Adaptation d'une métagrammaire du français contemporain au français médiéval

Mathilde Regnault[1, 2]
(1) Lattice
(2) Inria
mathilde.regnault@sorbonne-nouvelle.fr

RÉSUMÉ

Le français médiéval se caractérise par une importante variabilité langagière. Nous cherchons à étendre un corpus d'ancien français annoté en syntaxe de dépendance avec de nouveaux textes de cette période et de moyen français. Pour cela, nous voulons adapter des outils existants et non entraîner un *parser* avec des données annotées. Dans cet article, nous présentons un état de l'art pour ce projet et notre démarche : adapter FRMG (*French Metagrammar*) à des états de langue antérieurs.

ABSTRACT

Adapting an existing metagrammar for Contemporary French to Medieval French

Medieval French is characterized by strong language variation. Our purpose is to extend a corpus of Old French annotated with dependency syntax with new texts of this period and add texts of Middle French. In order to achieve this, we want to adapt existing tools instead of training a parser with annotated data. In this article, we present a state of the art for this project and our solution : adapting the French Metagrammar (FRMG) to former states of language.

MOTS-CLÉS : métagrammaire, analyse syntaxique automatique, français médiéval, données hétérogènes.

KEYWORDS: metagrammar, syntactical parsing, Medieval French, heterogeneous data.

1 Introduction

Dans le cadre du projet ANR Profiterole, nous développons des outils d'annotation syntaxique pour traiter le français médiéval, caractérisé par une importante variabilité langagière. Cela représente une difficulté pour le *parsing* (Guibon *et al.*, 2014, 2015). Pour gérer les différentes variations (diachroniques, dialectales...), nous avons choisi de ne pas nous servir d'un *parser* statistique, mais de développer une métagrammaire du français médiéval, que nous n'avons pas encore évaluée. Un système symbolique permet de mettre à profit les grammaires existantes (Hasenohr & Raynaud de Lage, 1993; Buridant, 2000; Marchello-Nizia, 1979). Une métagrammaire est une description modulaire et hiérarchisée des phénomènes syntaxiques, divisée en classes, qui nous permet de mettre en place une analyse fine de la langue et d'intégrer des contraintes activées ou non selon le contexte ou les métadonnées. Ce formalisme présente aussi une économie dans l'écriture de la grammaire d'une langue car ses classes permettent de factoriser les descriptions de phénomènes. Nous pensons qu'il est possible d'adapter une métagrammaire du français contemporain, FRMG (Villemonte de la Clergerie, 2005), à des états de langue antérieurs et hétérogènes. Nous pourrons ainsi étendre

le SRCMF (Prévost & Stein, 2013), un corpus d'ancien français (9e-13e s.) annoté en syntaxe de dépendance, pour en faire une ressource conséquente (env. un million de mots), incluant des textes en moyen français (14e-15e s.).

Dans cet article, nous présenterons d'abord les objectifs de nos travaux et les défis qu'ils soulèvent. Puis nous exposerons les problématiques liées à l'adaptation d'un outil du français contemporain au français médiéval, et, dans une dernière partie, nos solutions et travaux en cours.

2 Objectifs et défis de l'analyse syntaxique automatique du français médiéval

2.1 Objectifs du projet

Nous souhaitons multiplier par quatre la taille du SRCMF, un corpus arboré de 250 000 mots, en y ajoutant des textes annotés automatiquement. L'annotation manuelle est trop coûteuse pour envisager une nouvelle campagne de cette ampleur. Les nouveaux textes sont en partie choisis pour rendre la représentation de la langue médiévale plus complète. Six nouveaux dialectes feront ainsi leur apparition (pour un nouveau total de douze), comme le bourguignon, l'orléanais et le wallon. Les domaines seront mieux représentés, en particulier les textes historiques, dont on ne trouve qu'un exemplaire dans le SRCMF. Le changement le plus remarquable est l'ajout du moyen français à ce corpus, qui présente des différences notables avec l'ancien français.

2.2 Travaux antérieurs

Dans leurs premiers travaux sur le SRCMF, Guibon *et al.* (2014) ont fait des expériences d'étiquetage morpho-syntaxique et syntaxique avec des CRF et un *parser* statistique. Cependant, les résultats sont moins bons que pour une langue contemporaine en partie à cause de l'hétérogénéité des données, due notamment aux divers dialectes et aux différentes époques des textes. Ces techniques reposent essentiellement sur les données d'apprentissage, idéalement proches des ensembles de test. L'ancien français se caractérise par une grande variabilité, excluant une telle proximité. Nous voulons aussi ajouter des textes de moyen français, plus éloignés encore des textes du SRCMF. Il semble donc plus prometteur d'utiliser un système symbolique ou hybride pour analyser automatiquement de nouveaux textes.

Pour annoter automatiquement des textes de portugais médiéval (12e-13e s.), Rocio *et al.* (2003) ont quant à eux fait le choix d'adapter non seulement leur *parser*, mais aussi leur grammaire du portugais contemporain. Ils ont fait l'hypothèse qu'il était possible d'annoter une nouvelle langue avec des outils et ressources robustes d'une langue similaire. Nous avons choisi de procéder de manière semblable en adaptant un système basé sur une grammaire (voir section 4). Cependant, nous souhaitons couvrir une période plus vaste (9e-15e s.), dont les états de langue sont à la fois hétérogènes et éloignés du français contemporain.

3 Problématiques liées au français médiéval et différences avec le français contemporain

3.1 Syntaxe du français médiéval

Nous ne présenterons ici que quelques particularités du français médiéval qui ont un impact sur l'identification des constituants principaux. Leur ordre est libre, mais contrairement au latin, la déclinaison ne permet pas de déterminer la fonction syntaxique des syntagmes nominaux car elle est assez peu respectée et devient inefficace. D'autres critères permettent d'analyser les phrases.

ex. ordre SOV : "Tierce fiede Deu Samuel apela" (*Quatre Livres des rois*), trad. *Dieu appela Samuel trois fois.*

Tierce	fiede	Deu	Samuel	apela
Trois	fois	Dieu	Samuel	appela

"Deu" est ici sujet. L'analyse OSV (avec Samuel comme sujet) est exclue du fait que, jusqu'au 13e siècle, l'ordre OSV n'est possible qu'avec un sujet pronominal (Schøsler, 1984). L'analyse grammaticale est confirmée par le texte latin dont *Quatre Livres des rois* est la traduction :
"Et adjecit Dominus, et vocavit adhuc Samuelem tertio" [1], où "Samuelem" est un accusatif, objet de "vocavit".
Dans d'autres cas, aucune "règle" ne permet de lever l'ambiguïté :
ex. "Dolant et pansif *Lancelot* / **Vit** *la dame de la meison*" (*Le Chevalier de la charrete* de Chrétien de Troyes, tel que cité par Buridant (2000)), où *Lancelot* et *la dame de la meison* sont tous deux candidats à la fonction de sujet. C'est la suite du texte qui permet d'analyser cette phrase :
"Sel **mist** a consoil **a reison** : / Sire, por Deu et por vostre ame...", trad. de Buridant (2000) : *La dame du manoir le vit [vit Lancelot] triste et troublé. Aussi lui parla-t-elle en secret de la sorte : Sire, pour Dieu et sur votre âme...*
Ce n'est que le tour de parole suivant qui indique que le sujet des verbes "vit" et "mist a reison" est "la dame de la meison".

Il est habituel qu'un verbe transitif n'ait pas d'objet en français contemporain, mais très rare que le sujet fasse défaut. Le français médiéval est lui considéré comme une langue à sujet partiellement nul. La réalisation du sujet n'est donc pas prioritaire. L'analyse d'une phrase n'est pour autant pas nécessairement ambiguë :
ex. "Et quant il furent revenu si **fisent** savoir as barons qu'il avoient fait" (*Conquête de Constantinople* de Robert de Clari, tel que cité par T. Rainsford *et al.* (2012)), trad. *Et quand ils furent revenus, **ils firent** savoir au baron qu'ils l'avaient fait.*

Et quant il furent revenu	si	fisent	savoir	as	barons	qu'	il	avoient	fait
Et quand ils furent revenus		*ils* firent	savoir	au	baron	qu'	ils	*l'*avaient	fait

Dans cet exemple, aucun syntagme nominal ou pronom ne peut être sélectionné pour sujet de "fisent", car "barons" est précédé de "as", ce qui en fait l'objet second du verbe.

De nombreuses différences existent entre français contemporain et français médiéval. Par exemple, la graphie des mots n'était pas fixe et la variation dialectale était plus forte. Cependant, nous souhaitons

1. *Les Quatre livres des rois, traduits en français du XIIIe siècle, suivis d'un fragment de moralités sur Job et d'un choix de sermons de Saint Bernard*, de Lincy, L.R. (1841). Collection de documents inédits sur l'histoire de France, disponible à cette adresse : `https://books.google.fr/books?id=uW49AQAAMAAJ`.

tirer parti du lien entre les différents états de langue du français médiéval d'une part et le français contemporain d'autre part pour développer un *parser*.

3.2 La langue comme *continuum*

Les changements qui se sont produits du 9e au 15e siècle préfigurent la langue moderne. L'ordre des constituants principaux en est un exemple remarquable. Majoritaire en latin tardif, l'ordre SOV ne l'est plus dès le 12e siècle. C'est SVO, déjà présent dans les premiers textes d'ancien français, qui devient alors le plus fréquent. Cela a aussi un impact sur la place du complément de nom, qui a de plus en plus tendance à apparaître à droite de leur gouverneur syntaxique. Prenons l'exemple du complément déterminatif, tel que décrit par Buridant (2000). L'ancien français se passe parfois de préposition (ex. *a, de*) pour attacher le complément au nom qui le régit. Habituellement on observe cet ordre : "N1 (déterminé) + N2 (déterminatif)", mais on trouve parfois l'inverse, et plus particulièrement avec les mots *Dieu, Damedieu* ("Dieu"), *Dé hé* (malédiction), *Jesus*, ou *autrui*.
ex. "Seignors fait el por ***Deu*** merci Saintes reliques voi ici" (*Tristan* de Béroul, dans le SRCMF), trad. *Seigneurs, dit-elle, par la grâce* ***de Dieu****, je vois ici les saintes reliques.* Cependant, même avec ces mots, cet ordre n'est pas obligatoire. L'analyse N1 + N2 reste autorisée, et l'ambiguïté introduite ne peut être résolue que dans un second temps.

C'est également pendant cette période que les déterminants, apparus en latin tardif, commencent à se répandre. Le français contemporain a largement hérité ses structures du français médiéval, même si certaines ont évolué de façon conséquente. Ce sont surtout les fréquences d'apparition des phénomènes qui ont changé.

3.3 Problématiques liées à l'adaptation

Le français médiéval et le français contemporain semblent suffisamment similaires pour que nous puissions garder une certaine continuité entre le *parser* pour le français contemporain et celui pour le français médiéval. Adapter un système originellement développé pour un état de langue à des états bien antérieurs demande cependant de mettre au point une méthodologie pour le traitement des données hétérogènes.

Nous nous sommes demandé si une seule grammaire pouvait suffire à couvrir une période de sept siècles et si le système résultant ne serait pas inutilement complexe. Néanmoins, nous ne pouvons pas développer plusieurs grammaires couvrant chacune une période ou un dialecte. La langue ne saurait en effet être ainsi "découpée" car elle forme un *continuum*. De plus, les phénomènes syntaxiques ne connaissent pas nécessairement une évolution linéaire. Par exemple, l'ordre OSV (avec objet nominal) est très rare en ancien français, mais il connaît une hausse ponctuelle aux 14e-15e siècles. Il n'est donc pas possible de définir des étapes strictes que l'on décrirait dans différentes grammaires. La fréquence d'apparition de phénomènes est différente selon la date du texte, son dialecte, son domaine, son auteur et s'il est écrit en vers ou en prose. De nouveaux textes pourraient donc déjouer les prévisions, ce qui mènerait leur analyse à l'échec. Nous devons donc trouver un moyen de représenter à la fois la continuité et les variations possibles de la syntaxe à travers cette période.

Pour cela, nous avons envisagé plusieurs méthodes. La *Grammar Matrix* (Bender *et al.*, 2002) a inspiré une approche par librairies. Les descriptions de phénomènes syntaxiques seraient séparées, et chaque état de langue pourrait sélectionner les ressources dont il a besoin. Cependant, nous n'avons

pas l'ambition de développer une librairie universelle de grammaires, et nous risquerions de nous heurter aux problèmes définis précédemment (définition impossible de la syntaxe de chaque état de langue et incertitude face aux nouveaux textes). Nous avons choisi un formalisme plus modulaire : les métagrammaires.

4 Adaptation de FRMG

Les différences notables entre l'ancien et le moyen français d'une part et le français moderne d'autre part pouvent suggérer de développer un nouveau système pour traiter ces états de langue. Plutôt que développer une nouvelle métagrammaire spécifiquement conçue pour l'ancien et le moyen français, nous avons choisi d'adapter FRMG, une métagrammaire du français contemporain, qui représente une base intéressante pour décrire des états de langue anciens et hétérogènes.

4.1 Une métagrammaire pour le français médiéval

Les métagrammaires peuvent être écrites pour générer plusieurs types de grammaires. Elles ne sont pas intrinsèquement liées au formalisme des grammaires d'arbres adjoints, mais ce sont ces dernières qui nous intéressent et que nous allons présenter car FRMG en génère une.

4.1.1 Les grammaires d'arbres adjoints

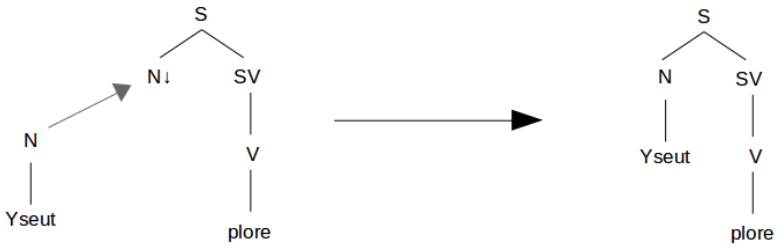
Plusieurs métagrammaires (Candito, 1999) sont basées sur les grammaires d'arbres adjoints (ou *Tree Adjoining Grammars*, désormais TAG) (Joshi *et al.*, 1975), un formalisme mathématique adapté aux langages naturels. Il est décrit formellement par le quintuple <N, T, I, A, S> où :

- N est un ensemble de non terminaux correspondant à des catégories syntagmatiques
- T est un ensemble de terminaux correspondant à des éléments lexicaux
- I est un ensemble d'arbres initiaux
- A est un ensemble d'arbres auxiliaires, qui ont chacun un noeud feuille appelé "noeud pied", symbolisé par une étoile (*) et étiqueté par un non terminal de même catégorie que le noeud racine de cet arbre
- S $\in$ N est l'axiome

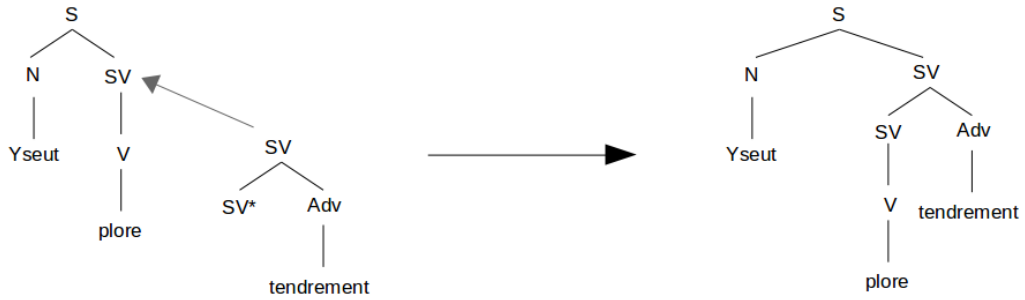
Une grammaire d'arbres adjoints est un ensemble d'arbres élémentaires. Les noeuds feuilles de ces arbres sont étiquetés par un terminal ou par un non terminal, et les noeuds non feuilles sont étiquetés par des non terminaux. Ce sont des noeuds de substitution (↓) ou des noeuds pieds.

Les deux opérations disponibles pour combiner des arbres sont :

- la substitution, obligatoire et unique, qui consiste à insérer un arbre initial ou dérivé sur un noeud à substituer dans un arbre élémentaire ou dérivé (ici N↓) :



— et l'adjonction, facultative et récursive, qui consiste à insérer un arbre auxiliaire ou son dérivé à un noeud de même catégorie (ici SV) :



Le noeud pied et la racine de l'arbre adjoint doivent aussi avoir la même catégorie (ici SV). L'adjonction peut être rendue obligatoire ou interdite, mais elle ne peut avoir lieu sur un noeud de substitution ou un noeud pied.

Lors de l'analyse d'une phrase, ses éléments terminaux sont intégrés à un arbre dérivé en tant que noeuds feuilles par les opérations décrites ci-dessus. La dérivation est complète une fois que tous les noeuds feuilles sont remplacés par des terminaux. Les TAG appartiennent à la classe des grammaires contextuelles (Joshi *et al.*, 1990) et sont plus puissantes que des grammaires hors contexte, tout en restant analysables en temps polynomial (Joshi, 1985).

Vijay-Shanker (1988) présente l'ajout de structures de traits, qui sont des couples attribut-valeur qui peuvent être attribués aux noeuds des arbres. Ils permettent d'introduire des contraintes sur leur unification. Un attribut peut recevoir une valeur absolue ou une variable pour être identique à un autre attribut. Les noeuds d'un arbre ont des structures de traits amont et aval qui doivent s'unifier lors de la dérivation, à l'exception des noeuds à substituer et des noeuds pieds. L'unification de deux structures de traits aboutit à une nouvelle structure, ou échoue, et ne permet pas l'opération de dérivation.

Sans contraintes supplémentaires, motivées linguistiquement, les TAG pourraient utiliser toutes sortes d'arbres qui ne répondraient pas aux réalisations syntaxiques d'une langue. Quatre principes ont permis de les adapter aux langages naturels (Kroch & Joshi, 1985; Abeillé, 1993) :

— principe d'ancrage lexical (ou lexicalisation) : un arbre élémentaire doit avoir au moins une ancre lexicale non vide
— principe de cooccurrence prédicat-arguments : chaque prédicat intègre ses arguments à sa structure élémentaire
— principe de consistance sémantique : un arbre élémentaire ne peut être sémantiquement "vide", les éléments fonctionnels n'étant pas perçus comme autonomes
— principe de non compositionnalité : un arbre élémentaire correspond à une seule unité sémantique

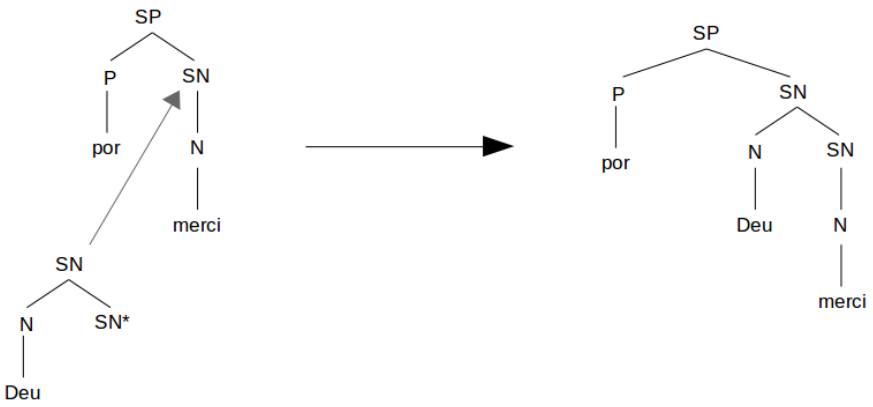


FIGURE 1 – Analyse N2 (déterminatif) + N1 (déterminé)

Il s'agit d'une convention habituellement acceptée par les linguistes. L. 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. La lexicalisation présente un intérêt pour le *parsing* car elle permet d'effectuer un tri avant l'analyse avec les mots de la phrase, qui ont des sous-catégorisations différentes.

Les structures de traits sont également mises au profit de la syntaxe, notamment pour faire peser des contraintes de tête, de pied et d'accord (en genre, nombre et cas selon la langue) sur l'unification de noeuds.

4.1.2 Avantages des TAG pour le français médiéval

Les TAG semblent avoir un pouvoir expressif suffisant pour traiter les états de langue du français médiéval, qui se caractérisent notamment par un ordre des mots plus libre qu'en français contemporain, ce qui pourrait poser problème. Cette hypothèse reste cependant à confirmer. Néanmoins, le domaine de localité étendu des TAG est un atout certain pour traiter les langues naturelles. Il permet une description locale de phénomènes, comme l'attachement d'arguments à leur prédicat, car les arbres élémentaires de la grammaire peuvent avoir une profondeur supérieure à 1, contrairement aux systèmes à base de règles. Il est ensuite possible d'insérer un nombre arbitraire de modifieurs entre ces éléments au moyen de l'adjonction, sans défaire l'arbre de départ. Le verbe peut par exemple se trouver éloigné de son objet, mais cela n'a pas de conséquence sur l'analyse.

Il est possible de lier des arbres élémentaires à certains mots du lexique, et ainsi de créer des arbres très spécifiques, n'étant utilisés qu'avec ces mots, comme dans le cas du complément déterminatif précédant son gouverneur syntaxique, qui n'est réalisé qu'avec certains mots (cf. section 3.2 et fig. 1).

Les grammaires d'arbres adjoints semblent donc tout indiquées pour l'analyse de langues naturelles, permettant à la fois la description générale de phénomènes syntaxiques et la gestion de cas particuliers. Une grammaire à large couverture est en revanche un objet complexe et difficile à maintenir à cause du grand nombre d'arbres qui la composent. Ceux-ci ont habituellement de nombreux points communs puisqu'ils représentent les différentes déclinaisons possibles de phénomènes. Il serait donc possible de factoriser ces arbres, grâce notamment au domaine étendu de localité. Un formalisme plus abstrait permet de générer ces arbres factorisés à partir d'une description modulaire de la syntaxe : les métagrammaires.

4.1.3 Modularité d'une métagrammaire

Une métagrammaire (Candito, 1999) fournit une description modulaire et hiérarchique d'une langue, organisée en classes. Celles-ci expriment des contraintes sur les constituants possibles des arbres de la grammaire et sur leurs relations (ex. accord, dominance...). Elle est destinée à être compilée en une grammaire d'arbres adjoints.

Pour ce faire, elle doit exprimer explicitement les trois "dimensions" (Candito, 1999) qui contrôlent les arbres. Selon le principe d'ancrage lexical, 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 catégorie grammaticale du lemme principal d'un arbre (son ancre) a ainsi une sous-catégorisation initiale, qui constitue la première dimension. Par exemple, le verbe "chasser", en ancien français comme en français contemporain, attend deux arguments (dans une diathèse (ou "voix") active, qui sert de cadre canonique) : un sujet et potentiellement un objet.
ex. "et là assortirent grant nombre d'artillerie, **qui** d'entrée *chassa* **tous les gens du duc de Calabre** hors du villaige de Charenton" (*Mémoires I* de Philippe de Commynes [2], p.61)
La valence de ce lemme peut être redistribuée, le sortant ainsi de son emploi "canonique", qui est décrit comme standard par la grammaire. Cette redistribution, la deuxième dimension, change la sous-catégorisation du lemme et potentiellement le nombre d'arguments représentés.
ex. avec nombre d'arguments conservé : "car **il** en *estoit chassé* **par le conte de Warvic**" (*Mémoires II*, p.139), où on trouve un complément d'agent qui serait sujet à la voix active
ex. avec modification du nombre d'arguments (sans complément d'agent) : "car **Pierre de Medicis** *fut chassé* ce jour" (*Mémoires VIII*, p.11)
Enfin, les fonctions syntaxiques peuvent être réalisées de différentes manières, selon l'hypothèse distributionnelle, ce qui constitue le dernier paramètre contrôlant la construction des arbres.

Ces informations doivent être représentées dans la métagrammaire, ainsi que l'architecture des arbres initiaux et adjoints. Par souci d'efficacité et pour décrire au mieux la langue, il est important d'être le plus général possible. En factorisant l'information linguistique, on peut répartir les contours généraux ou communs à plusieurs catégories dans des classes mères dont des classes plus spécifiques héritent. Par exemple, les modaux, auxiliaires et 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. Diviser ainsi l'information facilite la compréhension et la maintenance de la métagrammaire. Les différents arbres produits sont ensuite combinés par les opérations d'adjonction et de substitution pour former une TAG.

4.2 *French Metagrammar* (FRMG)

FRMG [3] est un ensemble de 451 classes produisant 381 arbres. Cette métagrammaire a été développée parallèlement au lexique *Lefff* (Sagot, 2010), avec lequel elle partage des types et la notion d'*hypertag*, rendant possible l'ancrage des arbres de la grammaire par des mots du lexique.

2. Cet exemple est issu de la *Base de Français Médiéval* (BFM - *Base de Français Médiéval* [En ligne]. Lyon : ENS de Lyon, Laboratoire IHRIM, 2016, <txm.bfm-corpus.org>).
3. FRMG est disponible à cette adresse : `http://alpage.inria.fr/frmgwiki/`.

4.2.1 Expressivité de la métagrammaire

FRMG est une grammaire à large couverture. Depuis 2004, les campagnes d'évaluation successives ont permis de la confronter à de nouvelles données, comme des écrits journalistiques ou scientifiques. L'entraînement du *parser* et l'apprentissage de poids de désambiguïsation sur le *French Treebank* l'ont rendue plus robuste (Villemonte de la Clergerie, 2013). Actuellement, les *tweets* constituent un défi parce qu'ils présentent une ressource non standard, difficile à traiter. Ils introduisent de fortes variations langagières, de nouveaux phénomènes syntaxiques, et un usage différent de la ponctuation. La métagrammaire a été enrichie pour intégrer ces nouveautés.

Les métagrammaires ont été longtemps mises en avant pour traiter la valence des verbes et les structures verbales standard. Elles proposent en effet un cadre de description simple grâce à leur modularité et au domaine étendu de localité (si elles sont fondées sur le formalisme des TAG). Cependant, E. Villemonte de la Clergerie (2012) souligne aussi leur apport pour l'analyse des modifieurs, dont la multiplicité dépasse celle des structures verbales. Ces dernières sont décrites par quelques arbres dans FRMG, tandis que les modifieurs, qui constituent un enjeu majeur, restent peu explorés. Leur variété et leur nombre en font des phénomènes difficiles à capturer, en particulier s'ils portent sur la phrase. Il est possible de les intégrer à l'arbre d'analyse à plusieurs noeuds différents. Les modifieurs peuvent prendre diverses formes, certaines étant difficiles à analyser, comme des syntagmes nominaux ou des locutions plus ou moins figées. Ce sont généralement des contraintes sémantiques qui permettront d'identifier de tels groupes. Ces cas spécifiques sont décrits au terme d'un long recensement. Les arbres de modifieurs sont très nombreux dans FRMG, mais il s'agit de descriptions simples. Ils pourraient être factorisés, à l'instar des arbres de structures verbales.

4.2.2 Expressivité du langage SMG

La métagrammaire FRMG s'appuie sur le langage de représentation *Simple Metagrammar* (SMG) et son compilateur, *mgcomp*, qui offrent de nombreuses possibilités d'expression pour décrire une langue. Tout d'abord, des contraintes comme l'égalité, la dominance ou la précédence entre noeuds permettent de décrire des arbres élémentaires.

Lors de la compilation de la grammaire, les arbres sont factorisés au moyen d'opérateurs (Villemonte de la Clergerie, 2010) qui sont des extensions du formalisme TAG. La disjonction permet d'introduire des alternatives dans la dérivation. Les gardes (Villemonte de la Clergerie, 2010) permettent la gestion des noeuds optionnels. La répétition est possible grâce à l'étoile de Kleene. Elle est surtout utilisée pour la coordination. Enfin, l'entrelacement permet un ordre libre entre noeuds frères. Il est notable que ces opérations n'ont pas d'impact sur l'expressivité ou la complexité des TAG, mais ils permettent de réduire la taille de la grammaire, rendant son usage plus efficace.

La description de la langue dans une métagrammaire repose en partie sur les *hypertags*, des structures de traits associées à chaque mot du lexique, contenant des informations syntaxiques comme la valence et les réalisations possibles. Ces éléments sont générés en même temps que la grammaire et permettent l'ancrage du lexique dans les arbres.

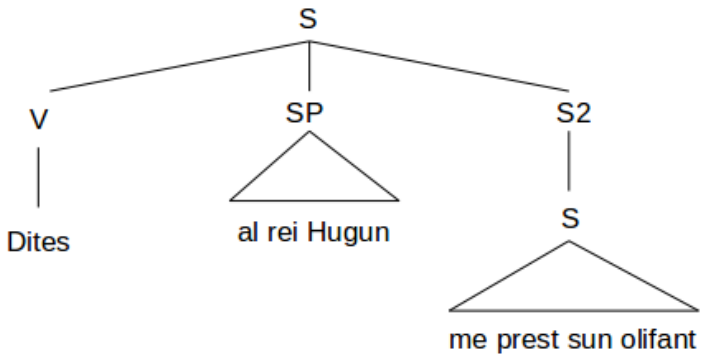


FIGURE 2 – Analyse d'une complétive sans "que"

4.3 Motivations et limites de la démarche

Plutôt que créer entièrement une nouvelle métagrammaire, nous avons choisi d'adapter FRMG au français médiéval parce que les états de langue qui le composent présentent un nombre conséquent de similarités par rapport au français contemporain. La langue moderne est aussi soumise au changement. Par exemple, pour traiter la disparition du "que" devant une complétive (Boutin, 2007), FRMG l'a rendu optionnel [4]. Son absence n'empêche plus l'analyse de la phrase. L'origine de cette tendance n'a a priori pas de lien avec le français médiéval, qui présente toutefois la même caractéristique.
ex. "Dites al rei Hugun me prest sun olifant" (*Le Voyage de Charlemagne à Jérusalem et à Constantinople*, 471, cf. fig.2), trad. par Buridant (2000) : *Dites au roi Hugon qu'il me prête son olifan.*
La conjonction peut être élidée car le subjonctif dans la subordonnée est un indice suffisant. Il semble donc que les différents états de langue présentent une syntaxe similaire et des cas de variation. Ceux-ci peuvent être inclus à la métagrammaire, qui est aisément modifiable.

En revanche, il n'est pas aisé de déterminer la probabilité d'une analyse. Lors de la désambiguïsation, des analyses sont préférées à d'autres en fonction du poids qui leur est attribué, obtenu grâce à un apprentissage sur *treebank* (Villemonte de la Clergerie, 2013). Ces poids seront appris sur le SRCMF, et seront très différents de ceux appris pour FRMG à partir du *French Treebank*. Cependant, le formalisme des *treebanks* ne correspond pas à celui des métagrammaires, ce qui limite l'apport d'une ressource pour déterminer les poids des analyses.

4.4 Leviers d'action pour adapter la métagrammaire

4.4.1 Relâchement de contraintes

Dans un premier temps, nous pouvons considérer la tâche d'adaptation de FRMG comme le simple relâchement de nombreuses contraintes. Une description plus générale, moins contrainte en amont, permet à la grammaire de traiter ces données hétérogènes. Le français médiéval peut être vu comme une succession d'états de langue préfigurant le français contemporain, avec plus de variations, une graphie non fixe et des règles d'accord moins respectées. Encore en moyen français, C. Marchello-Nizia (1979) relève des cas d'absence d'accord verbal avec le sujet, qu'il soit en position préverbale

4. Une garde porte sur la conjonction. Si celle-ci est absente, l'analyse de la subordonnée peut être faite, mais elle est pénalisée lors de la désambiguïsation car cette tournure reste rare en français contemporain.

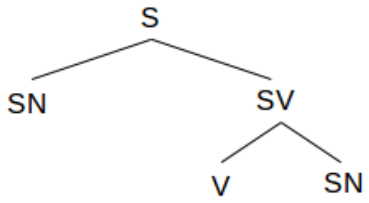


FIGURE 3 – Représentation traditionnelle des constituants principaux

ou postverbale.
ex. "si luy **avoit** *jeunesse et crainte* les yeulz si bandez que en rien il ne s'appercevoit du bien qu'on luy vouloit" (*Cent nouvelles nouvelles*, tel que cité par C. Marchello-Nizia), trad. *La jeunesse et la crainte lui **avaient** bandé les yeux, de sorte qu'il ne voyait pas qu'on lui voulait du bien.*
Il est plus fréquent, et ce dès l'ancien français, de trouver un substantif singulier à valeur collective (*gent, peuple...*) comme sujet d'une verbe au pluriel (Hasenohr & Raynaud de Lage, 1993; Buridant, 2000). Inversement, il est possible de trouver des verbes au singulier alors qu'ils ont plusieurs sujets exprimés (voir l'exemple de *Cent nouvelles nouvelles* ci-dessus). L'accord du participe passé varie beaucoup plus qu'en français contemporain. Employé avec *avoir*, il peut s'accorder avec l'objet placé avant ou après l'auxiliaire, ou même avec un substantif proche, rattaché à l'objet. Employé avec l'auxiliaire *être* (*estre*), il s'accorde avec le sujet, sauf, parfois, si celui-ci est postposé. Rendre plus souple la contrainte d'accord verbal semble donc être une mesure importante pour le traitement du français médiéval.

Des études ont été faites pour confirmer et quantifier l'impact des métadonnées sur le *parsing* (Guibon *et al.*, 2014, 2015). Les auteurs ont évalué des modèles appris sur différents ensembles d'apprentissage, et ont observé que le dialecte et la date sont les métadonnées les plus discriminantes (même si le domaine du texte et sa forme sont aussi une cause importante de variation). Dans leur article, T. Rainsford *et al.* (2012) tirent la même conclusion : "à des caractéristiques externes différentes peuvent être associées des propriétés langagières différentes". Leur étude de la zone préverbale en ancien français met en lumières des différences dues à la forme des textes. Sur l'ensemble du corpus étudié, la zone préverbale n'accueille habituellement qu'un élément, l'ancien français étant une langue majoritairement à verbe second (V2). Cependant, en prose les exceptions sont assez régulières. Les phrases à deux éléments préverbaux ou plus ont un circonstant (souvent *si*), ou une subordonnée précédant le sujet, ou un circonstant à valeur énonciative. Les textes en vers présentent plus de variété. On trouve par exemple plus souvent l'antéposition de deux satellites. Il semble donc qu'il vaut mieux privilégier une description très générale des phénomène syntaxiques pour rendre acceptables les multiples variations existantes.

4.4.2 Contraintes différentes

L'ordre libre des mots est une des principales caractéristiques qui différencie le français médiéval de la langue contemporaine. FRMG a une représentation traditionnelle des constituants principaux (cf. fig.3) : le sujet et le syntagme verbal (SV) sont des noeuds frères, tandis que l'objet est sous SV (Thomasset & Villemonte de la Clergerie, 2005).

Pour autoriser un ordre libre de ces constituants, nous avons tiré profit de l'entrelacement (symbolisé par #), et avons donc remonté l'objet au même niveau que les deux autres (cf. fig.4). La grammaire

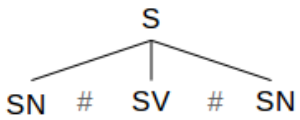


FIGURE 4 – Représentation des constituants principaux dans notre métagrammaire

peut ainsi reconnaître les six ordres possibles du français médiéval : SOV, SVO, OSV, OVS, VSO, VOS.

Pour cette métagrammaire du français médiéval, nous ne prenons pas en compte la ponctuation. Elle n'est pas présente dans le SRCMF, qui sera utilisé pour évaluer la métagrammaire et entraîner les modèles de désambiguïsation. Ce corpus est cependant segmenté en phrases. Les descriptions de phénomènes habituellement accompagnés de marques de délimitation, comme l'incise, ne peuvent donc pas reposer sur la présence de celles-ci. Nous n'excluons cependant pas d'ajouter la ponctuation à notre métagrammaire dans un second temps, si sa présence est souhaitable.

4.4.3 Facettes

Nous prévoyons d'implémenter un nouveau mécanisme pour filtrer l'analyse syntaxique selon les métadonnées des textes. Certaines contraintes ne s'appliquent qu'à des états de langue identifiés, et nous voulons restreindre leur activation à ces cas particuliers. Par exemple, les formes atones des pronoms objet sont placées avant le verbe, mais elles peuvent être postposées en picard, et même être élidées (Buridant, 2000). En nous appuyant sur des informations comme la date, le dialecte, la forme ou le domaine, nous voulons pouvoir permettre ou interdire certaines analyses, ou leur donner un poids plus ou moins important. Nous ne savons pas encore si nous souhaitons produire une grammaire par ensemble de facettes ou s'il est préférable d'appliquer des gardes à une seule grammaire et de les activer dynamiquement lors de l'analyse.

5 Conclusion et perspectives

L'adaptation de FRMG au français médiéval semble prometteuse, même si ces états de langue ancienne posent de nombreux défis pour l'analyse automatique. Des changements sont indispensables dans la métagrammaire. Il peut s'agir du simple relâchement de certaines contraintes pour ne décrire que ce qui est universel dans la grammaire, ou des changements de description. Nous souhaitons développer une métagrammaire assez générale pour traiter tous les phénomènes, quelle que soit leur fréquence dans les textes à annoter, mais en réduisant au mieux le nombre d'analyses possibles pour limiter la complexité.

De nombreuses problématiques restent à aborder, dont la description de phénomènes et l'inclusion d'un mécanisme de facettes pour filtrer l'application de ces contraintes. La fouille d'erreurs en sortie du *parser* permettra d'accélérer la mise au point de la grammaire et du lexique. A terme, nous voulons développer une méthodologie réutilisable pour traiter des données hétérogènes, incluant des arbres de décision en amont pour modifier le modèle ou la grammaire elle-même en fonction des métadonnées.

Références

ABEILLÉ A. (1993). *Les Nouvelles Syntaxes*. Armand Colin.

ABEILLÉ A. (2002). *Une grammaire électronique du français*. CNRS Editions.

BENDER E. M., FLICKINGER D. & OEPEN S. (2002). The grammar matrix : An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. *COLING-02 on Grammar engineering and evaluation*.

BOUTIN A. B. (2007). *De et que subordonnants, et variation en français*. *Linx* [En ligne], 57, mis en ligne le 15 février 2011. http ://journals.openedition.org/linx/280.

BURIDANT C. (2000). *Grammaire nouvelle de l'ancien français*. Sedes.

CANDITO M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées, application au français et à l'italien*. PhD thesis, Université Paris 7.

CRABBÉ B. (2005). *Représentation informatique de grammaires fortement lexicalisées, Application à la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2.

GUIBON G., TELLIER I., CONSTANT M., PRÉVOST S. & GERDES K. (2014). Parsing poorly standardized language dependency on old french. In V. HENRICH, E. HINRICHS, D. DE KOK, P. OSENOVA & A. PRZEPIÓRKOWSKI, Eds., *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, p. 51–61, Tübingen, Allemagne.

GUIBON G., TELLIER I., PRÉVOST S., CONSTANT M. & GERDES K. (2015). Searching for discriminative metadata of heterogenous corpora. In M. DICKINSON, E. HINRICHS, A. PATEJUK & A. PRZEPIÓRKOWSKI, Eds., *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, p. 72–82, Varsovie, Pologne.

GUILLOT-BARBANCE C., HEIDEN S. & LAVRENTIEV A. (2017). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques n.7*, p. 168–184.

HASENOHR G. & RAYNAUD DE LAGE G. (1993). *Introduction à l'ancien français de Guy Raynaud de Lage*. Sedes.

JOSHI A. K. (1985). Tree adjoining grammars : How much context-sensitivity is required to provide reasonable structural descriptions ? *Natural Language Parsing*.

JOSHI A. K., LEVY L. S. & TAKAHASHI M. (1975). Tree adjunct grammars. *Journal of Computer and System Sciences*.

JOSHI A. K., VIJAY-SHANKER K. & WEIR D. (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 L. (2010). *Parsing beyond context-free grammars*. Springer.

KINYON A. (2000). Hypertags : Beyond pos tagging. In D. N. CHRISTODOULAKIS, Ed., *Lecture Notes in Computer Science*, Berlin, Heidelberg : Springer.

KROCH A. S. & JOSHI A. K. (1985). The linguistic relevance of tree adjoining grammar. *University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-85-16.*

MARCHELLO-NIZIA C. (1979). *Histoire de la langue français aux XIV^{e} et XV^{e} siècles*. Bordas.

PRÉVOST S. & STEIN A. (2013). *Syntactic Reference Corpus of Medieval French (*SRCMF*)*. [version 0.92]. ENS de Lyon/ILR Stuttgart.

RAINSFORD T., GUILLOT-BARBANCE C., LAVRENTIEV A. & PRÉVOST S. (2012). La zone préverbale en ancien français : apport des corpus annotés. *SHS Web of Conferences*.

ROCIO V., ALVES M. A., LOPES J. G., XAVIER M. F. & VICENTE G. (2003). Automated creation of a medieval portuguese partial treebank. In A. ABEILLÉ, Ed., *Treebanks : Building and Using Parsed Corpora*.

SAGOT B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. *7th international conference on Language Resources and Evaluation (LREC 2010)*.

SCHØSLER L. (1984). *La déclinaison bicasuelle de l'ancien français : son rôle dans la syntaxe de la phrase, les causes de sa disparition*. Odense University Press.

THOMASSET F. & VILLEMONTE DE LA CLERGERIE E. (2005). Comment obtenir plus des méta-grammaires. In *Actes de TALN 2005 (Traitement automatique des langues naturelles)*, Dourdan : ATALA.

VIJAY-SHANKER K. (1988). *A Study of Tree Adjoining Grammars*. PhD thesis, University of Pennsylvania.

VILLEMONTE DE LA CLERGERIE E. (2005). From metagrammars to factorized tag/tig parsers. In H. BUNT, R. MALOUF & A. LAVIE, Eds., *Proceedings of the Ninth International Workshop on Parsing Technology*, p. 190–191, Vancouver, Canada : Association for Computational Linguistics.

VILLEMONTE DE LA CLERGERIE E. (2010). Building factorized tags with meta-grammars. *The 10th International Conference on Tree Adjoining Grammars and Related Formalisms - TAG+10*.

VILLEMONTE DE LA CLERGERIE E. (2012). Etude du traitement de certains compléments de phrase dans le cadre d'une méta-grammaire. In J. RADIMSKY, Ed., *LGC'12 - 31ème 30ème Colloque international sur le Lexique et la Grammaire*, Nové Hrady, République tchèque.

VILLEMONTE DE LA CLERGERIE E. (2013). Improving a symbolic parser through partially supervised learning. In *The 13th International Conference on Parsing Technologies (IWPT)*, Naria, Japan : Springer.