vous	le me	commandez
GMFort.	amour que tant vous ay couvert [...] Einsois	la m'	estuet descouvrir
GMFort.	Et veïstes vous Esperence [...]	le m'	avez devisé ?
GMFort.	Esperence	le me	donna, Quant a moy tant s'abandonna
GMPlour	La mort pri que la me maint, Car	la m'	ottroy.
prov	Bele fame est a poinne chaste. Bele	la me	fai, bele la te ferai.
hlanc	qe le deable par sa force ne par ses engyns	le me	fesoit faire
hlanc	Trois choses	le me	fra faire...
hlanc	Jeo le voille avoir : aletz, si	le me	queretz
hlanc	reconoistre	le me	covient a vous
hlanc	certes grande bosoigne	le me	fait faire
hlanc	si vous ne	le me	donetz
Berin1	le seneschal mesmes de ceste ville	le m'	a par pluseurs foiz demandé

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
Berin1	ainsi comme vous	le m’	avez en couvent
Berin1	pour tant que tu ne	le me	scés recorder, je te diray que tu feras
Berin1	Ainsy	le m’	avoit on dit.
Berin1	Honnis soit qui l’engendra par amours ! Or	le me	bailliez
Berin1	se mi baron	le me	loent
Berin1	ainsi que vous	le m’	avez creanté
Berin1	s’il vuelt faire la bataille, si	le me	face savoir briefment
Berin1	Sy vous prye que vous	le me	dïés, par quoy je sache
Berin1	Sy te conjur de quanques je puis que tu	le me	dïes tost et hastivement.
Berin1	car bien sçay que pour bien	le m’	avez dit.
Berin1	Et je l’ottroy [...] puis que vous tous	le me	loez.
Berin1	ne la vols donner a nulle personne qui	la me	requist
Berin1	Si vous pri [...] que vous en paix	la me	laissiez, si ferez sens et courtoisie
Berin1	Si lui pry pour Dieu et pour courtoisie qu’il	le me	perdoint.”
Berin2	Et puis que Dieu ne le veult, il	le me	convendra souffrir.”
Berin2	et qu’il	le me	baillast.
Berin2	et sachiez que ma mere	le me	donna afin que...
Berin2	Par amours [...] damoiselle, or	le me	baillez pour moy aydier.”
Berin2	que vous m’aiez en convenant que vous	le me	ferez tout vif bailler.
Berin2	nouvelles [...] que vous	le me	faciez orendroit rendre prins
Berin2	toille dont je estanchie mes plaies et	les me	benda moult debonnairement
Berin2	Si vous prie que vous	le me	vueilliez dire.
Berin2	le filz d’un poissant roy	le me	dist et me donna unes telles enseignes
Berin2	en alez a Dieu [...] et si	le me	salüez et je vous en prie.
Berin2	Et vous dy [...] que l’empereur	le m’	a fait forjurer
Berin2	si vous pry que vous	le me	dïez.
Berin2	Je vous jure sur sains que ceste vielle [...]	le me	dist et encusa, et ne m’en donnoie de garde.
Berin2	Chevalier, qui es tu et que quiers ? Or	le me	diz tost sanz celler.
Berin2	terre l’espace de sept ans et plus [...] et	la me	donna mon pere en heritage
regcrim1	ne le varlet aussi ; car il n’est pas à moy, et	le m’	a presté le bailli de Senz
regcrim2	Se vous ne	le me	dittes, je vous courrouceray jusques au corps.
regcrim2	vous prie et requier [...] que vous	le me	conseilliez.
dictier	Las ! qui	le me	fist penser ?
melusine	se je y mespren [...] qu’ilz	le me	veuillent pardonner
melusine	je prenoye toute la plaisance [...] et vous	la m’	avez tollue.
melusine	Mais ou alez vous a ceste heure, se vous	le me	povez bonnement descouvrir ?
melusine	ou il	le me	plaira a prendre
melusine	il n’a homme ou monde que je craingne qui	le me	puist oster
melusine	grant yre […]	le m’	a fait faire.
melusine	On	le m’	a bien tout compté et dit.
melusine	a ce diamant vous donrez au moinsné, et	les me	saluez beaucop de foiz.
melusine	Or alez, dist le roy, et	les me	faictes cy venir demain dedens prime.
melusine	se non qu’ilz	le me	facent savoir.
melusine	Et toutesfoiz, se tu ne tournes, faire	le me	convient.
melusine	mais raison naturelle	le me	deffent, pour ce que vous estes mon frere.
melusine	Or	le me	veulz faire comparer.
melusine	Or	le me	fault perdre par toy, faulse borgne
melusine	ces lecheours moynes de Malleres	le m’	ont enchanté
melusine	il n’a personne ceans qui oncques	le me	conseillast
melusine	Et ainsi	la me	fauldra porter et souffrir
melusine	ou il me rendra mes arrierages [...] ou vous	les me	rendrez.
melusine	si	le me	veullent pardonner
menagiera	tel	le me	fist faire et je ne m’en donnoie garde
menagierb	Mon mary	le m’	a commandé.
menagiera	car je suis certain qu’il	le me	deffendroit.
menagierb	Vous ne	le m’	avez point autrement commandé, etc.
menagierb	Vostre cousin	le me	conseilla ainsi a faire.
maniere1399	Dame, vous	le me	comanderez.

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
maniere1399	Ditez coment	le me	donrez vous a droit.
QJoyesKar	Vrayement, fait il, vous	le me	direz.
QJoyesKav	Par mon serement [...] si mon mary	le me	fasoit ainxin
QJoyesKar	puisque vous	le me	feistes.
QJoyesKav	car il	le me	doit bailler demain
QJoyesKav	Je ameroy mieulx [...] que vous	la me	baillassez du tout que la batre ainxin
QJoyesKav	comment les pourroye ge sçavoir si vous ne	les me	disiez ?
QJoyesKar	surs elles come surs les hommes, elles	le me	pardonront si leur plest
cdo_retenue	Sien estre vueil, se	le me	commandés
cdo_retenue	Je croy que non, car ainsi	le me	semble
cdo_ball.	Mais pardonner	le me	devés Et n'en devés autruy blasmer
cdo_ball.	Qu'il soit ainsi bien	le me	fist aprandre Ma maistresse
cdo_ball.	sitost que je le tenoye, Dangier	le me	venoit tolir Ce peu de plaisir que j'avoye.
cdo_ball.	Dieu	le me	sauve ce varlet ! Il est enroué devenu
cdo_ball.	Dieu	le me	sauve ce varlet ! Rompre ne sauroit un festu
cdo_ball.	Quant ou cellier sont en secret ; Dieu	le me	sauve ce varlet !
cdo_ball.	Et ses esperons d'un foret ; Dieu	le me	sauve ce varlet !
cdo_compl.	Mais les griefs maulx	le me	font faire
cdo_rond.	Ainsi Raison	le me	conseille.
cdo_rond.	On	le me	devra pardonner
cdo_rond.	Qu'il ne	le me	font Pour voir que feroye Et...
cdo_rond.	Qu'il ne	le me	font Pour voir que feroye !
cdo_rond.	Payer les vouldroye [...] ; Qu'il ne	le me	font !
cdo_rond.	En françoys	la m'	a translatee, Comme tressouffisant et saige
cdo_rond.	L'eaue d'Espoir [...], Soif de Confort	la me	fait desirer
monstre	Lequel [...]	le me	a bénignement otroié.
jouvencel1	car ilz	le m'	ont mandé
jouvencel1	qu'il arriva ung tas de gens sur moy, qui	le me	voulurent tuer
jouvencel2	monseigneur le conte	le m'	a laissé bon
jouvencel2	je ne vous donnay ma foy ne aussi vous ne	la me	demandastes pas.
jouvencel2	s'il plaist à Dieu	le me	octroyer.
jouvencel2	mettre une embusche [...] ainsi que	le m'	aviez chargé.
jouvencel2	Je vous pri, ne	le me	celez plus
jouvencel2	Si vous pri que	le me	pardonnez
quenouilles	je vous supplie	le me	pardonner et laditte faulte imputer...
quenouilles	celles qui par si tres grant haste	le me	disoient, que loisir ne temps n'avoie...
quenouilles	et aincoires puis que dire	le me	convient
quenouilles	quant je portoie ma fille [...] et	le me	fist et aprist ma tante
jehpar	si le courage	le me	conseille.
jehpar	Certes, mon redoubté seigneur, il ne	le me	fault pas demander
ressource	Et pour mon corps mieulx experimenter, Il	le me	fault de tous poincts tourmenter
commyn1	Ilz sont dedans ! Ainsi	le m'	ont compté plusieurs depuis.
commyn3	Tousjours	le m'	avoit accordé, jusques à celle heure
commyn4	Mais le roy	le me	compta, ne plus ne moins que je vous deïz
commyn5	Et le roy, qui depuis	le me	conta, l'entendoit bien
commyn5	qui ne veulx tenir oppinion [...]	le me	semble ainsi
commyn5	Il	le me	fault bien garder.
commyn6	avec ledict duc [...] et puis	le me	feït assavoir.
commyn6	et de peur de perdre obeyssance, car ainsi	le me	deïst luy-mesmes.
commyn7	combien que ne fusse present [...] si	le m'	ont compté le roy, ledit duc et aultres.
commyn7	Ceulx qui traictoient avecques ledit Pierre	le m'	ont compté, se mocquant de luy
commyn7	et presta au roy lors trente mil ducatz ; et	le m'	a dit [...] que on luy promist
commyn7	et	le m'	ont monstré les Chartreux
commyn7	Et ung natif de Bourge	le m'	appella sainct
commyn7	et veïs la lettre, car il	la me	monstra
commyn7	et ainsi	le me	disoit l'ambassade du roy Alphonce
commyn7	et m'en monstroient plusieurs lettres ou	le me	faisoient dire par ung de leurs secretaires.
commyn7	aymoyt fort ceulx du mont [...] (lesquelz	le m'	ont compté à Venise

Suite du tableau à la page suivante

Tableau 4 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
commyn7	autant en avoit mandé à Pierre de M., qui	le m’	a dit.
commyn8	car l’ambassadeur de Napples	le m’	avoit dict, cuydant que jà y fussent.
commyn8	disoit que en tout en avoit neuf mil ; et	le me	dist depuis nostre bataille, dont sera parlé.
commyn8	comme	le m’	ont compté des plus grans de la duchié.
commyn8	car des hommes [...]	le m’	ont compté
commyn8	je leur donneroye trop cueur et que on	le m’	avoit dit trop tard.
commyn8	et l’ung des providateurs s’y accordoit, qui	le m’	a compté, et l’autre non
commyn8	Depuis	le m’	a compté
commyn8	et tous les chiefz	le m’	ont confessé
commyn8	et que [...] il	le m’	escriproit de sa main,
commyn8	et	le m’	a maintes fois compté
commyn8	l’evesque d’A. et ses prouchains chambellans	le m’	ont compté
commyn8	luy a plussieurs foiz escript [...] ; et à moy	le me	dist de bouche quant je parlay à luy

TABLEAU 5 – Pronoms personnels conjoints *le* + *te*

Référence	Contexte gauche	Pivot	Contexte droit
guill1	Des hui matin	le t’	ai fait apareiller
guill1	Ainz que moussez,	le te	di jo assez Ja nel purriez soffrir ne endurer
ChrSMichel	Et de bon cuer l’enoreras De meie part,	le te	di bien, Ja nen auras besoig de rien
eneas1	com il naistront, en aprés	les te	nomerai et les batailles te dirai
eneas2	ce puez savoir que donc	la t’	estuet il avoir
ProvSerlo	Bele la me fai, bele	la te	ferai
ErecKu	Se tu viax avoir l’esprevier, mout	le t’	estuet conparer chier
ErecKu	Tien m’espee, je	la te	rant
becket	Les proveires ne deiz enseignier [...], Ensiwre	les t’	estuet, devant deivent aler
BenDuc1b	Il	le t’	eüst bien a desfendre E a delivrer
BenDuc1b	Pren l’ajue, s’il	la te	fait
CligesKu	Comant	le t’	a donc trait el cors
CligesKu	Comant	le t’	a il tret
CligesKu	Si ne	le t’	a crevé
CligesKu	Leisse ton duel, si te conforte, Car se vive ne	la te	rant, Ou tu m’oci ou tu me pant
CharretteKu	Et ausi avrai ge de toi, Car refuser ne	la te	doi Des que demandee la m’as
CharretteKu	Que ce que il est venuz querre Li done ainz qu’il	le te	demant
YvainKu	Rant li, qu’a randre	le t’	estuet
eracle	Dius	le te	mande ci par moi qui a te parole entendue
eracle	Dius	le te	sara bien merir qui ert as premiers cols ferir
eracle	Cil qui fu nés en Belleem	le te	mande del ciel lassus
PercevalKu	et tu viax que je	les t’	apraingne
PercevalKu	Et se tu nel trueves iqui, bien iert qui	le t’	anseignera
PercevalKu	Or me di, frere debonaire, ces armes, qui	les te	bailla
PercevalKu	Et se ge	le t’	ai desfandu, ge n’i ai nul mal antandu
PercevalKu	Au roncin	le t’	estuet changier don l’escuier as abatu,
PercevalKu	il vient por ta male avanture, ne	le te	celerai ge pas
PercevalKu	Et je, fet il,	le te	recroi sor ta fiance et sor ta foi
SBernAn	s’om ne	la te	donet tot en pardons
qlr	demain ú puis demain [...] hastivement	le te	manderai
qlr	Enquier des tuens, é issi	le te	dirrunt
qlr	é la vígne que tu desíres, jó	la te	durrái
qlr	é vá saisír la vígne Nabóth de Jezrael ki ne	la te	vólt otréiér ne par eschánge ne pur avéir
villehardouin	et te jurerons sor sainz, et	le te	ferons aus autres jurer
renart10	Por ce	le te	di
renart10	Et tu le faiz, s’i	le te	proie, si fera il ce sai ge bien
renart10	Se Lietart t’a eü pené, il	le t’	a bien guerredoné
renart11	Je	le te	dirai ja
renart11	ja part [...] male ne bone, s’autre raison ne	le te	done

Suite du tableau à la page suivante

Tableau 5 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
renart11	par foi, Renart, je	le t’	afi, ne m’as encor gaires servi
qgraal	Tant m’en as conjuré, fet li chevaliers, que je	le te	dirai, mes ce ne sera mie a toi sol
qgraal	Je	le te	dirai, fet li preudons, or m’escoute
qgraal	Si	le te	vint si tost dire
qgraal	Donc	la te	diré je, fait li preudons, or m’escoute
qgraal	que tu fusses herbergiez fors ou paveillon, ele	le te	fist apareillier, et quant ele t’apela si dist
qgraal	Si	le t’	ai dit et fet conoistre
qgraal	Donc Nostre Sires se dut corrocier a toi, et bien	le te	mostra en ton dormant quant il te vint dire
qgraal	tes freres n’est pas ocis [...] mes il	le te	dist por ce qu’il te vouloit fere entendant folie
rosel	ton servise en gré [...] se mauvestié ne	le te	tost
SLambert	aies merci de cels qui	le te	requierent
ChirAlbTb	ou par fil, si com je	le t’	ai fait savoir
ChirAlbTb	s’aucune chose ne	le te	deffent, ou se li os n’est muez
MirNDCh.	Se qu’en ton cuer li as promis Li porte, ge	le te	conseil
rosem1	se mauvestié ne	le te	tost
rosem1	Dire	le te	veill sanz demeure, car la te convient il aler
rosem1	car usurier, bien	le t’	afiche, ne porroient pas estre riche
rosem1	Si	la te	veill or ramantoivre por toi fere mieuz aperçoivre
grchron1	je	le te	rendrai maintenant
JAntRect	je demanderai l’exemple de ceste chose [...] et	le te	moustrerai
fauvel	Je	le te	monstre en tous estas Par argumens non intestas
fauvel	De Ninivé je	le te	preuve, Une cité de quoi l’en treuve
passbonnes	De l’uile de misericorde Aras de moy, je	le t’	acorde
passpal	S’il ne feut de mauvese vie, Nous ne	le t’	amenissions mie
moree	mais a toy le dy, qui n’es mie homme, et	le t’	afferme en verité que il est ainxi comme je le di
GMFort.	Se tu ne scez que c’est a dire, Monstrer	le te	vueil et descrire
GMFort.	« Amis, et je	le te	diray Volentiers, sans faire lonc plait
GMFort.	Prouvé	le t’	ay, se tu le nies
GMFort.	« Oïl, je	le te	prueve
prov	Bele la me fai, bele	la te	ferai
prov	Ne viel n’enfent, femme ne fol ne servir ja, je	le te	lo
daudin	requier qu’on	le te	moustre
Berin1	si prie a Dieu qu’il	le te	envoye selon ce qu’il scet qu’il t’est bien
Berin1	puis que tu m’as conjuré, je	le te	diray
Berin2	Maiz sachez que je ne	le te	puis dire ne par moy ne le saraz
Berin2	Et pour ce, je vueil que tu saches et	le te	jure par le souverain roy des cieulx
Berin2	que donnas tu a ce povre quant je	le te	commanday
melusine	croy qu’il est tout vray, tout aussi vray comme je	le t’	ay dit
melusine	Mais puis que mon nom veulz savoir, je	le te	diray, car pour toy ne le celeray je pas
melusine	je	le t’	avoye pardonné en cuer
melusine	et Dieu	le t’	eust pardonné
melusine	Je pry a Dieu qu’il	le te	veulle pardonner
melusine	Par foy, dist Gieffroy, je	le t’	asseure
melusine	Et quant de tes dix sols, je	les te	quicte
menagier	Mieulx vault que je	la te	donne que a ung autre homme
maniere1399	Je	le te	pardonne
quadrilogue	Or	le te	fault a present regreter

TABLEAU 6 – Pronoms personnels conjoints *le* + *te*

Référence	Contexte gauche	Pivot	Contexte droit
AlexisRa	reçut l’almosne quant Deus	la li	tramist, tant an retint dunt ses cors puet guarir
gormont	sil fiert sur la targe novele qu’il	la li	freint e eschantele ; sa hanste brise par asteles.
gormont	sil fiert sur sun escu bendé k’il	la li	ad freit e quassé, le hauberc rumpu et desafré ;
gormont	sil fert sur la targe novele qu’il	la li	freint e eschantele ; sa hanste brise par asteles ;
thebes1	et Thideüs	la li	otroie, car il de riens ne s’i foloie ;

Suite du tableau à la page suivante

Tableau 6 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
thebes1	s'il	le li	fet, qu'il ne s'en venge et sa terre ne li blastenge.
thebes1	Pollinicés la fin prendra, qui poursuivre	la li	voudra.
thebes1	l'enchace, et cil l'atent, fiert le en l'escu, tot	le li	fent ;
thebes1	mout tost, isnel et preuz et bien delivre ; il l'apele, cil	le li	livre ;
thebes1	il demande souvent Ysmaine et Jocaste	la li	ameinne :
thebes2	Il demande ou, cil	le li	dient, li rois en doute, cil li afient.
thebes2	La pucele	le li	otroie : « Ne sai, fet il, conment vous croie.
thebes2	Or	le li	a fait pardonner la proiere d'une meschine !
thebes2	il l'ont mout malement estors, freschement	le li	trest uns Mors.
thebes2	hante ot frete, s'espee tret, fiert l'en l'escu, tot	le li	fent ; Alissandres as poinz le prent.
thebes2	petit et petit	le li	emble, ne li dies pas tout ensemble ;
thebes2	ne li dies pas tout ensemble ; bien	le li	pourras atremper, ne te hastes pas du conter ;
thebes2	le branc d'acier [...] Acastus	le li	court tolir ;
ChrSMichel	A l'apostoile obeïsseit Qui bien mandei	le li	aveit.
becket	Robert de Herefort	la li	va demander
becket	Des mains	la li	voleit par vive force oster.
becket	Li reis dit qu'a dous cenz	les li	fera jurer Chevaliers e proveires.
BenDuc1b	Mais unc n'en voct li dus rien prendre, Ainz	les li	renveia arriere.
CharretteKu	Mes sanz ranpone et sanz vantance A chalongier	la li	comance Et dist :
CharretteKu	Ja Dex puis ne me doint Joie, que je	la li	randrai.
CharretteKu	Einz que s'ame alast devant Dé Je	le li	eüsse amandé Si richemant con li pleüst
YvainKu	Et cil	le li	dïent : « Par ci, tot droit.
beroul	Vient a Ogrin, il	la li	balle.
beroul	Ha ! roïne, donez	la li	! » Yseut la bele dist au roi :
beroul	Par mié l'uel	la li	fait brandir, Trencha le test et la cervele.
PercevalKu	Mes il dist qu'il la vangera, se Damedex	le li	consant.
PercevalKu	Si richemant apareilliee	la li	a li sires bailliee, et dist :
PercevalKu	et si li comanda s'espee, et cil	la li	garda.
PercevalKu	Et Percevax	le li	otroie, et li hermites li consoille une orison
conttyr	cuida trover Renaut de Sayete qui	la li	rendist. Il nen trova point.
villehardouin	l'impereres le conut bien [...] que il	le li	feroit mult volentiers.
rosel	il fu [...] plains de desdaing et de fierté, si ne	le li	vost ostroier ne por dire ne por proier.
SLambert	et habit de moines, et	la li	dona Nostre Sire Dex si grant
ImMondePr	n'i ait puissance par nature [...], tele comme Diex	la li	donne.
atrper	Et li sire	la li	otrie, Et li dist qu'il soit aseür
rosem1	se ses amis	le li	a doné ou tramis
rosem1	n'en a il pas mains de moleste que cil qui	la li	a requise, tant est d'amor grant la mestrise.
rosem1	qu'il croit que Dex	le li	present quant il lera l'essill present
rosem1	Nature bien	les li	nia.
rosem1	ou se sa robe trop s'enpoudre, souzlevez	la li	de la poudre.
rosem2	devant les voisins qui la vienent [...] et	la li	tolent a grant paine tant qu'il est a la grosse alaine.
rosem2	car trop avroit au queur angoisse quant el	les li	verroit porter, riens ne l'en porroit conforter ;
rosem2	il leur covient leur terres vendre ainz que tout	le li	puissent rendre.
rosem2	ainsinc fere	le li	convient [...] que bien raconter vos savré
rosem2	un blanc laz de fil pendues [...] ; donees	les li	ot uns freres qu'el disoit qu'il estoit ses peres
rosem2	s'il fust voirs, car il le seüst, qui que soit dit	le li	eüst. De soi le poïst il savoir :
rosem2	Faus Semblant ainsinc	le li	preuve ;
rosem2	et viengne trop celeement quant je	le li	feré savoir.
rosem2	Et la vielle es poinz	le li	lance et li veust fere a force prendre
rosem2	por soi mieuz escuser, que mieuz	le li	vient refuser.
rosem2	n'ait ja queur de servir echar, s'il est qui soffrir	le li	veille.
rosem2	Dame, por quoi tant atandez que vos ne	la li	demandez ?
rosem2	maintenant ravir la vosist, se plus fort ne	la li	tosist, et la lessast, s'il li pleüst, quant son voloir fet
rosem2	s'il cuidast que je le vousisse ou que, sanz plus,	le li	soffrisse. Ainsinc Nature nos joutise
rosem2	Mercurius	le li	trancha quant de Juno la revancha
rosem2	je ne bé pas qu'el soient moies, ainz	les li	quit.
rosem2	vos deïssiez certainement que Largece	le li	bailla et qu'el le paint et antailla
rosem2	si con Nature	les li	livre.

Suite du tableau à la page suivante

Tableau 6 – suite du tableau

Référence	Contexte gauche	Pivot	Contexte droit
rosem2	Et cil qui dit	le li	avra, s'il est tex, puis qu'el le savra, qu'il
rosem2	ja plus tost ne la touchera conme el	le li	reprouchera ; mes ce sera tout en apert.
rosem3	mes reson ainsinc	le li	preuve, qui les demontraisons i treuve
rosem3	Cil fist l'antandemant de l'ome, et, an fesant,	le li	donna.
rosem3	Lors escrit cil, et cele dite, puis la seele et	la li	baille, et li prie que tost s'an aille, mes qu'ele soit
rosem3	qui por ce	les li	vost baillier qu'el seüt autex antaillier
rosem3	s'il nouveaus homes ne fesoit, se refere	les li	plesoit, ou ceus feïst resouciter por la terre arriers habiter
rosem3	Puis	les li	roste, et puis ressaie con li siet bien robe de saie
rosem3	n'a poair de la chose oïr ne voair n'il n'est qui dire	le li	puisse, n'el n'a poair que ci vos truisse.
rosem3	Dames, je	la li	habandons, fet Bel Acueill, mout volantiers.
grchron1	il se soustrairoient de la bataille [...] et	le li	rendroient tost pris.
JAntRect	nos	la li	porrons legierement reprendre en tel guize.
JAntRect	se nature te constraint, tu	la li	rendes, mais se le païs te prie, tu ne la li donras pas
JAntRect	tu la li rendes, mais se le païs te prie, tu ne	la li	donras pas.
moree	Mais a la fin lui monstra tant de raisons que li princes	la li	ottroia, disant que il n'estoient ancores bien certains
GMLyon	la dame [...] S'a moy	le li	plaisoit a dire.
regcrim1	lui fu dit que se de ce ne disoit la verité, que l'en	le li	feroit dire, et seroit mis à question
regcrim1	se elle ne le disoit de sa volenté, que par force l'en	le li	feroit dire, et seroit mise à question.
regcrim1	dist [...] que elle	le li	baillast, en prenant de lui la somme de trois frans
regcrim1	de paour qu'il ne les perdesist, ou que l'en	les li	ostast.
regcrim1	fu dit que de ce que dit est elle deist verité, ou l'en	le li	feroit dire, et seroit mise à question.
regcrim1	elle deist verité, ou l'en	la li	feroit dire par sa bouche, à force, et seroit mise à question.
regcrim1	elle deist la verité, ou l'en	la li	feroit dire à force par sa bouche, et seroit mise à question.
regcrim1	il cogneut et deist verité, ou l'en	le li	feroit dire par sa bouche.
regcrim1	faite il deist verité, ou l'en	la li	feroit dire par force et contrainte
regcrim2	et ne scet, elle qui parle, se ledit prisonnier	le li	coppa ou non ;
regcrim2	ou s'il ne les disoit, qu'il seroit mis à question, et	le li	feroit-on dire par force.
regcrim2	il lui voulsist faire vendre icelle, et	la li	faire valoir ce que elle valoit de raison.
regcrim2	et, pour ce,	le li	baillerent.
regcrim2	dit à icelle Perrete comme prins l'avoit et qu'il	le li	rendroit ;
regcrim2	autrement que fait n'avoit, ou autrement il	le li	feroit dire par sa bouche et par voie de question.
QJoyesKa	car l'en lui fait acroire que son pere ou sa mere	les li	oust donnez de leur livree.
QJoyesKa	femme qui est bonne galoise et entent bien raison qui	la li	dit, la quelle croit auxi bien de son mary
QJoyesKa	Elle lui fait porter les enfans jouer ; elle	les li	fait bercier ;
SGenPr5	mais affin que sainte Geneviere ne	les li	demandast, il yssi hors de Paris

TABLEAU 7 – Exceptions à l'ordre de la séquence de pronoms conjoints

Référence	Contexte gauche	Pivot	Contexte droit
aucassin	Jel	te le	di et tu l'entens
thomas	Pur quei n'ai quis la vostre mort, quant	me la	quesistes a tort ?
QJoyesKa	S'il vous plaist, vous	me le	direz.
jouvencel2	Beau filz, il fault que vous	me la	laissiez pour ceste heure
jouvencel2	Et vous prie que vous	me les	dictes encore une foys.
jouvencel2	se le Roy vostre pere m'a fait aucunes promesses et il	me les	tenoit , ce seroit tollir à vostre frere son heritaige
menreims	et furent li enfant envoié [...], et	li les	fist bien garder.

Evaluation

Corpus de test

TABLEAU 8 – Phrases de test

Phénomènes syntaxiques	**Phrases de test**	**Sources**	**Modifications**	**Tokens originaux**
prédéterminant	**peu de** paroles rend[ent] le auditeur mains ententif	daudin	conjugaison	rend
idem + dét.	Vray est que ledict roy Henry valloit peu de sa personne	commyn2		
partitif de	Ainçois doit li derrains procureres conter **de** tout.	beauma1		
”	et il orent bien **de** quoi.	villehard.2		
relative partitive	il n’ont **de quoi** avoir nul livre	ImMdePr		
dét. numéral	Tut entur lui vint **milie** Sarrazins	roland		
adj. numéral	ansemble an vunt li [**trois**] pedre parler	alexis	voc.	dui
intervalle	**De** deux **a** trois cens chevaulx	jehpar		
dét + nc	**Le pape** et **le duc** [...] escripvoient	commyn8	ellipse	
nc sans dét.	In **figure** [...] volat a ciel	eulaliBfm	ellipse	
arg. du nc	[peur] **de perdre**	thebes1	voc.	poour
SN quantité	Dieu nous en donra **cent foiz plus que** nous n’en donrons pour lui	Berin2		
nc position	Li paien controverent Les nuns que as jurz dunerent **Li premier**	comput		
”	Comme je vous ay dit, le roy estoit arrivé **le premier**.	commyn4		
nc temps	**le matin** [ils] veneient à lui	oxfps	ajout	[ils]
nc + np	Li **reis** Marsilie esteit en Sarraguce	roland		
vocatif	**Rollant**, veez en alques	roland		
”	**Fille**, fet il, avant venez !	mf		
audience	**Ma tresdouce Dame**, il me semble	hlanc		
adj sur np	La **bele** Yseut	beroul	morphologie	belle
apposition	Romulus, **li reis**, A sun pere dunat	comput		
”	En Biture, **une cité**, Avint un fait mult renumé	adgar		
”	Guenes **li quens**	roland		
prep + prep	il fu partiz **de chiés** .i. vavasor	qgraal		
”	**dès avant** Pasques	baye2		
PP antéposé	**d’Espaigne** le regnet	roland		
”	**de la curt** le rei aloent	mf		
PP sur intj	Fy [...] **de** voz raisons !	cdo rond.	ellipse	
circonstancielle	**Quant** vit sun regne, durement s’en redutet	alexis		
”	**por ce que** ge la desiroie, son seignor ocirre voloie	Erec		
”	Tel cuntenance demerra **Endementers qu’** ele dormira	Lapidaire		
advneg modifieur	N’est hom **pas** sains	RutebZ		
adv sur adv	Chiés Gautier fu la joie **moult tres** grans	amiamil		
adv sur pro	car sanz ce ne puet **il aussi** vivre ne durer longuement	Berin1		
adv sur nb	enter mirra et aloén **quasi cent** liuras a doned	Passion		
adv sur adj	son **tres chier** neveu	ErecKu		
adv sur att.	Li amiralz est **mult de grant vertut**	roland		
”	Margariz est **mult vaillant chevalers**	roland		
adv sur csu	Sa vie estoit la plus belle du monde, **ainsi qu’**il se povoit veoir	commyn8		
adv sur SN	il ont prins (...) **environ XX escuelles d’estain**	regcrim1	ellipse	
adv sur mod. tps	**Uncore hui** ferrai de l’espee	guill1		
coord. *ni*	Il ne la list **ne** il dedenz **ne** guardet	alexis		
coord. de ph.	Aprés parlat ses filz envers Marsilies, **E** dist al rei	roland		
coord. de ptcp	Jusqu’al nasel li ad **fait e fendut**	roland		
coord. de v.	Ele volt **veïr et avoir** vitam eternam	CommPsia1a		
coord. de ph.	coment est anommé ceste ville, et ou demurt Guilliam Rorane ?	maniere1396		
coord. rel.	Ce fu cil **qui** mius se viestoit **et qui** se savoit mius avoir	eracle		
coord. csu	son pere bien l’asseüre, si conme il dit **et conme** il jure	thebes2		
coord. mod. tps	l’en plaidera les **lundi et mardi** des ordinaires	baye1		
coo + mod	ci comensent les aornemenz de paroles et **premierement** la repeticion	JAntRect		
ou non	ont il droitement fait ou non ?	JAntInv		
”	Et savra s’ele est vive ou non.	CligesKu		
clr + objet	Se une femme **se** (mes)passe **le pied**	quenouilles	morphologie	
clseq post	donrrai **le lui**	thebes2		
cld + adj	Molt **li** est **grief** a departir	roland		

Suite du tableau à la page suivante

Tableau 8 – suite du tableau

Phén. synt.	Phrases de test	Sources	Modifications	Tokens originaux
cld + obj	David **lui** coppa **la teste**	anglure		
cld + passif	la proprietes de la chose **li** sera acquise	beauma1		
SOV	Li hermite Tristran connut	beroul		
SVO	Nos avum dreit	roland		
VSO	dunc ad ele colur	lapidaire		
VOS	perdi alcun Fenénne	qlr		
OVS	Tutes cestes culurs ad ele ensemble.	lapidaire		
OSV	l'imagine Deus fist	alexis		
sujet passif	**la guige** en fu batue	guill1		
aux + obj. + ptcp	elle n'a riens fait	regcrim2		
question	Reis magnes, que fais tu ?	roland		
”	Cest enfant est elle vostre fille ?	SGenPr5		
”	De quoi fustes vous donc batue ?	bestam	voc.	dont
il impers.	**il** i a peril	CligesKu		
non accord ccomp	il [estoient] une sustance	SBernAn	conjugaison	sunt
abs. obj.	li chevaliers vont laver	thebes1		
relative	elle l'a de la personne **de laquelle** elle est	theologie		
”	C'est la lignee **qui** debouta et occist son souverain seigneur	quadrilogue		
”	Maleureux est celuy **qui** maleur quiert	ressource		
”	la science **qui** enflet ne vient mie del baisuel	SBernCant		
”	Vunt les ferir la **o** ils les encuntrent	roland		
relative sujet	**qui pert** paye	ressource	conjugaison	perdit
topic passif	En ceste fame sont trovez Cinq miracles	MirNDChartr		
modal + vinf + objet	Qui veut oir une aventure	beroul		
modal + objet + vinf	chascun veut le don aquerre	thebes1		
modal + nég.	Aussi ces gens d'armes ne font que endommaiger vostre royaume	jouvencel2		
modifieur de ph.	**Guaires** ne demura que li freres chaï	becket		
génitif post	l'arçun **li bers**	guill1		
génitif post	Li niés **Marsilie**	roland		
génitif ante	la **Deu** merci	yvain		
superlatif	**les plus** ancïens **que** vus porrez trover	becket		
comparatif	li boton [...] val[oit] **plus que** troi chastel	eneas1	ellipse	valoient
”	ele afereit **plus** a lui **que** a nul	conttyr		
”	Et si avra **plus** de deli **que** nus	ImMondePr	morphologie	delit
”	mais il (...) l'avoit **tellement** promiz **que** nullement il n'eust osé	Berin2	ellipse	
”	et orent **tellement** besoingnié **que** tous les hommes Giron furent mors	melusine		

Evaluations sur corpus

class. as =>	punct	root	det	advmod	nsubj	obl	obj	case	mark	cc:nc	conj	aux	cc	amod	advcl	iobj	nmod	acl : relcl	xcomp	case : det	cop	acl	ccomp	expl	obl : advmod	voca- tive	flat	aux : pass	num- mod	appos	para- taxis	csubj	dislo- cated	dis- course	nsubj : obj	fixed	dep	nsubj : pass	Total
punct	**57.8**	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	23004
root	-	**75.6**	0.1	0.8	2.0	0.7	1.3	1.0	0.1	-	2.2	0.8	-	2.5	1.3	-	0.3	0.1	1.8	-	0.4	1.3	0.7	-	-	0.3	0.2	0.1	-	1.9	2.6	0.1	0.1	0.2	-	0.1	1.3	0.1	15864
det	-	0.2	**85.2**	1.9	0.7	1.1	1.1	0.1	0.1	-	0.6	-	-	-	-	0.2	0.6	0.0	0.2	-	-	-	0.1	-	-	0.1	0.1	-	-	0.3	-	-	0.1	-	-	-	0.4	-	13452
advmod	-	2.1	1.0	**62.4**	0.7	0.4	0.5	1.9	0.4	0.2	0.5	0.1	1.5	0.3	0.1	-	0.4	0.1	0.2	0.1	-	0.2	0.1	-	-	0.3	0.3	-	-	0.8	0.1	-	0.1	0.1	-	0.2	0.6	0.1	13618
nsubj	-	1.7	0.8	0.2	**66.5**	0.4	1.5	0.2	0.2	-	1.3	-	0.1	0.4	0.2	0.1	0.5	0.1	0.2	-	-	-	-	0.1	-	0.5	0.5	-	-	2.7	0.2	-	0.3	-	-	0.2	0.6	0.1	13232
obl	-	1.6	0.8	0.6	1.2	**43.7**	1.7	0.4	0.3	-	0.9	-	0.3	0.6	0.1	0.1	9.5	0.1	0.3	0.3	-	0.2	0.1	-	0.9	0.1	0.3	-	-	3.7	0.1	-	0.1	-	-	0.2	0.8	-	11380
obj	-	1.8	1.8	0.7	1.6	1.1	**51.0**	0.2	0.4	0.1	1.0	0.1	0.7	0.7	0.1	0.3	0.8	0.2	0.3	-	-	0.1	-	-	-	0.5	0.3	-	0.1	4.2	0.1	-	0.1	-	-	0.4	0.3	0.1	11232
case	-	0.6	-	0.3	0.7	0.2	0.5	**84.8**	0.4	-	0.2	0.2	-	-	0.3	-	0.1	0.1	0.1	-	-	-	0.2	-	0.1	0.1	0.2	-	-	0.6	0.1	-	-	0.2	-	0.1	0.5	-	10644
mark	-	1.4	2.5	5.0	0.3	0.4	2.1	4.0	**34.4**	0.5	0.3	-	4.1	0.1	0.2	-	0.5	-	0.2	-	-	-	0.4	-	-	0.2	0.9	-	-	1.5	0.5	-	-	0.2	-	0.3	0.3	-	5278
cc:nc	-	0.9	0.7	0.8	0.1	-	0.2	0.7	0.9	**71.3**	-	-	0.4	0.1	-	-	-	-	0.1	0.1	-	0.1	-	-	-	0.2	-	-	-	0.1	-	-	-	0.3	-	-	0.2	-	4799
conj	-	1.7	0.4	0.9	1.7	2.0	2.7	0.2	0.1	-	**48.7**	0.1	-	3.5	0.5	0.1	2.0	0.2	0.6	-	-	1.0	0.4	-	-	0.6	1.1	-	0.2	5.3	0.6	0.1	0.1	0.1	-	0.4	0.6	0.2	3809
aux	-	6.6	-	0.1	0.1	0.2	0.5	1.7	-	-	1.3	**63.7**	-	0.8	1.1	-	0.2	1.3	0.5	-	0.6	0.4	1.0	-	-	-	-	0.1	-	1.4	0.4	0.1	-	0.2	-	0.2	0.4	-	3720
cc	-	0.8	-	0.2	0.3	0.1	0.5	0.6	0.1	1.5	0.1	-	**73.9**	-	-	-	0.1	-	-	0.1	-	0.1	-	-	-	0.3	0.5	-	-	1.1	-	-	0.1	1.2	-	0.1	0.1	-	3441
amod	-	2.6	0.2	1.6	2.6	2.3	1.8	-	-	-	1.1	-	-	**69.3**	0.3	-	0.7	0.1	1.4	-	-	0.2	0.2	-	-	0.5	0.1	-	-	1.2	0.3	-	0.1	0.1	-	0.3	0.7	0.1	3476
advcl	-	21.8	-	0.3	0.9	0.6	1.5	1.1	-	-	4.6	1.0	-	2.5	**21.0**	-	0.7	2.9	2.9	-	0.3	4.6	4.0	-	-	0.1	0.1	0.1	0.1	2.7	1.6	0.2	-	0.1	-	0.2	1.5	0.1	2917
iobj	-	0.7	2.9	0.1	0.6	1.2	1.6	-	0.3	-	0.9	0.1	-	0.8	-	**58.4**	1.4	0.2	0.2	-	-	-	-	-	-	0.1	0.2	-	-	0.8	-	-	-	-	-	0.2	0.1	-	2833
nmod	-	1.4	0.8	0.5	2.8	14.4	1.9	0.9	0.1	-	1.3	0.7	0.3	0.7	0.5	0.3	**40.5**	-	0.5	-	-	0.1	0.2	-	8.0	0.1	0.7	-	0.1	3.5	0.4	-	0.2	-	-	0.2	0.2	0.2	2622
acl:relcl	-	9.6	-	0.3	1.1	0.3	1.7	1.2	-	-	10.5	0.7	-	2.4	2.4	-	0.2	**32.8**	4.6	-	0.5	1.8	5.4	-	-	0.1	0.1	-	-	1.7	2.4	0.3	0.2	0.1	-	0.1	1.4	-	2054
xcomp	-	13.6	0.1	0.2	1.1	0.7	1.0	0.1	-	-	2.1	0.1	-	3.5	1.6	-	0.5	0.5	**28.2**	-	-	4.8	0.2	-	-	0.4	1.0	-	-	4.1	0.3	1.2	-	0.1	-	0.1	1.1	0.2	1777
case:det	-	0.7	0.1	0.4	3.8	0.3	0.7	1.7	0.2	-	0.8	0.6	0.1	-	0.2	-	0.4	0.1	2.0	**16.5**	-	-	0.1	0.3	-	0.3	0.1	-	-	2.6	0.2	-	0.2	-	-	0.2	-	0.1	1744
cop	-	17.1	0.1	-	0.8	0.3	0.9	0.2	-	-	2.2	0.4	-	-	2.5	-	0.2	1.4	0.7	-	**52.2**	0.1	2.4	-	-	0.5	0.1	0.2	-	2.5	1.1	-	0.5	-	-	0.1	1.0	-	1611
acl	-	13.9	0.1	0.5	1.0	1.0	1.3	0.6	0.6	-	10.8	0.4	-	7.4	8.2	-	0.7	1.8	6.1	-	0.2	**6.7**	6.1	-	-	0.3	0.5	-	-	2.6	2.5	0.8	0.1	0.1	-	0.1	1.2	-	1360
ccomp	-	6.9	0.1	0.3	0.8	0.3	1.8	0.3	0.2	-	3.1	0.4	0.2	2.3	1.9	-	0.3	2.1	2.7	-	0.3	0.6	**41.2**	0.1	-	0.2	0.1	-	-	1.5	0.7	0.3	-	0.2	-	-	0.6	-	1199
expl	-	1.2	3.6	3.0	3.1	0.1	2.5	-	0.6	-	0.2	-	0.4	-	0.1	0.2	0.1	0.1	0.1	-	-	0.1	0.2	**0.2**	-	0.1	-	-	-	0.1	0.2	-	-	-	-	0.1	0.3	-	1209
obl:advmod	-	0.4	-	0.2	0.8	0.4	0.4	1.9	0.4	-	1.9	-	-	0.1	-	-	0.6	-	-	-	-	-	-	0.1	**70.3**	0.1	0.2	-	-	1.3	-	-	-	-	-	0.2	-	-	1002
vocative	-	9.8	-	-	2.6	0.2	1.7	0.2	-	-	-	-	-	1.0	-	-	-	-	-	-	-	0.2	-	-	-	**35.8**	1.3	-	-	8.2	0.9	-	2.4	-	-	0.4	9.9	0.2	978
flat	-	0.8	0.9	-	1.4	3.6	0.6	6.3	-	-	0.5	-	0.1	-	-	-	0.8	-	-	-	-	-	0.2	-	-	0.4	**61.5**	-	-	1.1	0.1	-	-	-	-	0.1	0.2	-	853
aux:pass	-	8.4	-	-	-	0.4	0.2	-	-	-	1.4	1.9	0.2	0.2	1.8	-	0.2	2.2	0.5	-	3.3	0.2	1.2	-	-	0.2	0.1	**54.5**	-	1.2	0.5	-	0.1	-	-	0.1	0.5	-	837
nummod	-	4.5	0.2	0.5	9.0	5.7	4.2	-	-	0.2	2.0	0.5	-	0.3	-	-	1.5	-	1.3	-	-	0.8	0.7	-	-	0.8	0.3	-	**44.1**	2.8	0.2	-	0.3	-	-	0.2	0.8	0.8	599
appos	-	0.6	0.6	0.8	1.2	-	4.6	-	-	-	0.6	-	-	2.5	-	0.2	0.6	-	0.4	-	-	-	0.2	-	-	5.8	0.4	-	-	**65.8**	0.8	-	0.2	-	-	0.2	0.2	0.6	482
parataxis	-	15.9	-	-	0.2	-	0.7	0.5	-	-	-	-	0.2	3.6	6.3	-	-	-	0.5	-	0.2	2.4	-	-	-	1.2	-	-	-	1.2	**24.3**	-	-	-	-	-	2.9	-	415
csubj	-	17.9	-	-	2.0	-	1.5	1.5	-	-	3.5	1.0	-	3.0	2.5	-	0.5	12.4	3.5	-	-	3.0	8.5	-	-	-	-	-	-	3.5	1.5	**7.5**	1.0	-	-	-	0.5	-	201
dislocated	-	9.3	1.0	-	5.2	5.7	4.6	-	-	-	2.6	-	-	3.1	-	-	2.1	2.1	1.0	-	0.5	1.0	1.5	0.5	-	2.1	-	-	0.5	11.3	1.5	1.0	**9.8**	-	-	-	1.5	1.0	194
discourse	-	22.1	-	2.9	0.7	-	0.7	1.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.7	2.1	-	-	-	-	-	0.7	**1.4**	-	-	47.1	-	140
nsubj:obj	-	1.4	-	-	1.4	-	1.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.4	-	-	1.4	-	**82.2**	-	-	-	73
obj:advneg	-	-	-	3.6	-	-	7.1	-	-	-	-	-	3.6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	28
fixed	-	-	-	-	5.9	-	5.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5.9	-	-	-	47.1	-	-	-	11.8	-	-	-	-	17
obj:advmod	-	-	-	-	-	11.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	9
mark:advmod	-	16.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	33.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	6
orphan	-	-	-	-	-	-	-	-	-	-	20.0	-	-	-	20.0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	20.0	-	-	5
advmod:obl	-	-	-	33.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3
compound	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	50.0	-	-	-	-	-	2

	NOUN	VERB	PUNCT	PRON	ADP	DET	ADV	CCONJ	ADJ	AUX	SCONJ	PROPN	NUM	INTJ	Total
NOUN	**89,1**	5	-	-	-	-	-	-	2,6	-	-	-	-	-	124872
VERB	5,2	**80,2**	-	-	-	-	-	-	3	9,2	-	-	-	-	131697
PUNCT			**99,7**	-	-	-	-	-	-	-	-	-	-	-	114964
PRON	3,9	1,3	-	**81,3**	-	3,1	3,7	1,1	1,7	-	1,2	-	-	-	99407
ADP	1,4	1,7	-	-	**90,7**		1,7	-	-	-	-	1,2	-	-	81258
DET	1,3	-	-	6,1	1	**83,2**	3,7	-	2,4	-	-	-	-	-	78179
ADV	5,3	2,4	-	3,4	3,7	1,2	**76,2**	3,4	1,2	-	1,8	-	-	-	82078
CCONJ	2,1	-	-	-	3		2,2	**86,5**		1,4	1,1	-	-	1,1	49015
ADJ	14,7	6,7	-	-	-	1	2,4		**71,6**	-	-	1	-	-	31054
AUX	5,5	21,6	-	-	1,8	-	-	-	-	**68,2**	-	-	-	-	25377
SCONJ	1,3	-	-	20,1	3,5	2,3	11	12,1	-	1,1	**45,7**	1,5	-	-	22237
PROPN	25,5	2	-	-	-	-	-	-	1,2	-	-	**68,7**	-	-	23717
NUM	17,4	7,4	10,6	-	-	1,8	-	-	9,6	3,2	-	2,6	**44,2**	-	2912
INTJ	-	8,4	-	-	2,1	-	2,9	-	-	-	-	7	-	**74,6**	654

TABLEAU 9 – Confusions par partie du discours (au delà de 1%)

Annotation et analyses syntaxiques de corpus hétérogènes : le cas du français médiéval

Le français médiéval couvre les états de langue d'ancien français (9e-13e s.) et de moyen français (14e-15e s.). Nous disposons de données annotées pour ces états de langue, dont un corpus arboré d'ancien français (Stein et Prévost 2013). Il est cependant difficile d'obtenir plus de données annotées syntaxiquement, car les spécialistes sont peu nombreux et qu'il n'existe pas encore d'outil dédié pour l'ensemble de la période. Développer ce genre d'outil permet d'obtenir des annotations plus facilement et d'en contrôler la qualité. Cependant, ce n'est pas une tâche simple parce que les différents états de langue sont soumis à la variation, due à plusieurs facteurs, notamment l'absence de norme graphique, la variation dialectale, la souplesse de l'ordre des mots, l'évolution de la morphologie et de la syntaxe (sur sept siècles), qui fait passer le français d'une langue SOV à une langue SVO. La nature des écrits se diversifie aussi à mesure que la littérature évolue et que le latin est délaissé au bénéfice du français comme langue administrative et juridique. Les données à analyser sont donc hétérogènes, ce qui rend difficile le traitement automatique.

Pour obtenir un parseur du français médiéval, nous proposons d'adapter la métagrammaire du français contemporain FRMG (Villemonte de la Clergerie 2005). Bien que les différents états de langue présentent des différences manifestes, les points communs sont suffisants pour rendre possible la modification d'un système existant pour obtenir un outil dédié. Les changements concernent essentiellement l'ordre des mots (constituants majeurs, modifieurs du nom, position des pronoms conjoints). Pour utiliser cet outil sur corpus, il est nécessaire d'enrichir le lexique d'ancien français (Sagot 2019), d'une part pour obtenir une couverture lexicale satisfaisante sur les textes, et, d'autre part, pour y intégrer des informations syntaxiques et sémantiques nécessaires à l'analyse syntaxique.

Mots clés : annotation syntaxique, parsing, grammaire d'arbres adjoints, métagrammaire, français médiéval, ancien français, corpus hétérogène

Syntactic Analysis and Parsing of Heterogeneous Corpora : The Case of Medieval French

Medieval French is an umbrella term for Old French (9th-13h c.) and Middle French (14th-15th c.). We have annotated data for these stages, including a dependency treebank of Old French (Stein et Prévost 2013). However, obtaining more treebanks is difficult, because there are few experts of Medieval French and we do not have yet a dedicated parser for the whole period of Medieval French. A dedicated tool would make it easier to annotate new corpora and it would enable to control the quality of the annotation. Nevertheless, it is not a trivial task, because the states of language are subjected to variation. It comes from several sources, including the absence of standard spelling, dialects, flexible word order, evolution of morphology and syntax over seven centuries, with seminal phenomena like the transition from a SOV language to a SVO language. Text genres do also evolve as the number of literature writings rises and Latin is replaced by French for official texts such as treaties, contracts and chronicles. The data available for Medieval French are therefore heterogeneous, which makes it difficult to annotate them automatically.

We chose to adapt the *French Metagrammar* (FRMG, Villemonte de la Clergerie (2005)) in order to develop a parser for Medieval French. Even if the differences between Medieval French and Contemporary French are striking, there are enough similarities to obtain a satisfactory parser. The main changes ensure the word order is properly analysed (ex. major constituents, noun modifiers, position of clitics). In order to annotate a new corpus, adapting the lexicon OFrLex (Sagot 2019) is mandatory : new entries as well as new syntactic and semantic information were added.

Keywords : syntactic annotation, parsing, Tree-Adjoining Grammar, metagrammar, Medieval French, Old French, heterogeneous corpus

Ecole doctorale 622 – Langage et langues
Université Sorbonne Nouvelle, Maison de la recherche
4, rue des Irlandais, 75005 Paris