Approche multi-niveaux pour l'analyse des données textuelles non-standardisées : corpus de textes en moyen français

Mourad Aouini

THESE DE DOCTORAT DE L'ETABLISSEMENT UNIVERSITE BOURGOGNE FRANCHE-COMTE PREPAREE A L'Université de Franche-Comté

Ecole doctorale LECLA n°592

Ecole doctorale « LETTRES, COMMUNICATIONS, LANGUES, ARTS »

Doctorat en Ingénierie Linguistique

Par

Mourad AOUINI

Approche multi-niveaux pour l'analyse des données textuelles non-standardisées : corpus de textes en moyen français

Thèse présentée et soutenue à Paris, le 19 mars 2018

Composition du Jury :

Samir Mbarki, Professeur à l'université Ibn Tofail, Rapporteur

Céline Guillot, Maitresse de conférences à l'école normale supérieure de Lyon, Rapporteur

Kim Gerdes, Maître de conférences à l'université Sorbonne Nouvelle Paris III, Examinateur

Max Silberztein, Professeur à l'université Bourgogne Franche-Comté, Directeur de thèse

Jean-Philippe Genet, Professeur à l'université Paris I Panthéon-La Sorbonne, Codirecteur de thèse

Remerciements

En premier lieu, je tiens à remercier mon directeur de thèse, Max Silberztein, pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour ses multiples conseils et pour toutes les heures qu'il a consacrées à diriger cette recherche. Il a toujours été disponible, à l'écoute de mes nombreuses questions, et s'est toujours intéressé à l'avancée de mes travaux. Les nombreuses discussions que nous avons eues ainsi que ses conseils sont pour beaucoup dans le résultat final de ce travail. Enfin, j'ai été extrêmement sensible à ses qualités humaines d'écoute et de compréhension tout au long de ce travail doctoral.

Je souhaiterais exprimer ma gratitude à mon co-directeur de thèse Jean-Philippe Genet pour m'avoir permis d'intégrer l'équipe du projet « *Signs And States* » en me proposant un sujet très intéressant et m'a laissé la liberté de le re-orienter au cours du déroulement de ma thèse. Il a souvent attiré mon attention sur certains problèmes historiques, linguistiques et extralinguistiques. Je le remercie également pour son accueil chaleureux à chaque fois que j'ai sollicité son aide, ainsi que pour ses multiples encouragements, notamment lors de mes premières communications et premiers séminaires.

Je voudrais remercier madame Céline Guillot d'avoir accepté de me consacrer son temps en examinant le manuscrit et d'avoir accepté la fonction de rapporteur.

Mes remerciements vont également à monsieur Samir Mbarki pour l'intérêt qu'il a manifesté à l'égard de cette recherche en s'engageant à être rapporteur.

Je sais infiniment gré à monsieur Kim Gerdes de m'avoir prodigué maints conseils depuis mes études de master d'ingénierie linguistique et d'avoir accepté de participer à ce jury de thèse.

Je tiens à remercier l'ensemble du laboratoire LAMOP dans lequel j'ai passé quelques années qui ont été l'occasion de rencontres, d'échanges et de collaborations notamment avec Aude Mairie, Christopher Fletcher, Naomi Kanoaka, Chloe Morgan, Laura Albiero, Rachel Moss, Lauren Henras, Hicham Idabal, Stephane Lamassé, Anne Tournieroux, Claire Priol et Carla Bozzolo.

Je remercie également les membres du laboratoire LLACAN pour le soutien que l'on m'y a apporté durant mon passage de 2 ans, plus particulièrement Christian Chanard, Tahar Medour, Amina Metouchi, Jeanne Zerner, Magali Sansonetti, Danielle Bonardelle, Benoit Legouy et Bernard Caron pour leurs encouragements, leur sourire et tous les moments que l'on a passés ensemble. Mes plus vifs remerciements vont aussi à Paulette Roulon-Doko pour

sa relecture et ses conseils toujours très pertinents. À Marie-Claude Simeone-Senelle, merci pour son soutien et ses conseils.

Je tiens également à remercier l'UPS-2259 Cultures, Langues, Textes pour l'accueil, le soutien et les conditions de travail qui m'ont été offertes, et tout particulièrement par Fabrice Jejcic, Bernard Weiss et Isabelle Michel qui me donnaient de l'élan pour continuer malgré les difficultés rencontrées.

J'adresse aussi mes remerciements aux membres de la communauté NooJ avec lesquels j'ai beaucoup échangé durant les nombreux séminaires et conférences qu'on a passés ensemble.

Ma reconnaissance va à ceux qui ont plus particulièrement assuré le soutien affectif de ce travail doctoral : mes parents Salwa et Youssef, mes frères Bechir, Riadh et Bessem et ma fiancée Nissen. Merci.

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

> *Les phénomènes dont traite la grammaire sont à un certain niveau expliqués par les règles de la grammaire elle-même et par l'interaction de ces règles. A un niveau plus élevé, ces mêmes phénomènes sont expliqués par les principes qui déterminent le choix de la grammaire sur la base de l'expérience limitée et incomplète dont disposait la personne qui a acquis la connaissance de la langue et qui s'est construit cette grammaire particulière. Les principes qui déterminent la forme de la grammaire et qui choisissent une grammaire de forme appropriée sur la base de certains faits constituent un sujet qui pourrait, selon l'usage traditionnel, être appelé « grammaire universelle ». L'étude de la grammaire universelle ainsi comprise est une étude de la nature des capacités intellectuelles humaines.* *(Chomsky, 1990)*

Contexte et motivation

L'avènement de l'ère numérique a été marqué par une expansion considérable de la production de données. En effet, l'accessibilité des technologies de communication et de l'information a permis une mise à disposition de grands volumes de données produites. L'exploitation de ces données suppose une capacité à effectuer des analyses massives sur des gros volumes et à traiter toutes les variétés des données qui, pour la plupart, peuvent être hétérogènes. Nous constatons que la plus grande partie des données créées par ces systèmes, plus particulièrement sur la toile, sont des données textuelles non-structurées.

A l'heure de l'explosion de l'usage des réseaux sociaux, ces données se génèrent avec une fréquence très élevée souvent qualifiée d'exponentielle et une présence tout à fait remarquable des données textuelles non-standardisées. Ces dernières sont élaborées à partir d'un vocabulaire qui n'a pas été défini par des dictionnaires de référence. Elles ont une orthographe instable et déréglée et une grammaire descriptive dont les règles n'ont pas été décrites et ne permettent pas non plus de juger la grammaticalité de l'énoncé. Elles sont issues essentiellement des langues à tradition orale ou de variétés des langues standardisées.

L'analyse des données textuelles non-standardisées est devenue, comme nous l'avons déjà souligné, un enjeu majeur et indispensable pour l'analyse d'une grande partie des données produites en langues modernes sur la toile. Cependant, cette problématique concerne également les langues anciennes. Le moyen français, qui est une langue en pleine évolution dont l'orthographe, le système flexionnel et la syntaxe ne sont pas stables, en est une

illustration plutôt éloquente et exemplaire. En effet, les textes en moyen français se singularisent principalement par l'absence d'orthographe normalisée et par la variabilité tant géographique que chronologique des lexiques médiévaux. L'analyse de tel type de données textuelles nécessite leur enrichissement avec des informations de nature interprétative à l'aide de systèmes d'annotation automatique. Par conséquent, un système d'analyse multi-niveaux permettant une analyse lexicale, un étiquetage morphosyntaxique et une reconnaissance des entités nommées s'avère indispensable pour préparer ces données à des éventuelles exploitations par les systèmes informatiques.

Plan de la thèse

Ce manuscrit se compose de deux parties, la première s'intitule « Traitement automatique des langues et moyen français » et la seconde « Analyse automatique du moyen français ».

La première partie aborde les champs de recherches en traitement automatique des langues et présente les spécificités du moyen français considéré comme une langue non-standardisée. Elle commence par un chapitre consacré aux différents niveaux d'analyse linguistique automatique à savoir lexicale, syntaxe locale, syntaxe structurelle, sémantique locale, sémantique propositionnelle et pragmatique. Suit une introduction des principales approches classiques permettant la mise en place d'un système TAL, que ce soit par apprentissage ou à base de règles. Le deuxième chapitre expose l'état de l'art des méthodes d'annotation automatique des corpus, plus précisément les techniques d'étiquetage morphosyntaxique et les systèmes de reconnaissance des entités nommées. Le troisième et dernier chapitre de cette partie met en valeur les spécificités du moyen français et les différents projets TAL. Ceux-ci ont pour objectif la numérisation et la constitution d'un corpus en moyen français et la mise en place de quelques analyses principalement des étiqueteurs morphosyntaxiques. Pour finir, ce chapitre fait une présentation du MEDITEXT, un corpus regroupant des textes politiques en moyen français apparus entre la fin du XIIIème et le XVème siècle.

La seconde partie est consacrée à l'analyse automatique du moyen français. Le quatrième chapitre propose un tour d'horizon des principales caractéristiques des dictionnaires électroniques les plus répondus à savoir le réseau lexical WordNet, le standard Lexical Markup Framework LMF, les dictionnaires DELA et les dictionnaires NooJ. Suit une présentation détaillée du processus de construction d'un dictionnaire électronique en moyen français. Celle-ci utilise la description des éléments du vocabulaire proposée par le dictionnaire électronique NooJ qui s'inscrit dans une approche de formalisation des langues.

Le chapitre « Analyse lexicale des textes en moyen français » traite divers phénomènes linguistiques pour lesquels toutes les segmentations possibles d'une séquence des caractères sont proposées et sont enrichies par des descriptions morphosyntaxiques. Ensuite, une méthode symbolique d'étiquetage morphosyntaxique est proposée au chapitre 6 afin de désambiguïser les éléments du vocabulaire et de faire, selon le contexte, un choix automatique d'une des étiquettes morphosyntaxiques associées à la forme. Enfin, le dernier chapitre décrit un système multiple de reconnaissance des entités nommées. Il met en œuvre diverses ressources linguistiques qui permettent la modélisation des preuves internes et externes, dans le but d'identifier et de catégoriser des séquences d'unité de vocabulaire en noms de lieu, nom de personne, nom d'institution, expressions temporelles et expressions numériques.

Première Partie

Traitement automatique des langues et moyen français

Chapitre 1

Le traitement automatique des langues

1. Introduction

Avec l'importante évolution en puissance, en vitesse et en fiabilité des matériels informatiques, les capacités de stockage, de traitement et de partage des données ont explosé et les réseaux se sont accrus avec la naissance du réseau des réseaux « Internet ». Résultats : l'informatique s'est universalisée s'appliquant à toutes les activités humaines. La numérisation et la production des documents électroniques sont devenues de plus en plus importantes et « les flux d'informations transitant par internet ont une croissance exponentielle » (Campedel & Hoogstoël, 2011). Selon l'IDC[1], en 2014, le volume des données numériques produites devrait être multiplié par 10 en 2020 et, selon Sinequa[2], 80 % des données générées par les entreprises sont des données textes non-structurées. En outre, une grande part des données générées sur la toile sont des données textuelles non-standardisées produites par les utilisateurs des réseaux sociaux, des sites d'avis et les divers forums de chat et de discussion. Cette expansion des données textuelles a fait apparaître des développements applicatifs divers centrés sur les données afin de structurer, classer, détecter les informations pertinentes et synthétiser les documents électroniques. Par conséquent, le développement des systèmes d'extraction d'informations pertinentes fiables est devenu un enjeu majeur et des nouveaux besoins applicatifs sont apparus dans le but de résoudre des problématiques linguistiques bien précises telles que l'étiquetage morphosyntaxique et la reconnaissance des entités nommées.

---

[1] International Data Corporation (IDC) est une entreprise américaine spécialisée dans la réalisation d'études de marché dans les domaines des technologies de l'information et de la communication et de l'électronique grand public *(http://www.idc.fr/).* Dans une étude parue en avril 2014, elle estime que le volume de données produites dans l'Univers Digital devrait être multiplié par 10 entre 2013 (4,4 Zettabytes) et 2020 avec (44 Zettabytes). (https://france.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf)

[2] Sinequa est un éditeur de moteurs de recherche d'entreprise indépendants *(https://www.sinequa.com/).* Dans un article publié par la conférence BigData Paris 2017, la compagnie affirme que « *plus de 80% des données*

2. Objectifs

Le traitement automatique des langues (TAL) est le résultat du croisement et de l'interaction entre informatique et linguistique. Il permet de coder l'alphabet et les éléments atomiques d'une langue et de modéliser les phénomènes linguistiques. Depuis le « test de Turing »[3], le TAL avait comme objectif le développement d'un système capable de comprendre, de traduire et de générer un texte en langue naturelle. Dans cette optique, des systèmes ont été mis en place tel Eliza (Weizenbaum, 1966). L'évaluation des performances de ces systèmes a p ermis aux experts de tenir compte de l'ampleur et de la complexité des tâches à résoudre. Avec l'évolution des études sur la grammaire et la sémantique formelles, des nouvelles techniques de représentation des connaissances et de formalisation des raisonnements sont apparues permettant ainsi le développement des systèmes experts contextuels et pragmatiques capables d'étudier l'utilisation du langage dans un contexte bien déterminé. Ces systèmes de compréhension des langues sont fondés sur des modèles symboliques abstraits qui formalisent les règles à appliquer ainsi que les faits décrivant le problème à résoudre. Depuis les années 90, l'expansion des données textuelles a freiné le développement des systèmes experts afin de laisser place à des approches descriptives et empiriques qui essaient de répondre à des nouveaux besoins applicatifs fondés sur les données parmi lesquels figurent la classification des textes, l'analyse d'opinion et la recherche d'information textuelle. En effet, les formalismes et les algorithmes, qui constituent le moteur d'inférences permettant de traiter les données langagières, ont atteint un niveau de performance convenable. Or, il s'est avéré essentiel au bon fonctionnement de ces systèmes l'exploitation des données linguistiques capables de décrire exhaustivement une langue. Ces données appelées « ressources linguistiques », qui peuvent être des corpus, des dictionnaires électroniques, des lexiques et des grammaires, forment la représentation des connaissances du système. Elles sont un enjeu majeur et un élément central pour tout système TAL. Elles ne sont pas indépendantes des algorithmes qui les exploitent et leurs formats, leurs volumes ainsi que leurs structures sont pris en compte en phase de spécification et de conception pour une analyse fiable et robuste. Nous distinguons des approches symboliques, mises en œuvre à

[3] Le « *test de Turing* » ou « *jeu d'imitation* » est un test d'intelligence artificielle qui consiste à évaluer et à cerner la capacité d'une machine ou d'un agent conversationnel à assurer un échange textuel (conversation instantanée ou chat en ligne) avec un interlocutneur humain. Il a été décrit par Alan Turing en 1950 dans sa publication *Computing machinery and intelligence*.

l'aide des formalismes linguistiques ou à l'aide des plateformes à notation unifiée capable de traiter et faire communiquer plusieurs niveaux d'analyses linguistiques et des approches stochastiques qui sont principalement issues de l'apprentissage automatique et de la statistique descriptive multidimensionnelle.

3. Les niveaux d'analyse du langage

Les applications d'analyse des données langagières ont des besoins variés qui nécessitent souvent le développement de modules qui correspondent à plusieurs niveaux d'analyse linguistique. Ces modules sont souvent mis en place à l'aide des techniques différentes qui ne sont que partiellement compatibles. Par conséquent, ils ne visent pas la mise en œuvre d'un système en TAL cohérent à plusieurs composants capable d'établir une compréhension d'un énoncé en langue naturelle. En effet, comprendre artificiellement un énoncé nécessite une identification des différents niveaux de connaissances impliqués et une acquisition du raisonnement permettant d'exploiter ces connaissances. L'architecture d'un tel système en TAL procède à une formalisation des différents niveaux d'analyse linguistique, illustrés à la figure 1, à savoir lexical, syntaxique, sémantique et pragmatique. Quel que soit le niveau d'analyse, les résultats produits sont des informations associées à une portion du texte appelée « annotations ». Ces annotations peuvent être stockées en utilisant des systèmes de balisage du texte comme le XML-TEI qui associe à chaque portion du texte une balise ou en utilisant une structure multicouche comme la structure d'annotation du texte (TAS) définie dans le cadre du projet de formalisation d'une langue avec l'approche NooJ (Silberztein, 2015).

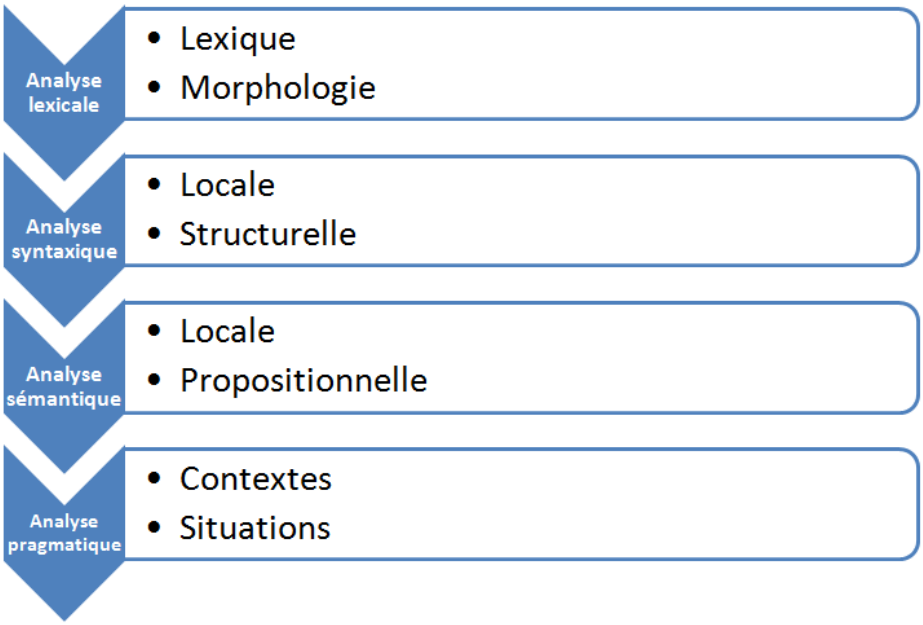


Figure 1. Les différents niveaux d'analyses du langage

3.1. Analyse lexicale

L'analyse lexicale est indispensable pour toute chaîne de traitement en TAL. Elle a pour but d'identifier les unités du vocabulaire du texte (ALU), en partant des séquences de caractères qui le constituent. Ces ALU (en anglais *Atomic Linguistic Unit*) sont des éléments non-analysables qui peuvent être des morphèmes, des mots simples, des mots composés ou des expressions discontinues. L'analyse lexicale contient une étape de normalisation et de segmentation qui permet de standardiser les différentes variantes typographiques, de reconnaître les valeurs numériques et de segmenter, selon un séparateur, les phrases puis les formes. Elle permet également la reconnaissance des formes simples et des formes composées ainsi que celles de leurs structures et de leurs caractéristiques afin de représenter toutes les hypothèses lexicales et les ambiguïtés potentielles. Elle fait appel, principalement, à des dictionnaires électroniques. Ces dictionnaires sont constitués des entrées lexicales auxquelles sont associées diverses informations linguistiques. Ces dernières sont définies selon les phénomènes linguistiques à résoudre et le besoin applicatif pour lequel le dictionnaire a été établi. Il existe donc plusieurs types des dictionnaires électroniques qui sont classés selon la nature des informations linguistiques associées à leurs entrées lexicales. Ils sont évalués selon leur robustesse c'est-à-dire leur capacité de couvrir le vocabulaire d'un texte et la précision et l'exhaustivité des descriptions employées pour décrire les entrées lexicales. En effet, les dictionnaires électroniques sont des formalismes qui offrent les outils permettant la mise en place d'une description des différents usages d'une forme. En d'autres termes, comme le montre l'exemple ci-dessous, pour une forme ambiguë, on distingue les différents emplois possibles selon le contexte et on les associe à des informations linguistiques distinctes. Quelle que soit la langue, les ambiguïtés sont réparties à plusieurs niveaux d'analyse et un dictionnaire électronique devrait fournir les informations permettant de les décrire avec précision.

Forme	**Signification**	**Exemple**	**Informations associées**
roy	Chef souverain de certains États, celui qui règne souverainement sur un pays, sur un	le roy a voulu, pour le bien commun, faire les translater en françois.	roi,NC+SENS=1

	royaume		
roy	Être en mesure de disposer de quelque chose.	ilh n'avoit roy d'iretage ne chevanche nulle	roi,NC+SENS=2

Tableau 1. Les emplois de la forme moyen français « roy »

Les dictionnaires électroniques sont mis en place en utilisant des lexiques et des analyses morphologiques auxquels on associe diverses informations linguistiques. Les lexiques sont généralement constitués d'une forme canonique et d'une liste des variantes associées tandis que l'analyse morphologique relie la forme canonique et les formes orthographiques par suppression et substitution de morphèmes. En effet, les règles morphologiques décrivent les morphèmes, qui sont les plus petites unités de sens, et la description des procédures de décomposition et de recomposition associées (Yvon, 2006). Deux types de règles morphologiques sont distingués : les règles de flexion et les règles de dérivation. Une règle de flexion est la description formelle de la combinaison des affixes flexionnels avec une entrée lexicale afin d'obtenir les différentes formes fléchies associées. À titre d'exemples d'affixes flexionnels, nous citons les affixes qui expriment les conjugaisons des verbes et les variations en genre et en nombre des noms et des adjectifs en français. Une règle de flexion est donc la formalisation d'un paradigme flexionnel. Chaque paradigme flexionnel correspond à une grammaire algébrique qui produit des formes fléchies à partir d'une entrée lexicale donnée en décrivant la combinaison des morphèmes.

Ces paradigmes peuvent être modélisés via une grammaire algébrique en utilisant les expressions rationnelles. Par exemple, l'expression rationnelle du paradigme « Artiste », décrite à l'aide d'un langage formel prédéfini par la plateforme NooJ, permet de décrire la flexion de tous les noms et les adjectifs en français qui sont invariables en genre mais prennent un « s » au pluriel.

Artiste = <E>/m+s | <E>/f+s| s/m+p | s/f+p;

Les paradigmes flexionnels peuvent également être modélisés sous forme de graphe comme le montre la figure 2. Ce graphe flexionnel NooJ représente la flexion en nombre des noms et des adjectifs. Il décrit deux suffixes, chacun étant associé à une production. Par exemple, l'application de ce paradigme sur le nom commun « ballon » donne deux résultats :

● La forme « ballon », qui reste identique à l'entrée lexicale, à laquelle on associe les propriétés linguistiques de genre masculin et de nombre singulier exprimés via le code « m+s ».

● La forme « ballons » est formée après l'addition du suffixe « s ». Elle est associée aux codes « m+p » qui indiquent, comme cela vient d'être dit, le genre et le nombre du nom « masculin-pluriel ».

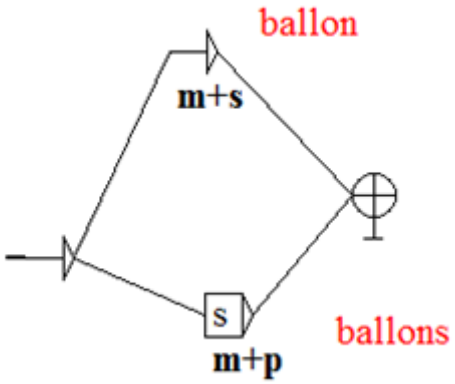


Figure 2. Graphe du paradigme « Ballon »

Une règle de dérivation est la modélisation des paradigmes issus de la morphologie dérivationnelle qui décrit les liens entre les formes de lemmes différents. En décrivant les affixes dérivationnels, une règle de dérivation peut établir des liens entre deux mots de même nature grammaticale ou de deux natures grammaticales différentes ; pour le français, par exemple, (i) les verbes manifester/remanifester ou (ii) un verbe et un nom comme restructuration/restructurer.

Avec NooJ, les mêmes formalismes et outils, qui permettent la description de la flexion, sont utilisés pour décrire la dérivation. A titre d'exemple le paradigme suivant « Ré » permet de dériver des verbes. Il permet de générer le verbe « réapprendre » à partir du verbe « apprendre ». La forme produite est associée à la partie du discours « Verbe » exprimé par le code « V ».

```
Ré = <LW>ré/V; # apprendre => réapprendre
```

La Figure 3 illustre un autre exemple d'une modélisation d'un paradigme dérivationnel en utilisant un graphe. Ce graphe permet de reconnaître tous les verbes qui sont dérivés d'un autre verbe et commencent par « re » comme dans l'exemple déjà cité manifester/remanifester. Il récupère les informations syntaxiques et flexionnelles liées au verbe pour annoter le verbe généré. Ainsi pour « remangeront », par exemple, l'analyseur reconnait « mangeront » comme « verbe », son nombre est à la 3ème personne du pluriel, son mode est l'indicatif, son temps est le futur et son lemme est « manger ». Il transmettra toutes

ces informations au terme produit « remangeront » qui sera donc annoté comme verbe à la 3ème personne du pluriel conjugué au futur de l'indicatif et ayant comme lemme « remanger ».

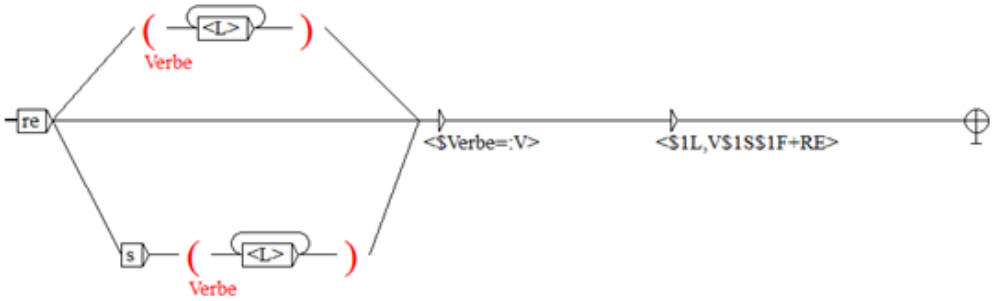


Figure 3. Graphe reconnaît les verbes en « re »

Nous notons également l'intérêt de décrire les entrées composées qui représentent une part importante du lexique des langues et qui ont une forte couverture avec un nombre d'occurrences non-négligeable dans le texte (Gross, 1986).

Finalement, l'analyse lexicale permet de regrouper les formes ayant une caractéristique commune en ajoutant à chaque entrée lexicale diverses informations linguistiques telle la partie de discours, des informations de flexion comme le genre et le nombre pour les noms et comme la personne, le genre, le nombre, le mode et la voix pour les verbes, et des traits sémantiques pour caractériser leur signification.

3.2. Analyse syntaxique

La syntaxe vise à étudier les structures des séquences linéaires des éléments du vocabulaire, comme les syntagmes ou les phrases, qui forment un énoncé. En effet, elle permet de fournir une description formelle de l'ordre des mots et de définir les règles qui gèrent les relations de combinaison et de dépendance des ALU et des séquences d'ALU au sein d'une phrase. Pour qu'une combinaison d'une séquence d'ALU soit syntaxiquement correcte, elle doit respecter diverses contraintes qui « correspondent à des propriétés sélectionnelles ou positionnelles » (Yvon, 2011). Selon Chomsky, on appelle « grammaire » une description de ces propriétés qui permettent un jugement de la grammaticalité d'une séquence d'ALU.

Figure 4. Le test de la grammaticalité

Dans un contexte TAL, la grammaire n'est pas une description normative qui doit être respectée à la production d'un texte mais une description des phénomènes syntaxiques et/ou syntactico-sémantiques utilisés par une communauté. Cette description permet de distinguer les caractéristiques et les spécificités d'une langue. De plus, la grammaticalité est une condition non suffisante et non nécessaire à la compréhension. A l'oral, par exemple, des phrases agrammaticales sont bien comprises par les locuteurs.

3.2.1. Analyse syntaxique locale

L'analyse syntaxique consiste à décomposer une phrase en constituants et à identifier les relations de dépendances qui existent entre eux. Les constituants peuvent être soit des ALU soit des suites d'ALU. En effet, certains courants linguistiques considèrent que l'ALU est une unité syntaxique et les relations syntaxiques se limitent aux dépendances entre les ALU d'une phrase. D'autres théories linguistiques définissent des constituants intermédiaires entre l'ALU et la phrase comme les syntagmes. Un syntagme est une séquence des éléments du vocabulaire à laquelle on associe une catégorie syntaxique. Il possède un ou plusieurs noyaux qui peuvent être des ALU ou des séquences d'ALU. Le noyau transmet au syntagme sa catégorie et sa fonction syntaxique. Kurdi (2017) décrit plusieurs tests linguistiques qui permettent l'identification d'un syntagme comme la commutation, l'ellipse, la coordination et les déplacements.

Steven Abney adopte les syntagmes élémentaires non récursifs appelés *chunks* ou syntagmes noyaux comme unité d'analyse syntaxique (Abney, 1991a). En effet, le *chunk* est la plus petite unité de syntagme à laquelle on attribue une catégorie syntaxique et elle ne pourra donc pas contenir d'autres syntagmes. Notons que certains courants linguistiques définissent d'autres niveaux plus hauts que le syntagme et inférieurs à la phrase comme la proposition pour les phrases complexes, et la phrase syntaxique lorsque la phrase est composée de plusieurs phrases syntaxiques reliées entre elles par une juxtaposition, une coordination ou une subordination.

Les différents constituants comme les ALU, les syntagmes ou les *chunks* peuvent faire l'objet d'un niveau d'analyse appelé analyse syntaxique locale. La syntaxe locale s'intéresse donc à la reconnaissance d'une ou des séquences d'ALU relativement petites et figées. A titre d'exemple d'analyses syntaxiques locales, nous citons la levée d'ambiguïté pour les ALU, la reconnaissance des entités nommées, des groupes nominaux et des expressions semi-figées pour les constituants intermédiaires entre l'ALU et la phrase.

Selon l'approche NooJ, l'analyse syntaxique locale est mise en œuvre grâce aux grammaires locales (Gross, 1997) modélisées sous forme d'un ou plusieurs graphes. Ces grammaires locales formalisent un ensemble de règles non-contextuelles définies par un expert ou observées à partir d'un corpus. Elles décrivent les phénomènes linguistiques et les variations terminologiques qui peuvent être traités sans une analyse des structures syntaxiques des phrases ou des énoncés. A titre d'exemple la figure 5 montre une grammaire locale pour la reconnaissance des séquences d'ALU qui constituent des dates, de type « le lundi 2 mars ». Pour catégoriser les séquences reconnues, les grammaires locales peuvent ajouter des annotations à la TAS. Notons qu'il est possible qu'une grammaire locale produise des filtres qui peuvent être utilisés pour supprimer de la TAS des annotations. Cette opération est utile pour le développement des grammaires de désambiguïsation automatique qui permettent la suppression des annotations incompatibles comme, par exemple, pour la levée d'ambiguïté lexicale des mots grammaticaux.

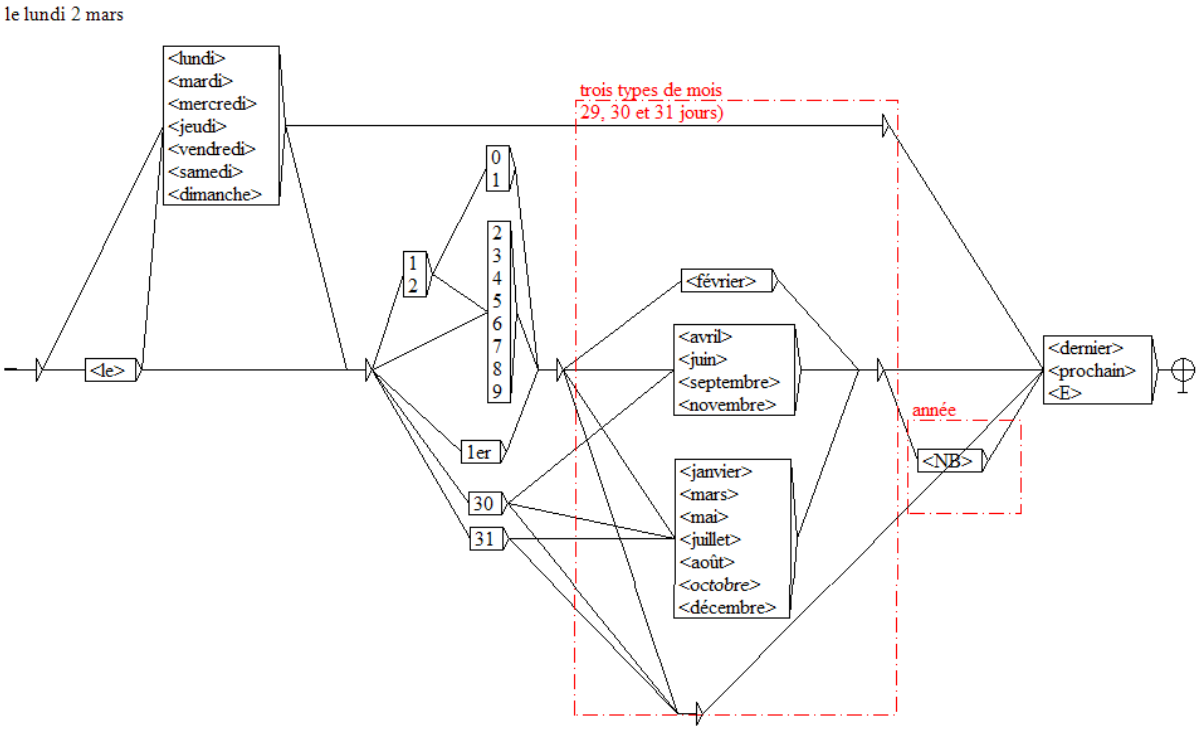


Figure 5. Grammaire permet la reconnaissance des dates de type « le lundi 2 mars »

Pour conclure, grâce à la disponibilité de corpus numériques volumineux, l'analyse syntaxique locale présente un grand intérêt pour les applications d'analyses des données langagières. En effet, elle permet de déterminer d'une façon précise et pertinente la catégorisation des séquences du vocabulaire sans faire appel à une analyse complète de toute la phrase ou de tout le contexte c'est à dire sans passer par une analyse syntaxique structurelle.

3.2.2. Analyse syntaxique structurelle

Dans la plupart des travaux en TAL, l'unité syntaxique est la « phrase » qui est définie comme une suite d'unités lexicales que l'on peut soumettre au test de grammaticalité. Elle peut être une idée ou une pensée qui raconte une vérité du monde réel ou imaginaire. Elle est considérée comme l'élément de base de l'analyse du sens d'où l'importance d'analyser chaque phrase, en déterminant sa structure. L'analyse syntaxique structurelle consiste donc à associer à la phrase segmentée en une suite d'unités lexicales, « une représentation des groupements structurels entre ces unités ainsi que des relations fonctionnelles qui unissent les groupes d'unités » (Fuchs & Victorri, 1993).

Les théories d'analyse syntaxique ont donc pour objectif d'associer à chaque énoncé sa structure de constituants. En effet, les grammaires syntagmatiques associent les unités pour construire les constituants qui s'emboîtent les uns des autres afin de créer une organisation syntagmatique. Cette organisation, qui a habituellement une structure hiérarchique, est visualisée par une représentation arborescente. L'arbre syntaxique, qui est donc produit par l'analyse syntaxique structurelle, contient des informations de structures, qui sont attachées aux nœuds et aux branches, et des unités lexicales qui sont attachées aux feuilles. La figure 6 montre un arbre syntagmatique qui représente la structure de la phrase « Cette contenance exclut toute arrière-pensée ».

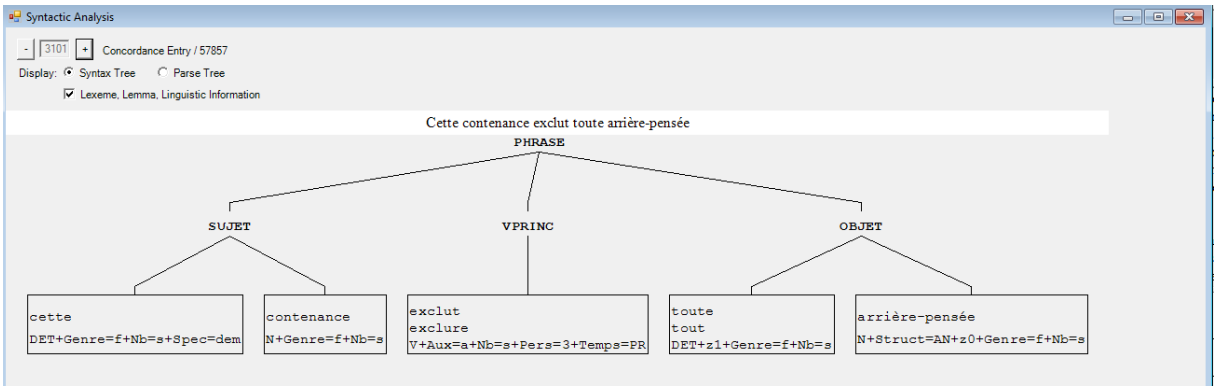


Figure 6. Arbre syntagmatique de la phrase « Cette contenance exclut toute arrière-pensée »

Dans cette figure, l'arbre syntagmatique est représenté de haut en bas :

- La racine de l'arbre est le nœud tout en haut étiqueté par le symbole PHRASE ;
- Les feuilles de l'arbre sont tout en bas et représentent les mots de la phrase ;
- La phrase se décompose en trois constituants : un SUJET (cette contenance) suivi du verbe principal VERBE (exclut) et d'un OBJET ;
- Cet OBJET est lui-même constitué des mots « toute » et « arrière-pensée » ;

Plusieurs arbres syntaxiques peuvent être associés à la même phrase, dans ce cas on parle de l'ambiguïté syntaxique. Ce phénomène a été bien mis en valeur par Chomsky[4] qui a proposé quatre analyses syntaxiques possibles pour la phrase « *Time flies like an arrow* ». Cet exemple de Chomsky nous éclaire sur deux types d'ambiguïtés syntaxiques :

a) L'ambiguïté liée à une analyse lexicale et/ou une analyse syntaxique locale ; lorsque, par exemple, plusieurs descriptions morphosyntaxiques sont attribuées à une seule ALU, plusieurs interprétations syntaxiques deviennent possibles pour l'analyse structurelle d'une phrase. La figure 7 montre que la phrase « *Time flies like an arrow* » est associée à quatre arbres syntaxiques différents à cause de l'appartenance des ALU à plusieurs catégories grammaticales. En effet, les mots « *time* » et « *fly* » peuvent être un verbe ou un nom et le mot « *like* » peut être un verbe ou adverbe.

b) L'ambiguïté liée à une interprétation sémantique ; la figure 7 illustre également deux arbres syntaxiques différents bien qu'ils contiennent la même succession de catégories grammaticales. Le troisième arbre contient un nom attaché à un groupe nominal et un groupe prépositionnel attaché à un groupe adjectival tandis que, pour le quatrième arbre, le nom et le groupe prépositionnel sont attachés à un groupe adjectival.

[4] La phrase « *Time flies like an arrow; fruit flies like a banana*» est apparue dans un film du comédien americain *Groucho Marx*. Elle est largement utilisée en linguistique anglaise pour illustrer des exemples d'ambiguïtés lexicales et syntaxiques. Chomsky a utilisé la première proposition « *Time flies like an arrow*» pour mettre en valeur l'ambigüité syntaxique en générant les arbres syntaxiques potentiels (Johnson & Erneling, 1997).

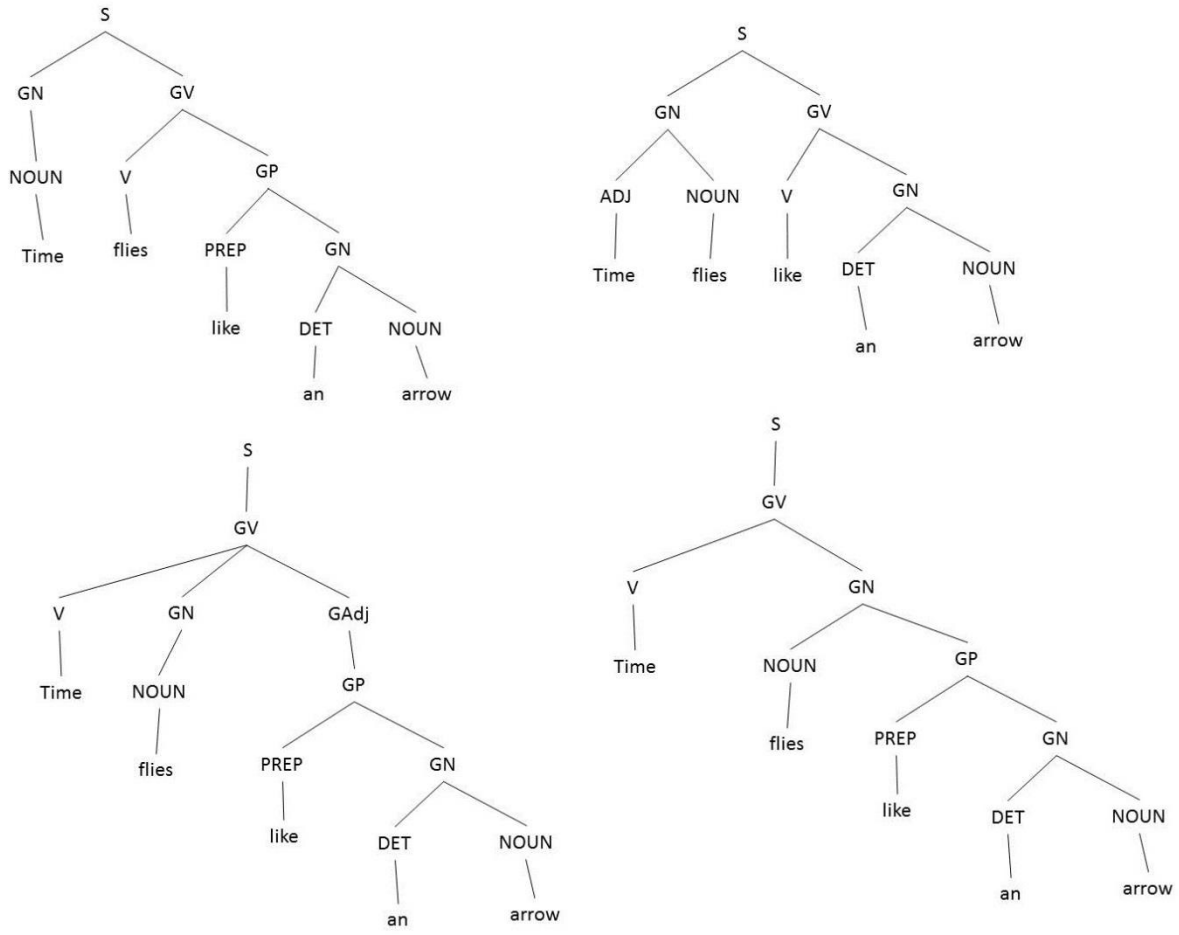


Figure 7. Quatre arbres syntaxiques associés à la même phrase

En opposition aux grammaires syntagmatiques, la grammaire de dépendance considère l'ALU comme unité syntaxique et nie l'existence des constituants supérieurs comme les syntagmes. Fondée par Lucien Tesnière (1959), cette théorie est de plus en plus répandue pour une analyse syntaxique structurelle fiable et robuste. Elle a été reprise et développée par plusieurs linguistes (Sleator et Temperley, 1991), (Hudson, 2000). L'un de ses nombreux avantages est qu'elle permet d'établir les relations qui pourraient exister entre le niveau syntaxique et d'autres niveaux d'analyse comme la sémantique. La grammaire de dépendance décrit donc la syntaxe d'une phrase à travers les relations existantes entre ALU. Ces relations sont représentées d'une façon arborescente par un graphe appelé « arbre de dépendance » ou « stemma ». Les arbres de dépendance diffèrent selon le formalisme de dépendance adopté. L'implémentation de l'*universal dependancies* (De Marneffe & Manning, 2008), (Nivre et al., 2016) est la plus utilisée en TAL avec des modules développés pour environ 60 langues comme le français (Gerdes & Kahane, 2017), l'anglais (De Marneffe et al., 2006), l'arabe (Taji et al., 2017) et le perse (Seraj et al., 2017).

L'approche NooJ considère l'analyse syntaxique structurelle comme une étape postérieure à l'analyse lexicale et à l'analyse syntaxique locale. Par conséquent, elle est appliquée sur tous les types des unités atomique de vocabulaire (ALU) enregistrées dans la structure d'annotation du texte (TAS) à savoir morphèmes, mots simples, mots composés et expressions discontinues. En effet, les types de grammaires et analyseurs intégrés à NooJ permettent le développement des grammaires syntaxiques qui produisent des annotations qui représentent des informations de structure. Ces grammaires syntaxiques structurelles sont des grammaires hors contextes et contextuelles, qui permettent d'associer plusieurs annotations à une séquences d'ALU. La figure ci-dessous montre un exemple d'une grammaire syntaxique structurelle qui reconnaît des phrases simples transitives.

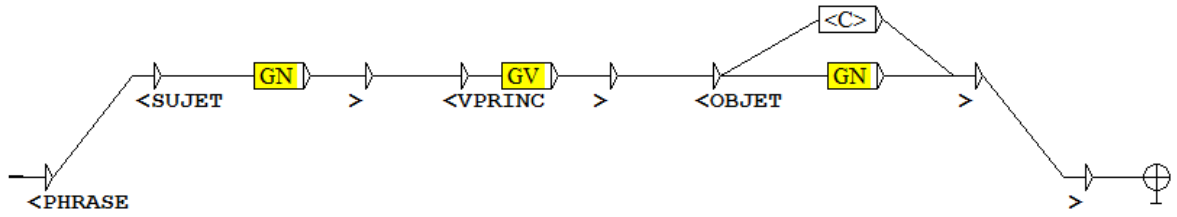


Figure 8. Grammaire syntaxique de reconnaissance des phrases simples transitives

Les annotations structurées produites sont indépendantes de l'organisation de la grammaire et elles permettent la génération d'arbres syntaxiques qui sont une représentation des structures de la phrase, ainsi que des arbres de dérivation qui sont une représentation de la structure de la grammaire. Elles sont rangées sur plusieurs niveaux dans la TAS et cette dernière les distingue des annotations lexicales même si elles sont stockées dans des niveaux parallèles, comme montre la figure 9. En effet, cette distinction, prise en compte techniquement par la TAS et par les analyseurs de NooJ, est en réalité une distinction sémantique qui différencie les annotations lexicales, dont la présence dans des niveaux parallèles souligne les ambigüités, des annotations syntaxiques qui, bien qu'elles soient rangées parallèlement, sont considérées comme des annotations valides.

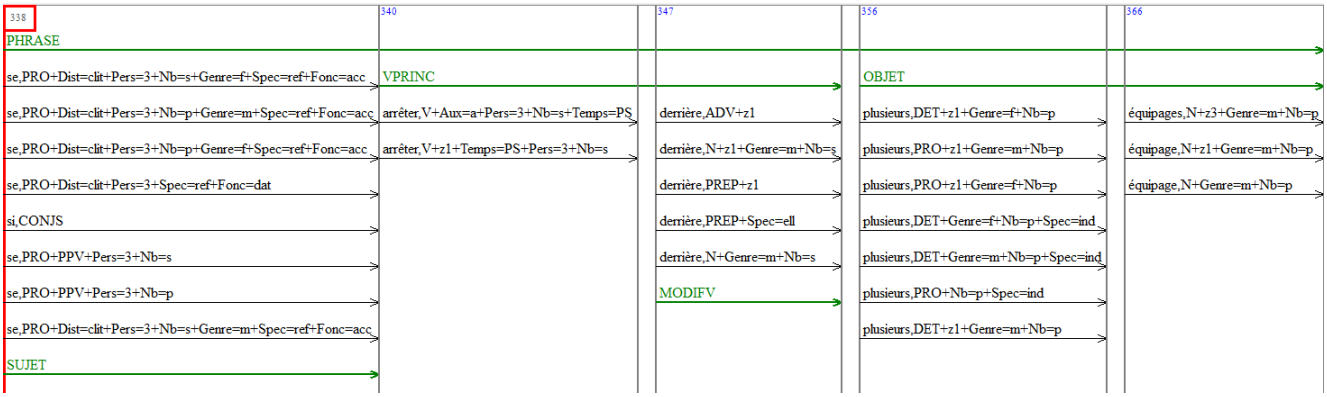


Figure 9. Les annotations syntaxiques dans la TAS

Bien qu'elle permette en premier lieu une description de la structure des phrases, l'analyse syntaxique structurelle peut aussi être utilisée comme une méthode de désambiguïsation lexicale. En effet, si des annotations lexicales présentes dans la TAS ne figurent pas dans un des arbres syntaxiques produits suite à une analyse syntaxique structurelle, on peut considérer que ce sont des hypothèses lexicales incompatibles et les supprimer de la TAS. Par exemple, dans le cas où une phrase est représentée par un seul arbre syntaxique, on peut ne garder dans la TAS que les annotations des ALU présentes dans l'arbre syntaxique.

3.3. Analyse sémantique

Tout énoncé en langue naturelle est composé des séquences d'unités syntaxiques appelées « proposition » ou « phrase »[5] qui est considérée comme l'élément de base de l'analyse du sens. La signification d'une phrase fait l'objet d'un niveau d'analyse linguistique : la sémantique. L'analyse sémantique vise donc à étudier le « sens » associé à une unité ou à une suite d'unités lexicales. Elle consiste en une procédure de deux étapes : l'analyse sémantique locale qui s'intéresse à étudier le sens des unités lexicales et l'analyse sémantique propositionnelle qui cherche à identifier le sens véhiculer par une séquence d'unités lexicales.

[5] La notion de « phrase » fait toujours débat. Elle est souvent définie par des critères formels comme les signes de ponctuation. Nous constatons que les sémanticiens préfèrent le terme « proposition ».

3.3.1. Analyse sémantique locale

La sémantique locale est l'étude du sens des unités du vocabulaire d'une langue. Elle cherche à associer un sens à une unité lexicale monosémique et plusieurs sens à une unité lexicale polysémique. Cette définition nous amène à poser cette question : comment décrit-on le sens d'une unité lexicale ? Pour représenter le « sens », Ferdinand de Saussure distingue entre signifiant qui désigne la représentation mentale de la forme de signe linguistique et signifié qui désigne la représentation mentale du concept associé au signe. Cette représentation est illustrée par le schéma de la figure 10 qui montre la forme « cheval » et le concept exprimé par l'image d'un type de cheval.

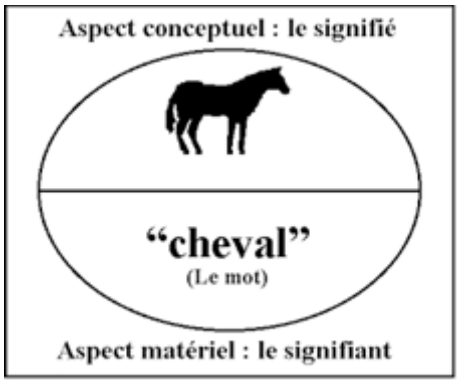


Figure 10. Exemple d'un signifiant et d'un signifié

A cette dualité signifiant/signifié, Gottlob Frege[6] ajoute la notion de référence ou la dénotation qui décrit la portion de réalité que le signifiant désigne. En effet, pour Frege, ce qu'on appelle « sens » n'est que la fonction mentale qui permet, à partir d'un signifiant, de retrouver sa dénotation. La relation entre signifiant, signifié et référence forment le triangle sémiotique qui figure dans le schéma 11.

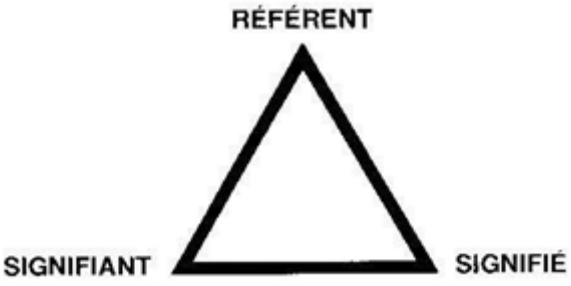


Figure 11. Triangle sémiotique

[6] Friedrich Ludwig Gottlob Frege est un mathématicien, logicien et philosophe allemand. La distinction entre sens et dénotation est considérée comme l'une de ses importantes contributions en philosophie du langage.

Traditionnellement, on représente le sens d'une unité lexicale en utilisant d'autres unités lexicales. Par exemple, les dictionnaires éditoriaux décrivent le sens associé à une unité lexicale en utilisant d'autres unités lexicales. Depuis l'élaboration de plusieurs théories sur la sémantique, nombreuses méthodes de représentation sémantique d'une unité lexicale ont vu le jour. La plateforme NooJ dispose de formalismes et d'outils comme « les dictionnaires électroniques » qui permettent la mise en place de plusieurs de ces méthodes. Un dictionnaire électronique est un formalisme permettant une description exhaustive, riche et précise du vocabulaire d'une langue. En effet, chaque unité lexicale est représentée par une entrée, à laquelle on associe des informations linguistiques mentionnant explicitement toutes les propriétés orthographiques, morphologiques, syntaxiques et sémantiques. Un dictionnaire électronique permet donc d'établir une représentation de sens en se basant sur deux approches différentes :

- A l'aide des « primitives sémantiques ». Il s'agit de décomposer un sens d'une ALU en unités de sens élémentaire. Ces dernières souvent appelées « sèmes », correspondent aux traits sémantiques minimaux qui, sont attribuées à une unité lexicale. Les unités lexicales partageant des « sèmes » sont considérées proches sémantiquement.

- A l'aide d'un « réseau de sens ». Cette méthode remonte à Platon qui a définit la notion d'archétype. C'est un modèle général représentatif d'un sujet ou d'une catégorie. Il s'agit de définir des archétypes. Chaque unité lexicale qui a une proximité avec un archétype sera associée à ce dernier en utilisant un trait sémantique. Une catégorie peut être la sous-catégorie d'une autre catégorie. Par conséquent, comme le montre la figure 12, l'ensemble des catégories définies peuvent avoir une organisation hiérarchique qui permet une description plus riche et détaillée du sens de l'unité lexicale.

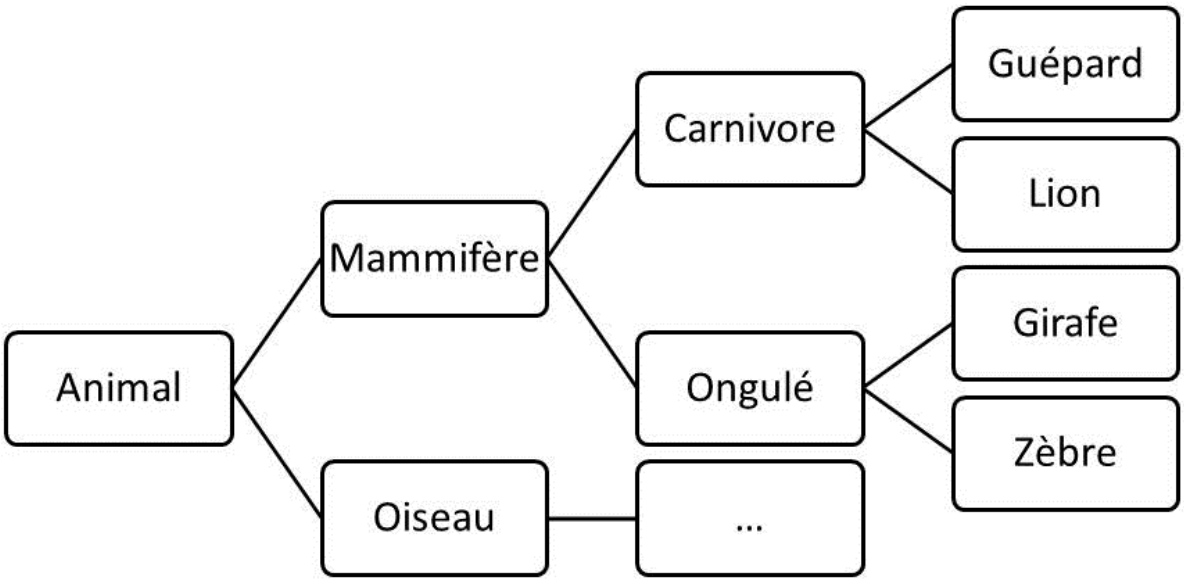


Figure 12. Taxonomie d'un ensemble de catégories

3.3.2. Analyse sémantique propositionnelle

Après une représentation sémantique des unités lexicales, ces dernières vont être agencées intelligiblement afin de permettre la construction du « sens » d'une phrase. L'étude du « sens » d'une séquence d'unités lexicales hors contexte est l'objectif de l'analyse sémantique propositionnelle. Cette analyse, qui a un lien étroit avec la logique, permet la production des prédicats logiques qui peut être écrit sous forme des formules logiques comme *[NOT(aimer(Luc, Léa))]*.

L'approche NooJ permet la mise en place d'une analyse propositionnelle en utilisant des grammaires transformationnelles qui permettent la construction de représentations logiques du sens d'une phrase. Inspirée de la théorie compétence-performance de Chomsky qui démontre qu'une structure de surface qui correspond à la performance n'est qu'une transformation d'une structure profonde, la grammaire transformationnelle permet, entre phrases qui partagent le même matériau lexical et un invariant de sens, d'identifier si une phrase est ou non la transformation d'une autre, telle une phrase et sa forme négative ou une phrase et sa forme passive. Cette grammaire est donc capable de produire toutes les phrases transformées à partir d'une phrase donnée. De la même manière, une grammaire transformationnelle peut générer la représentation logique d'une phrase.

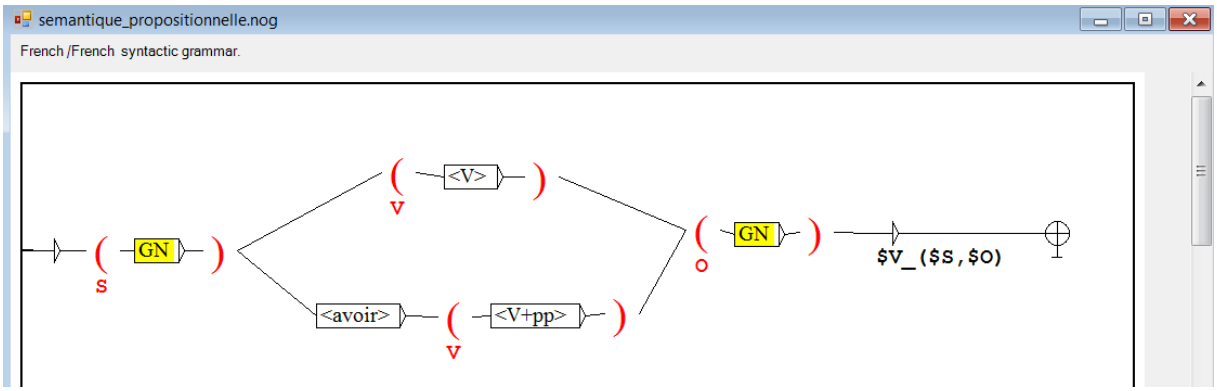


Figure 13. Analyse sémantique d'une phrase à deux arguments

La grammaire transformationnelle de la figure ci-dessus permet, à partir d'une phrase, la production d'une formule logique de la forme *[verbe(sujet,objet)]* dans laquelle le prédicat est le verbe et ses deux arguments sont le sujet et l'objet. Par exemple, pour la phrase «*Pierre salue Jean* », notre grammaire de la figure 13 génère la formule logique suivante *[saluer(Pierre,Jean)]*.

Ces représentations sémantiques produites par les grammaires transformationnelles sont puissantes et elles peuvent être utilisées par toute application qui s'intéresse à la génération automatique des données langagières en reformulant, en résumant ou en traduisant un texte.

3.4. L'analyse pragmatique

Comme nous avons vu, l'analyse sémantique étudie le sens de l'énoncé hors contexte en supposant qu'il soit vrai. Par conséquent, elle ne tient pas compte du locuteur et du contexte d'énonciation, on dit qu'elle traite seulement des aspects vériconditionnels de l'énoncé (Siouffi & Van Raemdonck, 2009).

Au contraire de l'analyse sémantique, l'analyse pragmatique étudie les sens non vériconditionnels de l'énoncé. Selon certains linguistes, on peut dissocier l'analyse sémantique de l'analyse pragmatique puisque cette dernière ne peut en aucun cas influencer les résultats des analyseurs sémantiques. En effet, les informations existantes dans l'énoncé et les informations sur le locuteur et sur le contexte sont utilisées dans le cadre d'une analyse pragmatique dans le but d'élaborer un modèle capable de poser des hypothèses sur l'intention de locuteur.

Généralement, on trouve ces modèles pragmatiques dans des applications de raisonnements à base de cas qui essaient d'identifier l'intention du locuteur afin de lui associer un sous cas du système. Ces systèmes sont orientés vers la résolution d'une tâche précise dans un domaine bien déterminé telles que les applications de gestion de relation

client comme l'analyse des feedback client, la gestion de dialogues, le résumé automatique etc.

Les techniques de l'analyse pragmatique ne sont pas encore stabilisées. On parle généralement des programmes qui reposent sur des opérations logiques permettant de respecter le « principe de coopération »[7] définit par Paul Grice. Il s'agit de vérifier si le propos du locuteur ou auteur est conforme à la situation et au contexte de communication.

4. Les approches de TAL

Dans cette section, nous présentons les différentes approches de TAL à savoir *(i)* l'approche statistique qui profite de la disponibilité des corpus et des ressources linguistiques et du progrès réalisés en intelligence artificielle plus précisément en apprentissage automatique et *(ii)* l'approche symbolique qui permet une description détaillée et précise des différents phénomènes linguistiques.

4.1. Le TAL statistique

L'approche statistique en TAL est de plus en plus utilisée ces dernières années. Elle est rendue accessible grâce à la disponibilité de volumineux corpus annotés dans diverses langues. Par la suite, nous exposerons les méthodes par apprentissage qui permettent l'analyse et l'annotation d'un texte en langue naturelle et les méthodes d'analyse textométrique qui présentent une approche intéressante pour l'exploration des corpus volumineux.

4.1.1. Les méthodes par apprentissage

Les méthodes par apprentissage sont issues principalement de l'apprentissage automatique, qui est une discipline de l'intelligence artificielle. L'apprentissage automatique étudie les algorithmes qui permettent aux programmes de s'améliorer automatiquement par expérience à partir des données (Mitchell, 1997). Les techniques d'apprentissage automatique peuvent être divisées en deux types : supervisé et non supervisé. L'apprentissage supervisé permet de construire des règles de prédictions à partir des données d'entraînement constituées à la fois des classes prédéterminées et des exemples connus. Le système apprend donc à

[7] Paul Grice (1979) a défini le principe de coopération comme suit : « *Que votre contribution à la conversation soit, au moment où elle intervient, telle que le requiert l'objectif ou la direction acceptée de l'échange verbal dans lequel vous êtes engagé.* » (Traduction française de Wilson et Sperber, 1979)

classer les données selon un modèle prédéfini. Par exemple, à partir d'un corpus annoté manuellement, un système d'annotation morphosyntaxique apprend à attribuer à une forme une partie de discours. Les méthodes d'apprentissage non supervisé (*clustering*) ne disposent pas des classes prédéterminées et cherchent, à partir des exemples, à regrouper les données afin d'identifier des classes. Nous commencerons cette section par présenter les débuts de la théorie de l'information ainsi que quelques notions de probabilité. Ensuite nous ferons une description détaillée de la mise en œuvre d'un modèle de langue qui montre comment les algorithmes d'apprentissage automatique permettent l'analyse d'une séquence d'ALU d'une langue naturelle. Enfin, nous exposerons les limites de ces méthodes.

4.1.1.1. Théorie de l'information et probabilité

En 1930, dans le but d'analyser la distribution des formes du roman « Ulysse »[8] de James Joyce, George Kingsley Zipf a classé les formes distinctes de texte par ordre décroissant selon leur nombre d'occurrences. Il a remarqué que les cent premières formes les plus fréquentes correspondent à la plus grande partie du texte et que le reste des formes n'apparaissent que très peu. Il a observé également que la fréquence d'occurrence d'une forme était inversement proportionnelle à son rang. Cette observation empirique appelée la loi de Zipf[9] est caractérisée par la formule suivante :

$\mathrm{f} \times \mathrm{r} = \mathrm{k}$ où *f* est le nombre d'occurrence, *r* est le rang et *k* est une constante.

L'application de la loi de Zipf en pratique donne des résultats approximatifs. De ce fait, en prenant compte l'existence de plusieurs distributions de probabilités, la loi de Zipf a pris une nouvelle forme, telle que la loi de Mandelbort.

$f(n) = k/(a + bn)^c$ où *a*, *b* et *c* sont des constantes.

Ces observations ont démontré qu'il existe certaines régularités dans l'utilisation des formes d'un texte qui peut faire l'objet d'une étude statistique et probabiliste.

On appelle variable aléatoire, une variable dont les valeurs possibles sont distribuées selon une loi de probabilité donnée. Par exemple, pour calculer la probabilité de lancement

---

[8] « Ulysse » est un célèbre roman de l'auteur irlandais James Joyce sorti sous forme de feuilleton dans un magazine américain entre mars 1918 et décembre 1920, avant d'être publié dans son intégralité le 2 février 1922 à Paris.

[9] La loi de Zipf a d'abord été formulée par Jean-Baptiste Estoup et a été par la suite démontrée à partir de formules de Shannon par Benoît Mandelbrot.

d'une pièce de monnaie, la variable aléatoire vaut 1 si la pièce indique face et 0 si la pièce indique pile.

Une estimation d'une distribution de probabilités est possible à partir des fréquences des formes d'un texte. Cette distribution appelée « maximum de vraisemblance » permet de calculer la probabilité d'une forme du texte tirée au hasard.

$P(w) = count\,(w)/\sum_{w'} count(w')$ Où $P(w)=P(W=w)$ est une distribution de probabilités qui indique avec quelle vraisemblance la variable W prend la valeur de la forme w.

Le « maximum de vraisemblance » ne permet pas de prédire la probabilité d'apparition d'une forme sur la base des observations des variables passées (à partir d'un corpus d'apprentissage par exemple). Pour pallier cette limite, la probabilité conditionnelle permet de calculer si la valeur d'une variable donnée dépend des variables de la séquence qui l'a précédée. En effet, on dit que deux événements E et F sont dépendants si la réalisation de E nous donne des informations sur la réalisation de F. En effet, si la variable aléatoire W_1 correspond au mot w_1 est conditionné par l'apparition de la variable aléatoire W_2 qui correspond au mot w_2, le calcul de la probabilité conditionnelle s'effectue mathématiquement comme suit :

$P(w_2/w_1) = P(w_2,w_1)/P(w_1)$

Si les deux mots w_1 et w_2 sont indépendant, notre expression mathématique devient $P(w_2/w_1) = P(w_2)$. Par exemple, pour l'étiquetage morphosyntaxique de la forme « *il* » de la phrase « *Il aime la lecture* », la probabilité conditionnelle permet de calculer l'attribution de la partie du discours « *pronom* » à la forme « *il* » conditionnée par l'événement : la forme « il » est annotée uniquement comme « *pronom* » dans le corpus d'apprentissage.

Le théorème de Bayes nous permet de calculer la probabilité d'un événement E conditionné par un autre événement F si nous disposons des informations sur la probabilité de F conditionné par E.

$P(E/F) = P(F/E)P(E)/P(F)$

Par exemple, le théorème de Bayes nous permet de calculer la probabilité conditionnelle $P(w_2/w_1)$ ainsi :

$P(w_2/w_1) = P(w_1/w_2)P(w_2)/P(w_1)$

Notons que dans les méthodes par apprentissage en TAL, nous sommes souvent amenés à traiter une séquence de variables aléatoires qui sont dépendantes les unes des autres afin d'analyser des séquences de caractères, de mots ou de phrases d'où l'utilité de la probabilité conditionnelle et du théorème de Bayes.

4.1.1.2. Le modèle de langue

Un modèle de langue est un concept probabiliste utilisé dans plusieurs applications en TAL comme la traduction automatique, la recherche d'information, la reconnaissance automatique de la parole et la classification automatique des textes. Il correspond à la composante qui permet l'analyse d'une séquence d'ALU. En effet, c'est une distribution de probabilité sur des séquences d'ALU qui permet d'estimer la probabilité d'une unité en tenant compte de la séquence dans laquelle elle apparaît.

$$P(w_n|w_1, \ldots, w_{n-1})$$

En effet, la distribution d'une séquence des unités qui apparaissent l'une après l'autre comme $w=w_1w_2w_3...w_n$ sera modélisée par une probabilité dite « conjointe » qui sera formulée mathématiquement ainsi :

$$P(w_1, w_2, w_3, \ldots, w_{n-1}, w_n) = P(w_1)\, P(w_2|w_1)\, P(w_3|w_2) \ldots P(w_n|w_{n-1})$$

Selon la théorie de la probabilité, nous utilisons la « règle de la chaîne » qui permet de décomposer notre probabilité conjointe en un produit de probabilité conditionnelle. En effet, ce qui différencie un modèle de langue d'un autre c'est la modélisation de la probabilité conditionnelle $P(w_k|w_1^{k-1})$.

Par exemple, pour un modèle N-grammes, notre probabilité conditionnelle peut être calculé selon l'équation :

$$P(w_n|w_1^{n-1}) = P(w_1, w_2, w_3, \ldots, w_{n-1}, w_n)/P(w_1, w_2, w_3, \ldots, w_{n-1})$$

En réalité, un modèle N-gramme estime ces probabilités à partir des unités d'ordre N-1. Il peut correspondre à un modèle de Markov d'ordre N ou à un réseau bayésien d'ordre N. Les modèles séquentiels N-grammes sont utilisés par un bon nombre de méthodes par apprentissage en TAL. Ces modèles prennent en considération seulement les N-1 qui précédents l'ALU à analyser comme les *uni-grams* pour N=1, les *bi-grams* pour N=2 et les *tri-grams* pour N=3. En effet, la réduction de données nécessaires à l'analyse des séquences permet d'améliorer la précision et l'efficacité des modules d'apprentissages en évitant le sur-apprentissage.

$$P(w_n|w_1^{n-1}) = P(w_n|w_{n-N+1}^{n-1})$$

Par exemple, pour N=2, l'analyse de la séquence « *le petit garçon aime la lecture* » :

$P(le\ petit\ garçon\ aime\ la\ lecture) = P(le|\#)\ P(petit|le)\ P(garçon|petit)\ P(aime|garçon)\ P(la|aime)\ P(lecture|la)P(\#|lecture)$ avec # comme marqueur de début et fin de phrase.

Suite à plusieurs implémentations (Baker, 1976) en utilisant les chaînes de Markov, un facteur supplémentaire de complexité a été ajouté qui est la variable dite cachée. Ces variables permettent de traiter les ambiguïtés dans le cas où une variable d'une séquence présente plusieurs hypothèses possibles. Par exemple, un mot peut avoir plusieurs descriptions morphosyntaxiques possibles. Ces modèles permettent donc de calculer la probabilité d'un état à partir d'un comportement observé. Cela se fait généralement à l'aide de la formule de Bayes. Mais, puisqu'en TAL, nous analysons des séquences d'ALU comme les formes et les morphèmes, notre formule Bayes s'applique comme suit :

$$P(w_1 \ldots w_n / T_1 \ldots T_n) = P(T_1 \ldots T_n / w_1 \ldots w_n) P(w_1 \ldots w_n) / P(T_1 \ldots T_n)$$

Bien qu'il existe d'autres modèles tels que les modèles à base de classe implémentés en utilisant les algorithmes de classification comme les machines à vecteurs de support (SVM), avec l'émergence des méthodes de l'apprentissage profond *Deep Learning,* le modèle à base de réseaux de neurones présente une vraie alternative aux modèle N-grammes (Hin et al., 2012).

4.1.1.3. Limites des méthodes par apprentissage

Toutes les méthodes par apprentissage s'appuient sur des corpus de référence qui sont annotés manuellement ou semi-automatiquement. Nous constatons que ce travail manuel laborieux peut produire un nombre important d'erreurs d'annotations. C'est un problème bien connu par les chercheurs en TAL (Green & Manning, 2010) qui estiment que la plupart des corpus de références contiennent des erreurs qui nuisent à la qualité et aux performances de l'analyse.

Selon Chomsky, un nombre infini de phrases peut être généré à partir d'un nombre limité d'unités lexicales. Les corpus des références ne peuvent donc pas être exhaustifs et ne peuvent prétendre être représentatifs d'une langue. De ce fait, les méthodes d'apprentissage automatique ne peuvent être fiables que si le texte à analyser est en cohérence avec le corpus de référence. Par exemple, si le corpus de référence est constitué d'un ensemble des textes littéraires et le texte à analyser ou le corpus de test est un article journalistique sur l'intelligence artificielle, la performance des algorithmes d'apprentissage automatique risque de ne pas être spectaculaire.

Une étude réalisée par E. Tzourkermann[10] en 1996 sur deux corpus[11] en français extraits du journal Le Monde montre que plus de la moitié des mots ne sont pas ambigus et qu'une grosse part de l'ambiguïté est faite d'un petit nombre des mots fréquents qui sont généralement des mots outils. Si le taux d'ambiguïté pour des langues standardisées comme le français se limite à un nombre fini de mots, pour les langues à traditions orales et les langues anciennes ayant une orthographe non-standardisée, il existe un grand nombre de possibilités pour une même forme. Par conséquent, les corpus de référence ne parviennent pas à contenir une liste finie et exhaustive du vocabulaire standard de la langue et les méthodes par apprentissage automatique risquent d'être inefficaces.

4.1.2. Les méthodes d'analyse textométrique

« Traiter les mots comme des nombres »[12] est le principe de base de tout système d'analyse textométrique. La textométrie est une discipline, dont les contours sont difficiles à établir, riches en méthodes et en approches. Elle permet principalement d'explorer les corpus textuels. Il s'agit de décrire, comparer, classer, analyser des ensembles de textes en utilisant des méthodes statistiques sans réduire le texte à un sac de mots en supprimant les morphèmes grammaticaux comme pour le français « *par* », « *ou* » et « *que* » et sans négliger les informations linguistiques secondaires tels le genre, le nombre, l'usage de la majuscule et de la minuscule, etc.

Les méthodes d'analyse textométrique sont le résultat de l'évolution d'un des courants de la statistique moderne « la statistique multidimensionnelle[13] » et de l'application de la méthode statistique à l'étude des textes « statistique descriptive ». Ces méthodes ne se contentent pas seulement de produire des résultats d'analyses figées, elles proposent de plus des analyses diverses et offrent une grande palette d'outils qui permettent l'élaboration d'un grand nombre de scénarios d'analyse. Implémentées généralement dans des logiciels appelés

[10] Etude citée dans le livre « les linguistiques de corpus » de Benoit Habert, Adeline Nazarenko et André Salem.

[11] Les deux corpus sont constitués à partir des articles du journal *Le Monde* du septembre à octobre 1989 et du janvier 1990. Le premier corpus contient de 94 882 tandis que le deuxième contient 200 182 occurrences.

[12] Citation du Livre "Statistique Textuelle" (Lebart & Salem, 1994)

[13] La statistique multidimensionnelle est une branche de la statistique qui s'intéresse aux méthodes permettant de traiter simultanément un nombre quelconque de variables au-delà de l'étude d'une seule ou de deux variables.

« plateforme d'analyses des données textuelles »[14] tels que *Lexico3*[15] ou *Le Trameur*[16] ou *TXM*[17], certaines méthodes de textométrie peuvent être complémentaires, d'autres peuvent communiquer entre elles dans le cas où les entrées/sorties sont compatibles et d'autres sont interdépendantes.

Afin de guider vers la solution d'un problème, les plateformes d'analyse des données textuelles donnent la possibilité d'effectuer deux démarches différentes : une analyse exploratoire qui permet de décrire et d'explorer le corpus afin d'avoir une idée des différents phénomènes que les données représentent et une analyse confirmatoire qui permet d'émettre des hypothèses afin de confirmer certains phénomènes et/ou d'infirmer des cas particuliers posés par certaines propriétés.

Nous distinguons ainsi deux types de plateformes d'analyse textométrique :

- Plateforme de lexicométrie qui effectue des analyses sur les formes de corpus. Les plus connues sont *Lexico 3* et *Hyperbase*[18].

- Plateforme de textométrie qui effectue des analyses en croisant les formes textes ainsi que les informations linguistiques associées tels que les lemmes et les parties de discours. Les plus répandues sont *Le Trameur* et *TXM*.

Ces différentes plateformes ont comme objectif de compter des éléments (contenus) dans des ensembles (contenants). Les contenus se réalisent sous forme de ressources textuelles à savoir une séquence des caractères organisée en phrase ou en paragraphe tandis que les contenants existent sous la forme des masques ou de calques qu'on peut définir sur les contenus (Fleury, 2007).

Des programmes existent dans les plateformes d'analyse textométrique permettant, en utilisant une liste de délimiteurs, de découper le corpus en plusieurs types d'unités

---

[14] « *Plateforme d'analyses des données textuelles* » est un terme utilisé par (Lebart & Salem, 1994) pour désigner les applications qui met en œuvre des méthodes d'analyse textométrique.

[15] Lexico 3 : http://lexi-co.com/

[16] Le Trameur : http://www.tal.univ-paris3.fr/trameur/

[17] TXM : http://textometrie.ens-lyon.fr/

[18] Hyperbase : http://ancilla.unice.fr/

statistiques. Cette segmentation de corpus permet de déterminer des contenants et des contenus afin de permettre des analyses en plusieurs niveaux (comme l'illustre la figure suivante) : le niveau du texte, du paragraphe, de la phrase, d'une séquence de mots et du mot.

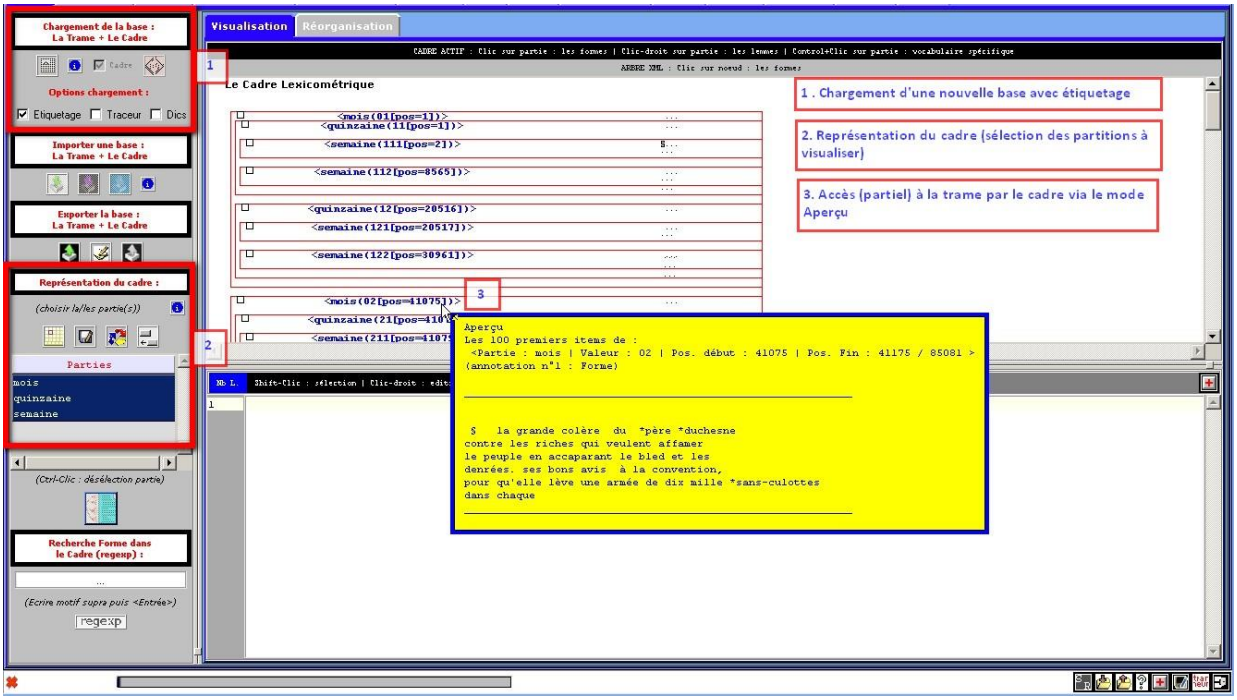


Figure 15. Exemple de découpage de texte en unités statistiques

Par la suite, nous présentons les principales méthodes d'analyse textométrique qui permet d'exploiter les différents contenants et contenus (la figure 15) afin d'effectuer des analyses multiniveaux sur un corpus :

- Un concordancier permet d'afficher des concordances sur divers niveaux d'annotations : forme, partie de discours, lemme, etc.

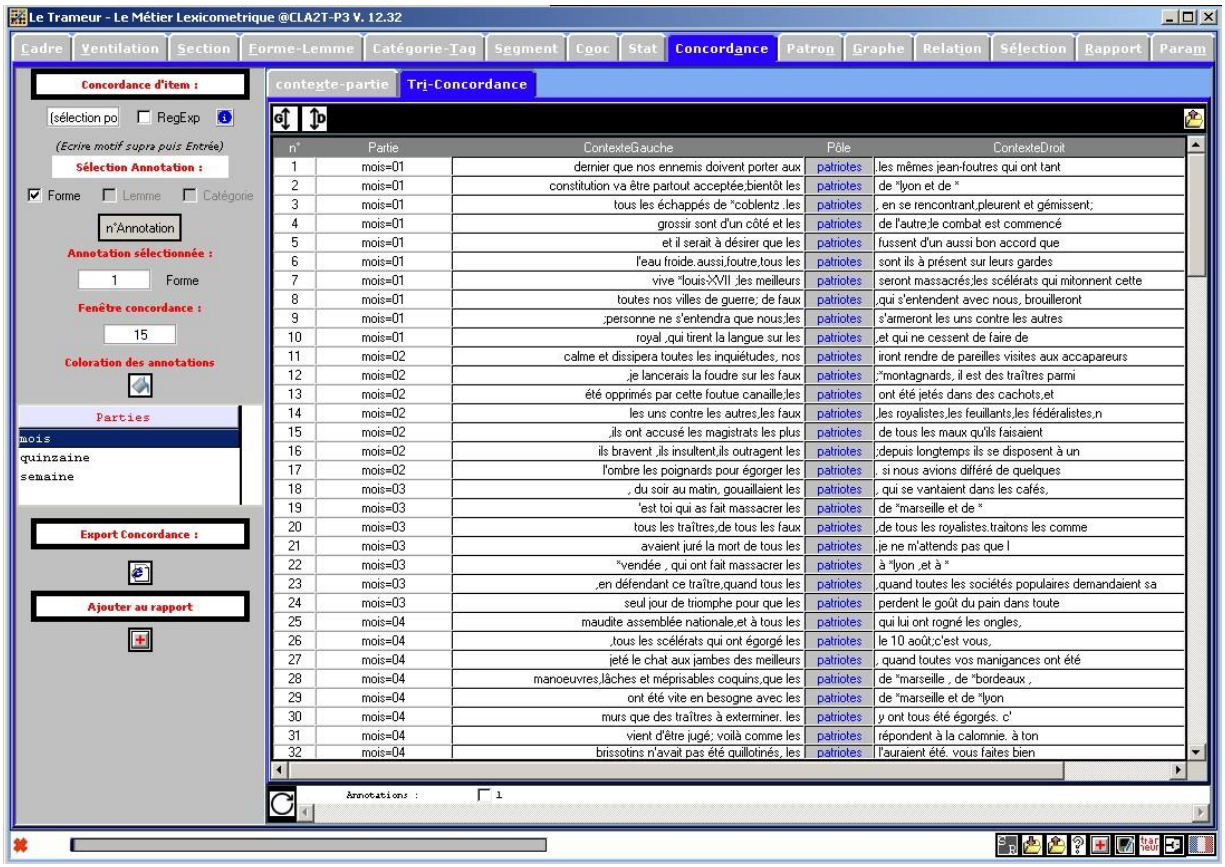


Figure 16. CONCORDANCE, affichage édition + affichage tri

- La ventilation permet d'afficher des graphiques de ventilation d'une sélection d'annotations par partie.

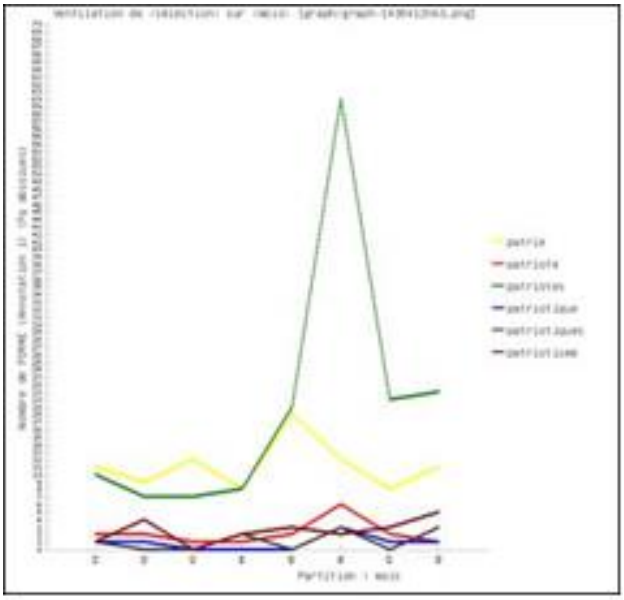


Figure 17. Ventilation d'une sélection d'items

- Les segments répétés permettent de calculer des séquences d'annotations conjointes selon la longueur maximale des segments recherchés et leur fréquence dans le corpus.

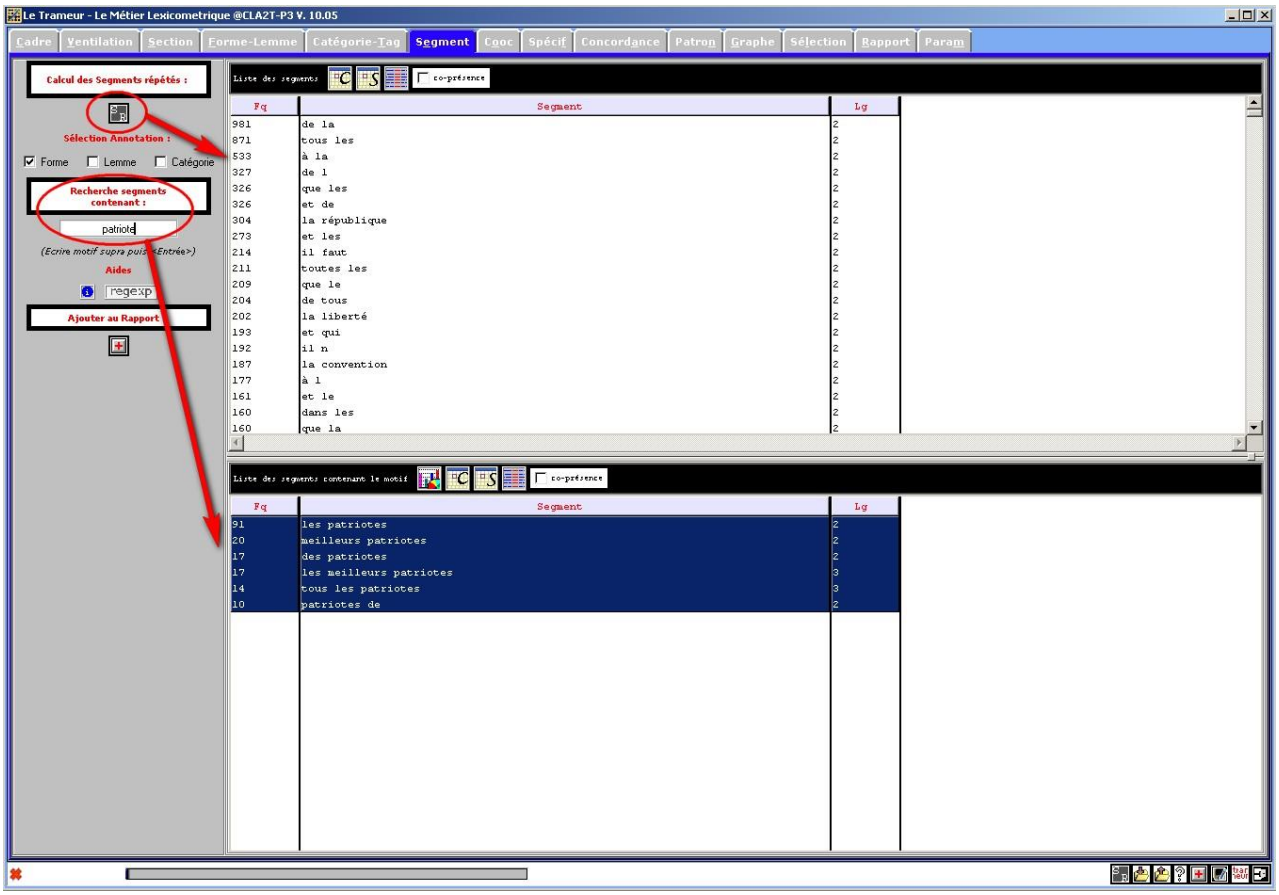


Figure18. Segments répétés

- Les méthodes de calculs de cooccurrences qui détectent l'apparition de deux mots, en même temps et dans le même contexte, et de poly-cooccurrences qui détectent les attractions lexicales au-delà de la cooccurrence binaire.

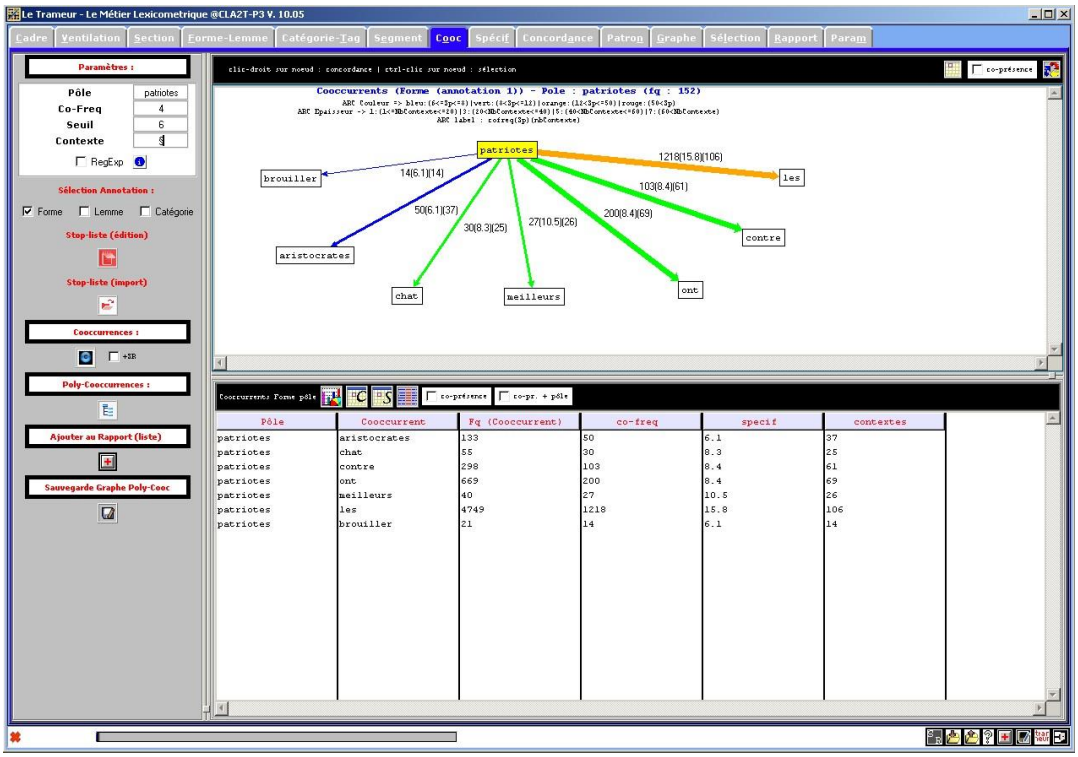


Figure 19. Réseau de cooccurences autour d'un pôle

- Méthodes d'extraction de patrons et graphes de mots qui permettent d'identifier les séquences d'annotations qui répondent au patron recherché.

Figure 20. Extraction de patron (croisement d'annotation)

- L'analyse factorielle des correspondances sur l'ensemble des parties du corpus.

Figure 21. Analyse Factorielle des correspondances (AFC)

4.2. Les méthodes symboliques

Une méthode symbolique consiste à modéliser des connaissances linguistiques qui seront exploitables par un système d'analyse automatique afin de traiter un texte en langue naturelle. Partant de cette définition, il est nécessaire de produire une description formelle et unifiée des différents phénomènes linguistiques. Cette description est l'objectif des grammaires formelles. En effet, Chomsky (1957) a suivi une démarche scientifique en linguistique appelée « la mathématisation en linguistique ». Elle permet de représenter les connaissances linguistiques via un langage symbolique dans le but de les interpréter en utilisant un modèle mathématique. Cette démarche a conduit Chomsky à s'interroger pour savoir : « De quel modèle mathématique a-t-on besoin pour décrire les langues ? ». Au-delà, Chomsky a supposé l'existence d'un modèle « universel » de grammaire qui permette la description de toutes les langues naturelles. La grammaire, considérée comme un ensemble de règles, doit permettre la reconnaissance des phrases d'un langage et la production à volonté de toutes les phrases du langage. La grammaire n'est donc qu'une représentation formelle qui doit être interprétée par un analyseur et elle est ensuite appliquée à un texte en langue naturelle. De ce fait le test de grammaticalité peut être schématisé comme suit.

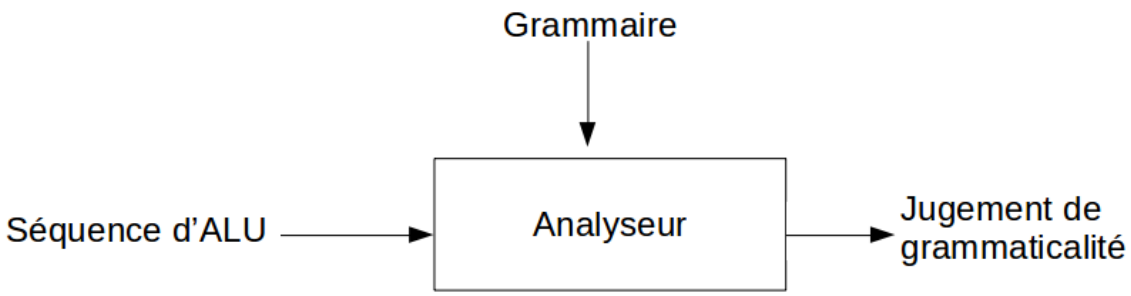


Figure 22. Le test de grammaticalité en intégrant l'analyseur

4.2.1. Les grammaires génératives

Chomsky a développé une grammaire fondée sur la distinction compétence-performance. Elle doit être capable non seulement d'analyser la structure de l'énoncé (performance) mais aussi de formaliser les mécanismes permettant de formuler des énoncés (compétence). Cette grammaire nommée « grammaire générative » est composée d'un ensemble de règles de réécritures. Une règle de réécriture est constituée de deux membres : selon laquelle chaque membre gauche se réécrit en un membre droit. Chaque grammaire générative contient un et un seul symbole de départ qui correspond à l'axiome.

La grammaire générative noté G est définie par quatre d'éléments $G=<V,N,P,S>$ où:

- V est le vocabulaire terminal de G qui correspond aux mots de la langue.
- N est le vocabulaire non-terminal de G qui sert à décrire une langue.
- P correspond aux règles de réécriture.
- S correspond à l'axiome de la grammaire.

Prenons l'exemple de la grammaire générative suivante dont P est le symbole de départ.

P → GN GV
GN → DET ADJ NOM | DET NOM ADJ
GV → V GN
NOM → prince | thème
DET → le | des
ADJ → petit | universel
V → aborder

Le vocabulaire terminal V de cette grammaire est constitué de 7 éléments {prince, thème, le, des, petit, universel, aborder}. L'ensemble d'élément du vocabulaire non-terminal est le suivant {P, GN, GV, DET, ADJ, NOM, V}.

L'exemple *P* illustré ici contient 7 règles de réécritures : la première permet de décomposer la phrase en un groupe nominal suivi d'un groupe verbal. La deuxième règle

analyse le groupe nominal en deux suites : un déterminant suivi d'un adjectif suivi d'un nom ou un déterminant suivi d'un nom suivi d'un adjectif. La troisième règle décrit le groupe verbal comme la succession d'un verbe et d'un groupe nominal. Les règles quatre, cinq, six et sept donnent des exemples respectifs des noms, des déterminants, des adjectifs et d'un verbe.

Par exemple, notre grammaire permet de générer la phrase « Le petit prince aborde des thèmes universels » en appliquant une succession des règles de réécriture comme suit :
P → GN GV → DET ADJ NOM GV→Le ADJ NOM GV→ Le petit NOM GV → Le petit prince GV →Le petit prince V GN→ Le petit prince aborde GN → Le petit prince aborde DET NOM ADJ → Le petit prince aborde DET NOM ADJ → Le petit prince aborde des NOM ADJ → Le petit prince aborde des thèmes ADJ → Le petit prince aborde des thèmes universels.

Cette suite de règles de réécritures s'appelle une dérivation. Elle peut être représentée par l'arbre de dérivation suivante :

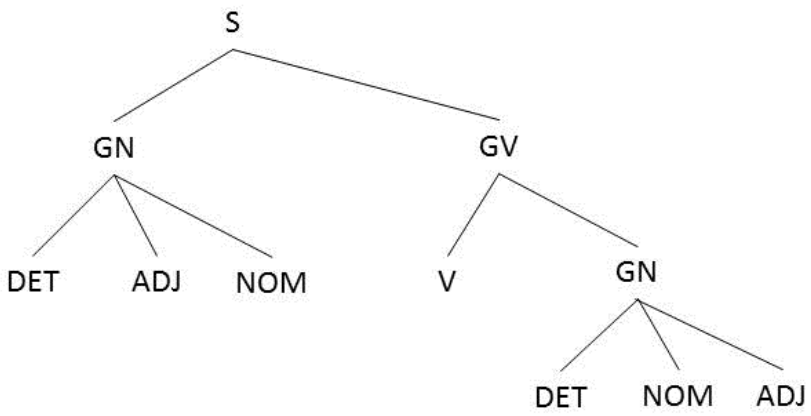


Figure 23. Arbre de dérivation générée par une grammaire

L'une des caractéristiques des langues naturelles est la possibilité de générer un nombre infini de phrases en répétant la même structure plusieurs fois d'une façon imbriqué. Par exemple, si nous supposons le développement d'une grammaire capable d'analyser l'ensemble de ces phrases :

Des fleurs rouges décorent l'assemblée
Des fleurs rouges blanches décorent l'assemblée
Des fleurs rouges blanches jaunes décorent l'assemblée

Les règles de réécritures qui décrivent le syntagme nominal sont :

SN → DET ADJ NOM
SN → DET ADJ ADJ NOM
SN → DET ADJ ADJ ADJ NOM

Mais si nous ajoutons un quatrième adjectif de couleurs à notre exemple, nos règles de réécritures ne permettent pas de reconnaître une phrase comme « Les fleurs rouges bleus jaunes blanches décorent l'assemblée ». La solution à ce problème consiste à utiliser une « règle récursive » afin de reconnaître un nombre indéfini d'adjectifs.

SN → DET NOM SAdj
SAdj → ADJ SAdj
SAdj → ADJ

On dit que le syntagme nominal contient une récursivité à droite. Cette grammaire permet donc de reconnaître tous les syntagmes nominaux composés d'un déterminant, d'un nom et d'un nombre indéfini d'adjectifs. En effet, la première règle décompose le syntagme nominal en déterminant, nom et syntagme adjectival. La deuxième règle est une règle récursive qui décrit le groupe adjectival en adjectif suivi d'un groupe adjectival. La dernière règle est la condition d'arrêt de notre récursivité qui consiste à transformer le groupe adjectival en un adjectif. Cette grammaire peut être mise en place en utilisant la grammaire NooJ de la figure 24.

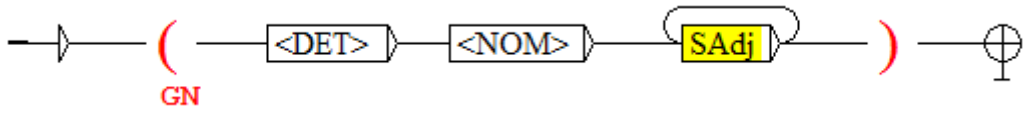


Figure 24. Grammaire de reconnaissance d'un ensemble de groupes nominaux

En guise de conclusion, les règles de réécritures et les arbres de dérivation sont des éléments de base pour la formalisation d'une grammaire d'une langue.

4.2.2. La hiérarchie de Chomsky

Les grammaires formelles décrivent des langages. La portée d'une grammaire est déterminée par les langages engendrés par cette dernière. Selon le degré de complexité des phénomènes linguistiques à décrire, Chomsky définit une typologie composée de quatre types de grammaires formelles. A chaque grammaire correspond une classe de langues. En effet, le type d'une grammaire dépend des règles de réécritures R utilisées par cette dernière. Elles sont classées par valeurs numériques de 3 à 0 où la grammaire de type 3 correspond à la grammaire la plus contrainte et la grammaire de type 0 correspond à la grammaire la moins contrainte de toutes les grammaires.

- Type 3 : Les grammaires rationnelles ou régulières contiennent des règles de réécritures de type $A \rightarrow a$ ou $A \rightarrow a\ B$ avec $A, B \in N$ et $a \in V$. Ces grammaires décrivent les langages rationnels et peuvent être mis en place par des automates à états finis. Le langage typique d'un tel type de grammaire est $a^n b^m$. Bien qu'elles soient puissantes, l'utilité de ces grammaires est limitée et elles sont généralement utilisées pour décrire le lexique d'une langue.

- Type 2 : Les grammaires hors contexte (ou CFG : *Context Free Grammar*) contiennent des règles de réécritures décrites mathématiquement ainsi : $\exists A \in N$ *et* $a \in$ VUN, $A \rightarrow a$. Ces règles sont donc composées du membre gauche qui est un symbole non-terminal et du membre droit qui est à son tour composé d'une suite quelconque des symboles terminaux et non-terminaux. Ces grammaires décrivent les langages algébriques. Le langage typique engendré par des grammaires de type 2 est $a^n b^n$. Elles sont souvent utilisées pour effectuer des analyses syntaxiques locales et des analyses morphologiques. Elles peuvent être modélisées grâce aux réseaux de transition récursive RTR.

- Type 1 : Les grammaires contextuelles comme leur nom l'indique décrivent des langages contextuels. Les règles de réécritures de ce type de grammaire peuvent contenir un contexte dans leur membre droit et un contexte dans leur membre gauche à titre d'exemple PLURIEL GN → PLURIEL DET ADJ NOM. Le Langage typique généré par cette grammaire est de forme $a^n b^n c^n$. Elles peuvent être modélisées en réseaux de transition augmentés RTA et elles sont utilisées pour décrire des phénomènes morphologiques et syntaxiques complexes.

- Type 0 : Les grammaires non restreintes ou générales peuvent décrire tous les langages. Leurs règles de réécritures de type $A \rightarrow a$ *avec* $a \in$ VUN et A $\in$ VUN. Ces règles ne contiennent donc aucune « restriction » c'est-à-dire que n'importe quelle combinaison de symboles terminaux et non-terminaux est possible dans les membres droits et gauches. En effet, la principale spécificité d'une telle grammaire est qu'elle admet des règles dont la composante droite est plus longue que la composante gauche. Les langages engendrés par les grammaires non-restreintes sont reconnus par une machine de Turing.

La figure suivante montre que les quatre grammaires de la hiérarchie de Chomsky sont emboîtées les unes dans les autres. En effet, les grammaires sont de moins en moins

contraintes c'est-à-dire que les grammaires rationnelles sont moins contraintes que les grammaires hors-contextes, elles-mêmes plus contraintes que les grammaires contextuelles, elles-mêmes plus contraintes que les grammaires non restreintes.

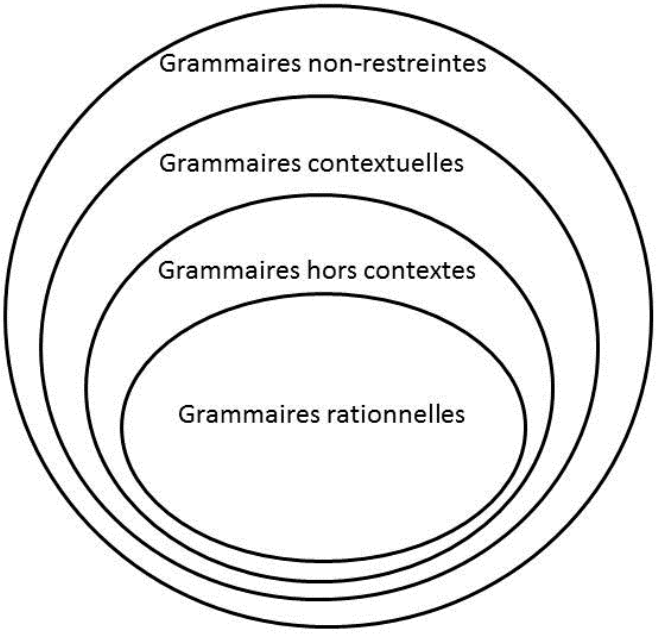


Figure 25. Hiérarchie de Chomsky

Selon Chomsky, les grammaires contextuelles de type 1 sont les mieux adaptées pour engendrer les langues naturelles. Mais ces dernières sont difficiles à implémenter, ce sont donc les grammaires hors contextes de type 2 qui ont été longuement utilisées dans les années 1960 pour décrire les langues naturelles.

4.2.3. Plusieurs formalismes linguistiques

Bien que les grammaires génératives aient permis des résultats importants sur les langages, sur la calculabilité et sur la complexité (Carton, 2008), elles sont très difficilement exploitables pour une description linguistique. Elles permettent de représenter les dépendances hiérarchiques des constituants et de leur ordre. Cependant, elles exigent un travail manuel coûteux et difficile à implémenter. En effet, la description de certains phénomènes linguistiques nécessite un développement lourd pour représenter les traits linguistiques d'un constituant ainsi que les relations d'accord entre les constituants. La représentation des traits linguistiques d'un constituant est assurée par une structure appelée « structure de trait ». On distingue des structures de traits atomiques dont tous les traits ont une valeur simple et les structures de traits complexes dont un trait au moins contient comme valeur une structure de traits (Kurdi, 2017). Les relations entre les structures de traits sont

assurées par l'opération d'unification qui permettent de combiner des ALU et des syntagmes pour représenter des unités syntaxiques supérieures. Appliquée à un ensemble de structures de traits compatibles, l'unification a donc pour but de déterminer une nouvelle structure qui englobe toutes les informations présentes dans l'ensemble des structures de traits.

Le choix de type d'ALU que ce soit morphème, forme, mot composé ou syntagme comme unité syntaxique atomique, la détermination des traits linguistiques utilisés pour décrire les constituants et l'identification des différentes relations qui peuvent exister entre les constituants sont les principales raisons de l'existence de plusieurs formalismes créés pour une description linguistique de la langue. Ces formalismes appelés « formalismes linguistiques » sont le résultat des travaux pluridisciplinaires associant linguistique, logique, mathématique et intelligence artificielle. Nous distinguons des formalismes linguistiques « procéduraux » qui, pour chaque phénomène linguistique, représentent les connaissances linguistiques et définissent les mécanismes d'analyses simultanément et dans la même structure, et les formalismes linguistiques « déclaratifs » qui cherchent à distinguer entre la représentation des connaissances et le raisonnement.

Bien qu'il existe des formalismes linguistiques génériques qui essaient d'établir une description linguistique complète de la langue telle la grammaire *Head Driven Phrase Structure Grammar* (HPSG)[19], la plupart des formalismes se contentent de décrire un ou plusieurs langages de la hiérarchie de Chomsky. Par exemple, la grammaire *Xerox Finite State Tools* (XFST)[20] n'engendre que les langages rationnels, la grammaire *Generalised Phrase Structure Grammar* (GPSG)[21] engendre des langages hors contexte et la grammaire *Lexical Functional Grammar* (LFG)[22] engendre des langages contextuels.

Ces formalismes contiennent une représentation des informations linguistiques et une méthode d'analyse. Ils sont conçus pour illustrer des travaux théoriques, par conséquent, ils ne sont centrés ni sur les besoins applicatifs ni sur les expériences des linguistes. En effet, ils demandent un développement des ressources « à la main » et ces ressources sont difficiles à

[19] HPSG (Polar & Sag, 1994)

[20] XFST (Karttunen et al, 1997)

[21] GPSG (Gazdar, 1985)

[22] LFG (Kaplan & Bresnan, 1982)

accumuler, à maintenir et à faire évoluer. Puisque la même grammaire formalise divers phénomènes linguistiques, l'analyse est souvent jugée lente et inefficace pour être utilisée par des applications informatiques. Au point que des formalismes moins lourds et suffisamment puissants ont été développés appelés sous le nom de *« grammaire légèrement sensibles au contexte »*. Ces grammaires engendrent l'ensemble des langages de type 2 et elles ont une portée strictement inférieure à l'ensemble des langages de type 1, tels la grammaire *Tree Adjoining Grammars* (TAG)[23], *Linear Indexed Grammar* (LIG)[24] et *Head Grammar* (HG)[25].

Bien que ces différents formalismes soient conçus sur l'idée d'un formalisme universel, leurs efficacités et leurs puissances sont variables selon les phénomènes linguistiques à traiter. Par exemple, XFST est un formalisme adapté pour traiter les phénomènes morphologiques, LFG est mieux adapté pour la syntaxe, RG est bien adapté pour l'analyse sémantique et SFG pour l'analyse pragmatique. Bien qu'ils paraissent complémentaires, ces formalismes sont incompatibles entre eux et aucune combinaison de ces grammaires n'est possible. En effet, chaque formalisme linguistique a défini une description unifiée c'est-à-dire un langage symbolique avec lequel les phénomènes linguistiques sont décrits. Par conséquent, il n'est pas possible par exemple d'utiliser une grammaire XFST avec une grammaire LFG.

4.2.5. L'approche NooJ

L'approche NooJ a pour objectif de recenser et de décrire le plus grand nombre possible de phénomènes linguistiques dans leur diversité. Elle suppose l'inexistence d'un formalisme « universel » capable de formaliser tous les phénomènes des langues naturelles. Elle intègre de ce fait plusieurs formalismes en utilisant le mieux adapté pour le phénomène linguistique à décrire. Ces formalismes permettent la construction des différentes grammaires de la hiérarchie de Chomsky à savoir des grammaires rationnelles, des grammaires hors contexte, des grammaires contextuelles et des grammaires non restreintes. A ces grammaires génératives « standard », NooJ permet l'ajout de quelques mécanismes afin de faciliter la description de certains phénomènes linguistiques, en construisant des grammaires

[23] TAG (Joshi, 1987)

[24] LIG (Gazdar, 1988)

[25] HG (Pollard, 1984)

intermédiaires entre deux types de grammaires comme les grammaires « légèrement sensibles au contexte ». NooJ se base sur un environnement « multiple » de développement des grammaires qui intègre des outils de formalisation selon la nature de phénomène linguistique à décrire. Ces outils formalisent donc différemment les phénomènes linguistiques, par exemple les phénomènes orthographiques sont décrits différemment des phénomènes syntaxiques.

L'idée est d'effectuer des analyses complexes d'un texte à partir des analyses de types différents, chacune très simple. En effet, NooJ contient des formalismes qui permettent la description d'alphabets, des dictionnaires, des grammaires flexionnelles et dérivationnelles, des grammaires d'agglutination, des grammaires syntaxiques locales, des grammaires syntaxiques structurelles et des grammaires transformationnelles. Par exemple, le formalisme utilisé pour décrire la variation orthographique en chinois est différent de celui qui sert à décrire la dérivation morphologique en arabe, et le formalisme utilisé pour décrire l'agglutination des langues sémitiques ne permettrait pas de traiter la dérivation des langues latines.

La méthode multiple adoptée par NooJ ne suppose pas l'utilisation d'une notation distincte par formalisme. En effet, la TAS, dans laquelle toutes les informations linguistiques issues de tous les types d'analyse sont rangées, permet d'assurer la communication et la comptabilité entre les différents analyseurs de NooJ à chaque niveau linguistique. D'où la possibilité d'utiliser une seule notation unifiée pour tous les formalismes. La même notation permet donc d'établir des grammaires rationnelles, des grammaires hors contexte, des grammaires contextuelles et des grammaires non restreintes. De plus, « elle s'enrichit de façon ascendante : ainsi, une grammaire rationnelle NooJ peut être intégrée sans modification à une grammaire hors contexte, à une grammaire contextuelle ou à une grammaire non restreinte » (Silberztein, 2015) permettant ainsi de construire et d'accumuler des ressources linguistiques de nature très diverses.

La notation unifiée de NooJ s'affiche donc comme un standard qui simplifie le développement, la gestion, l'accumulation et le partage des grammaires. En effet, elle apporte plusieurs mécanismes à savoir des variables, des contraintes d'accord et des traits spéciaux qui permettent de décrire les constituants et les phénomènes linguistiques triviaux qui seraient décrits difficilement par des grammaires génératives. De plus, cette notation offre une flexibilité pour passer d'un type de grammaire à un autre lors du développement des grammaires, car une grammaire hors contexte NooJ n'est rien d'autre qu'une grammaire rationnelle à laquelle on a ajouté un mécanisme supplémentaire permettant de décrire les

règles de réécriture avec un membre droit composé d'une suite quelconque de symboles terminaux et non-terminaux et une grammaire contextuelle n'est qu'une grammaire hors contexte à laquelle on a ajouté des contraintes.

Finalement, l'approche NooJ adopte donc une architecture en « cascade » qui repose sur plusieurs analyseurs. En effet, les différents analyseurs traitent le texte niveau par niveau depuis l'analyse des caractères jusqu'à l'analyse sémantique. Selon la grammaire adoptée pour la description du phénomène linguistique, NooJ choisit l'analyseur le plus efficace pour compiler la grammaire. Une telle architecture permet de gagner en efficacité et en puissance permettant ainsi l'analyse massive de grands volumes textuels.

5. Conclusion

Dans ce chapitre nous avons introduit le traitement automatique des langues (TAL) et ses finalités. Nous avons présenté les quatre niveaux d'analyse du langage. L'analyse lexicale permet l'identification des unités du vocabulaire du texte en faisant appel à des lexiques et à des analyses morphologiques. Puis, pour étudier les structures des séquences linéaires des éléments du vocabulaire, nous avons distingué l'analyse syntaxique locale, qui s'intéresse à des séquences d'éléments du vocabulaire relativement petites et figées, de l'analyse syntaxique structurelle qui consiste à associer une représentation des groupements structurels entre les unités ainsi que des relations fonctionnelles qui unissent ces groupes d'unités. Puis, nous avons abordé l'analyse sémantique qui consiste en une procédure en deux étapes : l'analyse sémantique locale qui à étudie le sens des unités lexicales et l'analyse sémantique propositionnelle qui cherche à identifier le sens véhiculé par une séquence d'unités lexicales. Finalement, nous nous sommes intéressé à l'analyse pragmatique qui étudie les aspects non-vériconditionnels de l'énoncé c'est-à-dire ceux qui prennent en compte le locuteur et le contexte d'énonciation.

Nous avons exposé les différentes approches en TAL :

- L'approche statistique dans laquelle nous avons présenté l'exploration de corpus avec les méthodes d'analyse textométrique et les méthodes empiriques capables d'analyser et d'annoter un texte en langue naturelle en utilisant des modèles de langue. Ces méthodes, qui s'appuient sur un corpus de référence, ne peuvent couvrir tout le vocabulaire fréquemment utilisé des langues non-standardisées comme les langues à tradition orale ou les langues anciennes à cause de l'absence d'une orthographe

normalisée et de la présence d'un vocabulaire influencé par des variations chronologiques et géographiques.

- L'approche symbolique consiste à modéliser les connaissances linguistiques qui seront exploitables par un système d'analyse automatique afin de traiter un texte en langue naturelle. Nous avons exposé les grammaires génératives classées selon la hiérarchie de Chomsky en quatre types : grammaires rationnelles, grammaires hors contextes, grammaires contextuelles et grammaires non restreintes. Ces grammaires sont très difficilement exploitables telles quelles pour une description linguistique. C'est la raison pour laquelle les formalismes linguistiques ont été créés. Ces formalismes sont incompatibles entre eux et leur complexité ne les rend pas pratiques pour une formalisation exhaustive de la langue. Nous adoptons, pour notre part, l'approche NooJ avec sa méthode multiple, sa notation unifiée et son architecture en cascade afin d'analyser des langues non standardisées telles que le moyen français.
- Nous signalons également l'existence et l'émergence des méthodes hybrides qui combinent des techniques issues des méthodes empiriques pour avoir une meilleure capacité de prédiction et par conséquent augmenter le rappel, et des techniques issues des méthodes symboliques pour améliorer la précision du système. En effet, de nombreux travaux en TAL n'opposent pas l'approche statistique à l'approche symbolique, bien au contraire de nombreuses méthodes intègrent des règles symboliques et des modèles probabilistes (Tellier, 2010).

Chapitre 2

L'annotation automatique de corpus

1. Introduction

Au début du XXème siècle, les analyses manuelles des corpus ont données naissance à la théorie de l'information (Zipf, 1930). De fait, l'utilisation de corpus dans diverses disciplines n'est pas une pratique récente. Les premiers corpus électroniques ont vu le jour vers la moitié du XXème siècle et leur utilisation se limite aux analyses lexicales ou à des intérêts pédagogiques. De nos jours, l'intérêt de l'usage des corpus se généralise à plusieurs disciplines comme la médecine, l'écologie et diverses sciences humaines et sociales. En outre, les corpus sont des ressources linguistiques indispensables pour le développement des applications en TAL. En effet, ils sont utilisés pour entraîner les modules d'analyse automatique basée sur les algorithmes par apprentissage ou pour développer les grammaires utilisées par les analyseurs symboliques. De ce fait, un corpus est une collection de données sélectionnées dans un but descriptif et/ou applicatif ayant une bonne représentativité des phénomènes à traiter et une taille minimale nécessaire à l'analyse ainsi qu'une sauvegarde en format électronique qui facilite son exploitation et son développement en utilisant des standards pour la structuration des données comme TEI (Text Encoding Initiative) et EAGLES (Corpus Encoding Standard XCES) (Kurdi, 2017). La taille du corpus est souvent prédéterminée lors de la conception du système en fonction de l'architecture matérielle et logicielle mise en place. Notons qu'avec l'émergence des nouvelles techniques issues de la science des données, les tailles des corpus sont de plus en plus importantes et la représentativité de corpus devient le critère le plus pris en considération lors du développement du corpus. Depuis des décennies, des corpus volumineux ont été constitués, tels que Brown Corpus[26] (Kucera & Francis, 1967) pour l'anglais contenant un million de

[26] Brown Corpus ou The Brown University Standard Corpus est un corpus de l'anglais américain qui a été constitué pour des études en linguistique de corpus dans les années 1960 par un ensemble des échantillons des œuvres publiées en 1961 aux Etats-Unis.

mots et Le trésor de la langue française[27] (Imbs P., 1971) contenant 160 millions de mots. L'apparition des « linguistiques de corpus »[28] montre l'intérêt que les linguistes portaient aux corpus, en particulier pour ceux annotés pour les études linguistiques. En effet, les corpus annotés ont gagné en productivité vu leur utilité pour l'étude du langage. Des corpus annotés, comme le British National Corpus avec 100 millions de mots annotés (Burnard L., 1995), variés et de grande taille, ont ainsi vu le jour. Malgré les critiques Chomskyennes de l'approche distributionnelle, les méthodes basées sur corpus de type exploratoire et de type confirmatoire se sont vite imposées et plusieurs types d'annotations ont été définis pour l'annotation automatique par les systèmes informatiques décisionnels. Nous allons présenter ces différents types d'annotation linguistique puis, deux systèmes d'annotation automatique feront l'objet d'une étude approfondie à savoir l'annotation grammaticale établie grâce aux étiqueteurs morphosyntaxiques et l'annotation sémantique effectuée en mettant en place un système de reconnaissance des entités nommées. Enfin, nous présenterons les métriques que nous avons adoptées pour l'évaluation des systèmes d'annotations automatiques.

2. Les types d'annotations

L'annotation de corpus est définie comme l'enrichissement des données textuelles brutes avec des informations de nature interprétative (Habert et al., 97). Les corpus annotés sont ainsi utilisés dans plusieurs thématiques de diverses disciplines telles que la linguistique, la littérature, la philosophie et l'histoire. Des annotations de plusieurs types sont utilisées parmi lesquelles les annotations linguistiques qui se décomposent selon Véronis (2000) en quatre catégories à savoir : phonétique, grammaticale, sémantique et multilingue.

[27] Le Trésor de la langue française (TLF) a été élaboré par le laboratoire Analyse et traitement informatique de la langue française (ATILF). Il a été constitué à partir des textes littéraires des XIXe et du XXe (1789-1960) dans le but d'élaborer un dictionnaire de la langue française.

[28] « Les linguistiques du corpus » présentent une diversité d'approches qui font appel à plusieurs types de ressources langagières et à des outils d'analyse pour étudier le langage à travers les contenus de textes réels. (Habert et al., 97)

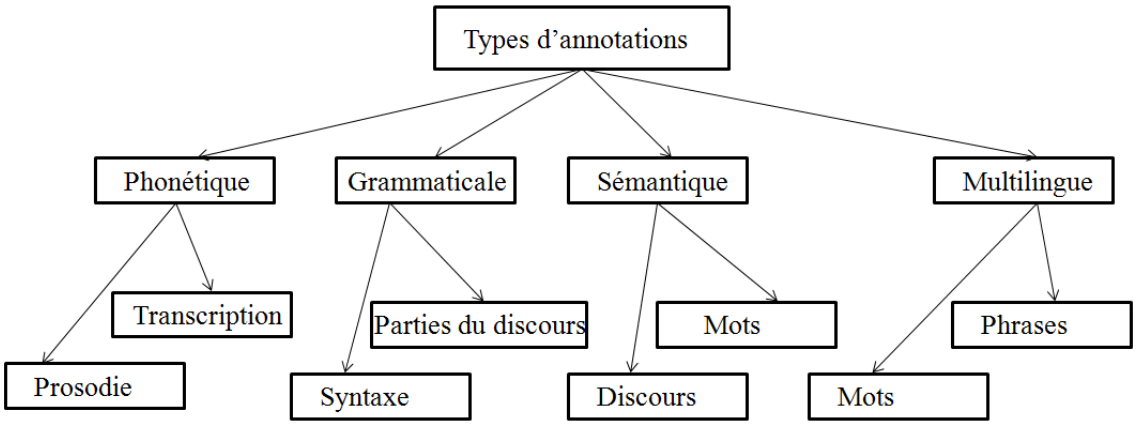


Figure 26. Les différents types d'annotations

La construction des corpus annotés manuellement est un processus long en termes de temps et coûteux en termes de ressources humaines. Il demande un bon niveau de qualification pour garantir une bonne qualité d'annotation et un nombre d'erreurs négligeable afin que le corpus annoté soit utilisable pour diverses analyses. Par conséquent, plusieurs outils ont été développés dans le but de rendre cette tâche humainement faisable. Nous distinguons deux types d'outils d'annotation :

- Les outils d'aide à l'annotation semi-automatique, qui proposent un ensemble de fonctionnalités pour faciliter l'exploration du corpus, et l'annotation manuelle. Ils sont largement utilisés pour l'annotation des corpus des langues à tradition orale. Citons à titre d'exemple *ELAN*[29] (Brugman et al., 2004), *EXMARaLDA*[30] (Schmidt, 2004) et *Le Trameur* (Fleury, 2007).

- Les outils d'annotation automatique qui permettent la mise en place d'un module capable d'annoter automatiquement un corpus en développant des grammaires ou en utilisant des techniques d'apprentissage automatique. Par exemple, la plateforme NooJ donne la possibilité de développer en toute souplesse des chaînes de traitement en utilisant diverses ressources linguistiques afin de produire plusieurs types d'annotations. Des étiqueteurs

[29] ELAN est un logiciel d'annotation et de transcription manuel ou semi-automatique des enregistrements audio ou vidéo. Il dispose d'un modèle de données basé sur les niveaux qui prend en charge l'annotation multi-niveaux et multi-participants de médias basés sur le temps. Il est développé par l'Institut Max Planck pour la psycholinguistique à Nimègue. (https://tla.mpi.nl/tools/tla-tools/elan/)

[30] EXMARaLDA (Extensible Markup Language for Discourse Annotation) est un logiciel qui offre un ensemble d'outils pour créer, gérer et analyser des corpus de langue parlée. Il se compose essentiellement d'un outil de transcription, d'un outil d'administration des métadonnées du corpus et d'un outil pour effectuer des recherches (recherches KWIC) sur les corpus de langue parlée. (http://exmaralda.org/en/about-exmaralda/)

probabilistes existent également, comme *Treetagger*[31] (Schmid, 2013) et *wapiti*[32] (Gahbiche-Braham, 2013) permettant, à partir d'un corpus de référence, la mise en œuvre d'un modèle probabiliste pour l'annotation automatique.

Nous allons nous intéresser aux annotations de type grammatical avec l'étiquetage morphosyntaxique et aux annotations de type sémantique avec la reconnaissance des entités nommées.

3. L'étiquetage morphosyntaxique

3.1. Objets

L'étiquetage morphosyntaxique est utilisé dans la plupart des applications en TAL car il facilite d'autres analyses variées postérieures telles que la lemmatisation, l'analyse syntaxique locale, comme la reconnaissance des entités nommées et l'extraction terminologique ou l'analyse syntaxique structurelle. Il consiste en une opération qui permet, à partir d'un ensemble de couples (forme, étiquette morphosyntaxique), de choisir pour chacun des mots du texte parmi ses étiquettes morphosyntaxiques associées celle(s) qui correspond(ent) au contexte (Fleury, 2009). Ainsi l'exemple Chomskyen de la phrase « *Time flies like an arrow* » pourrait être annoté morphosyntaxiquement comme suit : *Time*[Nom] *flies*[Verbe] *like*[Préposition] *an*[Déterminant] *arrow*[Nom].

L'étiquetage morphosyntaxique permet donc une catégorisation des unités lexicales, qui composent le texte, dans des classes appelées « classes grammaticales » ou « parties du discours ». Cette tâche implique la résolution des difficultés liées à l'ambiguïté posée par les unités lexicales. Dans notre exemple précédent trois unités lexicales sur cinq sont ambiguës, à savoir les formes « *time* » et « *fly* » qui peuvent être un verbe ou un nom et la forme « *like* » qui peut être un verbe ou un adverbe. De même, si nous examinons une phrase en français tel que « *L'élève est retardé à cause de la neige* ». Cette phrase présente cinq mots ambigus sur

[31] Treetagger est un étiqueteur morphosyntaxique qui offre un module qui permet d'entraîner l'étiqueteur sur un corpus de référence afin de générer un modèle d'une langue donnée avec lequel il est possible d'annoter des textes. Techniquement, il se base sur les arbres de décision et sur les chaines de Markov cachées. (www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)

[32] Wapiti est une implémentation des CRF linéaires qui permet, à partir d'un corpus de référence, de générer des modèles permettant une annotation automatique qui supporte plusieurs types d'annotations. (https://wapiti.limsi.fr/)

neuf : les formes « élève », « cause » et « neige » peuvent être un nom ou un verbe et les formes « l' » et « la » peuvent être étiquetés comme « pronom », «article défini » ou « déterminant ».

En effet, la levée d'ambiguïté morphosyntaxique nécessite l'exploitation de l'ensemble des informations mises en œuvre lors de la compréhension automatique du langage (Paroubek & Rajman, 2000). Par exemple, en utilisant une description morphosyntaxique indiquant la classe grammaticale, le genre et le nombre pour annoter les deux phrases : « *Gabrielle était à la salle de conférence. On l'a vue à l'entrée* », la levée d'ambiguïté du pronom « l' » nécessite l'utilisation d'une information sémantique qui est la nature du sujet du verbe « être » et la réquisition de la connaissance pragmatique du sexe de la personne mentionnée. Par conséquent, un système capable d'effectuer de telles analyses sémantiques et pragmatiques doit requérir la résolution de compréhension automatique de texte. Or, dans une chaîne de traitement TAL, l'étiquetage morphosyntaxique est une tâche qui est supposée antérieure aux analyses sémantiques et pragmatiques. Les méthodes, que nous exposons par la suite, ne font donc pas appel à l'analyse syntaxique structurelle, sémantique locale, sémantique propositionnelle et pragmatique pour effectuer l'étiquetage morphosyntaxique.

3.2. Jeu d'étiquettes

L'ensemble des classes grammaticales utilisées pour effectuer l'étiquetage morphosyntaxique souvent nommé « jeu d'étiquettes » (*tagset* en anglais) constitue un formalisme qui permet de décrire les propriétés morphosyntaxiques d'une unité lexicale dans son contexte d'énonciation. Ces étiquettes peuvent contenir diverses informations linguistiques comme, par exemple pour le français, le genre et le nombre pour les noms et les adjectifs et la personne, le genre, le nombre, le mode et la voix pour les verbes. Le choix des étiquettes dépend de plusieurs critères définis généralement lors de la mise en place de la chaîne de traitement. Finalement, la définition d'un jeu d'étiquettes adéquat est indispensable pour assurer une meilleure performance de l'étiqueteur morphosyntaxique. Par exemple, Chanod & Tapanainen (1995) ont démontré qu'il est possible d'améliorer considérablement la performance d'étiquetage en réduisant la description morphosyntaxique à quelques classes grammaticales.

La définition d'un jeu d'étiquettes est donc le résultat d'une phase expérimentale sur corpus durant laquelle plusieurs annotations manuelles et des mesures de performances sont effectuées. Pustet (2003) définit les étiquettes principales et universelles indispensables à toute analyse linguistique à savoir les noms, les verbes et les adjectifs. (Taylor, 2003) divise

l'ensemble des classes grammaticales d'un jeu d'étiquettes en étiquettes et sous-étiquettes. Tandis que Güngör (2010) les catégorise en classes ouvertes qui peuvent être associées à des unités lexicales régulièrement ajoutées à la langue tels que les noms, les verbes, les adjectifs et les adverbes et en classes fermées contenant principalement des classes grammaticales qui constituent l'ensemble des mots-outils (fonction *words*) auxquels s'ajoutent rarement de nouvelles unités lexicales. Ces différentes classifications des catégories grammaticales peuvent être prises en compte lors d'un étiquetage automatique, par exemple, pour la reconnaissance des mots inconnus et pour le traitement des mots ambigus.

Plusieurs standards ont été développés pour normaliser les descriptions morphosyntaxiques de diverses langues. Parmi les plus populaires, nous citons le standard Multext (Ide&Veronis, 1994), qui définit les jeux d'étiquettes de quatorze langues principalement des langues romanes dont le français, et le jeu d'étiquettes proposé par la communauté de l'universal dependancies qui établit une liste de partie du discours utilisées pour le développement des copus annotés par un formalisme de dépendance. Plusieurs autres jeux d'étiquettes ont été définis pour le français comme celui mis en place au cours de la campagne d'évaluation des étiqueteurs de français appelée l'« action Grace » (Adda et al., 1999). Il existe également un jeu d'étiquettes défini pour l'ancien français nommé « Cattex 2009 »[33] (Mazziotta, 2012) qui dispose d'un guide d'annotation mis en place par le projet « *Syntactic Reference Corpus of Medieval French (SRCMF)* »[34] (Prévost & Stein, 2013). Le tableau 2 montre un échantillon des descriptions morphosyntaxiques, utilisées par le standard « Cattex 2009 », associées aux verbes, aux noms et aux adjectifs.

[33] Cattex 2009 est la liste des étiquettes morphosyntaxiques adoptée, par l'équipe IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) de l'ENS de Lyon pour l'annotation du corpus de la base du français médiéval qui a été fondé par Christiane Marchello-Nizia (http://bfm.ens-lyon.fr/spip.php?article176)

[34] Le projet *Syntactic Reference Corpus of Medieval French* (SRCMF) a comme objectif le développement d'une version annotée du corpus d'Amsterdam. En effet, il s'agit d'annoter manuellement une partie de ce corpus. Les annotations manuelles effectuées ont été utilisées pour entrainer des analyseurs automatiques essentiellement l'étiquetage morphosyntaxique avec Treetagger. Finalement, un outil a été mis en place qui permet l'interrogation du corpus annoté.

CATEG	*TYPE*	*MODE*	*TEMPS*	*PERS.*	*NOMB.*	*GENRE*	*CAS*	*DEGRE*	*Contr.*	*Commentaires, Exemples*
VER	cjg [1]	ind/imp/con/ sub [2]	pst/ipf/ fut/psp [3]	0/1/2/3 [4]	s/p [5]					Quant il la **voit** venir
	inf				s/p		n/r/- [6]			Quant il la voit **venir**
	ppe				s/p	m/f [7]	n/r			ce que j'ai toz jorz **celé**
	ppa				s/p	m/f	n/r			Et aussi fist il en **veillant**
NOM	com				s/p	m/f	n/r			en ceste **nuit**
	pro				s/p	m/f	n/r			**Boort**
ADJ	qua				s/p	m/f/n	n/r	p/c/s [8]		en pechié **mortel**
	ind				s/p	m/f/n	n/r			une **autre** nef
	car				s/p	m/f/n/-	n/r			apres ces **.ii.** vertuz
	ord				s/p	m/f/n	n/r			li **quarz** jorz
	pos			1/2/3	s/p	m/f/n	n/r			contre .i. **suen** voisin

Tableau 2. Description du jeu d'étiquettes pour l'ancien français « Cattex 2009 »

3.3. Processus d'étiquetage morphosyntaxique

Les travaux sur l'étiquetage morphosyntaxique ont débuté dans les années soixante suite à la constitution d'un corpus anglais « *Brown corpus* » (Kucera & Francis, 1967). Des outils ont été mis en place afin d'aider les annotateurs humains à étiqueter le million de mots du *Brown corpus*. Depuis, les travaux se sont multipliés en utilisant diverses techniques pour annoter des langues variées. De nos jours, les étiqueteurs morphosyntaxiques affichent des résultats satisfaisants, de l'ordre de 95 %. Bien que l'évaluation des performances d'un étiqueteur soit une tâche complexe car elle dépend de plusieurs facteurs liés généralement à la langue, à la complexité du jeu d'étiquettes utilisé et au choix du corpus de test (Véronis et al., 1995), nous constatons que les étiqueteurs sont relativement efficaces quelle que soit la technologie et la méthode d'évaluation adoptées. Quelques études et observations nous permettent d'interpréter ces résultats. En effet, selon Véronis (2000), un bon nombre des unités lexicales ne sont pas ambigües et l'application d'un lexique contenant l'unité lexicale et la description morphosyntaxique suffit pour produire une performance de 60 %. Pour les unités lexicales ambiguës (environ 40 %), Church & Mercer (1993) montrent que la sélection de l'étiquette la plus fréquente permet de résoudre efficacement 90 % des ambiguïtés. De plus, selon la loi de Zipf un petit nombre d'unités lexicales les plus fréquentes représentent environ 80 % des unités lexicales présentes dans un texte anglais. Toutes ces observations et ces expériences montrent qu'il est possible d'atteindre des bons résultats dépassant largement 90 % en se concentrant seulement sur l'ensemble des unités lexicales qui sont à la fois fréquentes et ambiguës. Cette hypothèse a été prouvée par Tzourkermann et al. (1996) suite à une étude réalisée sur deux corpus en français extraits du journal *Le Monde*. Selon cette étude,

plus de la moitié des mots ne sont pas ambigus et une grosse part de l'ambiguïté est détenue par un petit nombre de mots fréquents qui sont généralement des mots outils.

Le processus d'étiquetage morphosyntaxique, illustré par la figure 27, se décompose généralement en trois phases : une phase de segmentation, une phase d'application des lexiques et des analyses morphologiques, et une phase de désambiguïsation.

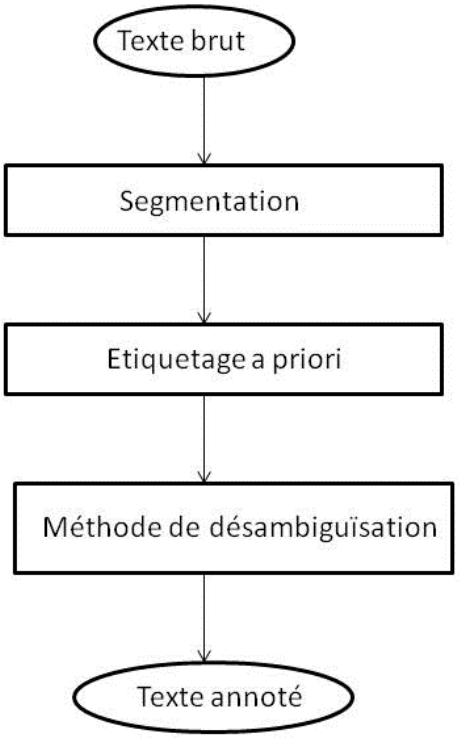


Figure 27. Processus d'étiquetage morphosyntaxique

La phase de segmentation aussi appelée « phase de prétraitement » consiste à standardiser les différentes variantes typographiques et à transformer, selon des séparateurs, un texte en une suite d'unités lexicales. Cette opération n'est pas triviale puisqu'une mauvaise segmentation entraîne des problèmes pour effectuer des analyses ultérieures. En effet, cette phase peut contenir des analyses permettant la reconnaissance des unités lexicales simples ou composées. Par conséquent, l'ensemble des unités résultats et les ambiguïtés produites par cette segmentation typographique seront prises en compte à la suite du processus d'étiquetage. Une fois les différentes unités identifiées, l'étape suivante consiste à attribuer pour chaque unité lexicale l'ensemble des descriptions morphosyntaxiques associées sans tenir compte du contexte d'énonciation. Par exemple, pour la forme « *neige* » on attribue l'ensemble des descriptions morphosyntaxiques associées à savoir « *Verbe à la première personne de singulier* », « *Verbe à la troisième personne de singulier* » et « *nom féminin singulier* ». Cette

tâche est assurée par un lexique contenant l'unité lexicale et l'ensemble des descriptions morphosyntaxiques associées, et par des analyses morphologiques qui peuvent être utilisées pour la construction du lexique et pour identifier et prédire les unités lexicales dites « inconnues »[35] qui n'existent pas dans le lexique. A cette phase, pour le français par exemple, nous pouvons obtenir un étiquetage avec des performances qui peuvent dépasser les 60% si nous utilisons des jeux d'étiquettes minimales qui ne contiennent que les classes grammaticales principales telles que nom, verbe, adverbe, etc. Cependant, avec un jeu d'étiquettes très fin par exemple contenant plus de 300 descriptions morphosyntaxiques, le taux de performance dépasse rarement les 50% (Paroubek & Rajman, 2000). Suite à l'application des lexiques et des analyses morphologiques, les unités lexicales ambiguës auxquelles plus d'une étiquette morphosyntaxique sont associées, sont traitées, c'est la phase de désambiguïsation. Celle-ci consiste à choisir et associe parmi les étiquettes morphosyntaxiques associées aux unités lexicales ambiguës celle(s) qui correspond(ent) au contexte.

3.4. Les méthodes de désambiguïsation

Plusieurs méthodes ont été développées en utilisant des techniques variées afin d'effectuer la tâche de désambiguïsation. Nous les classons en deux catégories : les méthodes par apprentissage et les méthodes à base de règles.

3.4.1. Les méthodes par apprentissage

Rappelons ce que nous avons vu avec les modèles de langue[36] qui permettent la mise en place d'une distribution d'une séquence d'unités. En utilisant un modèle tel que les N-grammes, la distribution d'une séquence d'ALU est calculée avec la formule suivante :

$$P(w_1,w_2,w_3,\ldots,w_{n-1},w_n) = P(w_1)\,P(w_2|w_1)\,P(w_3|w_2,w_1)\ldots P(w_n|w_{n-1},w_{n-2},\ldots w_1)$$

Un modèle de Markov suppose que seul un voisinage restreint permet le calcul de la probabilité conditionnelle de $P(w_n|w_1^{n-1})$.

Par conséquent, nous pouvons nous limiter au bigramme et notre formule sera ainsi :

---

[35] Les unités lexicales inconnues est une notion utilisée pour désigner les formes du texte qui ne sont pas reconnues par une analyse lexicale en utilisant deux principales ressources linguistiques à savoir le lexique et les règles morphologiques.

[36] Le modèle de langue est un concept probabiliste vu à la section 4.1.1.2 du chapitre 1

$$P(w_1, w_2, w_3, \ldots, w_{n-1}, w_n) = P(w_1)\,P(w_2|w_1)\,P(w_3|w_2)\ldots P(w_n|w_{n-1})$$

Ou au trigramme, pour avoir la formule suivante :

$$P(w_1, w_2, w_3, \ldots, w_{n-1}, w_n) = P(w_1)\,P(w_2|w_1)\,P(w_3|w_2, w_1)\ldots P(w_n|w_{n-1}, w_{n-2})$$

La méthode par apprentissage pour l'étiquetage morphosyntaxique consiste à définir une distribution de probabilités qui permet d'associer à une séquence d'unités lexicales une séquence d'étiquettes morphosyntaxiques. Il s'agit donc de maximiser la probabilité conditionnelle $P(t_{1,n}|w_{1,n})$ afin de permettre de choisir la séquence d'étiquettes la mieux appropriée pour être associée à une séquence d'unités lexicales donnée. La mise en œuvre d'un tel modèle nécessite d'énoncer des hypothèses probabilistes afin de pouvoir décrire ces probabilités avec un nombre fini de paramètres. Dans ce but, le théorème de Bayes est appliqué afin d'exprimer notre probabilité conditionnelle avec la probabilité d'une séquence d'unités lexicales sachant une séquence d'étiquettes $P(w_{1,n}|t_{1,n})$, la probabilité d'une séquence d'étiquettes $P(t_{1,n})$ et la probabilité d'une séquence d'unités lexicales $P(w_{1,n})$.

$$\arg_{t1,n} \max P(t_{1,n}|w_{1,n}) = \arg_{t1,n} \max(P(w_{1,n}|t_{1,n})P(t_{1,n})/P(w_{1,n}))$$

Puisque la séquence d'unités $P(w_{1,n})$ est fixe c'est-à-dire qu'elle a la même valeur pour toutes les parties de discours, cette dernière peut être négligée et éliminée du calcul.

$$\arg_{t1,n} \max P(t_{1,n}|w_{1,n}) = \arg_{t1,n} \max(P(w_{1,n}|t_{1,n})P(t_{1,n}))$$

La probabilité $P(w_{1,n}|t_{1,n})$ peut être décomposée en produit de probabilités conditionnelles en utilisant la règle de la chaîne :

$$P(w_{1,n}|t_{1,n}) = P(w_1|t_1..t_n)P(w_2|w_1,t_1\ldots t_n)\ldots.P(w_n|w_1\ldots w_{n-1},t_1\ldots t_n)$$

Afin de simplifier le calcul de cette probabilité, on considère que les unités lexicales sont indépendantes les unes des autres et que chaque unité lexicale dépend uniquement de sa propre étiquette. Cette hypothèse signifie que la description morphosyntaxique associée à une unité lexicale contient toutes les informations linguistiques qui permettent le conditionnement probabiliste de l'unité lexicale dans son contexte. Par conséquent, la probabilité d'association d'une unité lexicale à une étiquette peut être simplifiée ainsi :

$$P(w_i|w_1\ldots w_{i-1},t_1\ldots t_n)=P(w_i|\,t_i)$$

Linguistiquement, cette hypothèse n'est pas toujours vérifiée et sa validité dépend du jeu d'étiquettes utilisé. En effet, si nous utilisons un jeu d'étiquettes très fin qui contient des descriptions morphosyntaxiques détaillées permettant d'associer plusieurs informations linguistiques à chaque unité lexicale, cette hypothèse peut être vérifiée. Sinon, avec l'utilisation d'un jeu d'étiquettes contenant des classes grammaticales génériques telles que

nom, verbe et déterminant sans spécifier le genre et le nombre par exemple, cette hypothèse ne tient pas dans la mesure où cette description ne permet pas de vérifier l'association entre des séquences d'étiquettes et des séquences d'unités lexicales. Par exemple, lorsqu'il s'agit d'annoter avec un modèle bigramme la séquence « les cause» qui peut être associée à la séquence d'étiquettes « Déterminant Nom » ou à la séquence d'étiquettes « Pronom Verbe », une telle égalité est possible P(cause|les, Déterminant Nom) = P(cause|Nom) alors qu'elle est linguistiquement incorrecte et cela à cause de l'absence d'accord entre le déterminant et le nom.

La probabilité d'association d'une séquence d'unités lexicales à une séquence d'étiquettes peut donc être exprimée en utilisant cette formule :

$P(W_{1,n}|T_{1,n})= P(w_1| t_1) P(w_2| t_2)... P(w_n| t_n)$

Concernant la probabilité d'une séquence d'étiquettes P($t_{1,n}$), elle se transforme en un produit de probabilité conditionnelle en appliquant la règle de la chaîne :

$P(t_{1,n}) = P(t_1)P(t_2| t_1)....P(t_n| t_1... t_{n-1})$

Cette probabilité peut être simplifiée si nous utilisons des contextes limités en supposant que chaque probabilité $P(t_n|t_1... t_{n-1})$ dépend de k étiquettes précédentes avec N=1 ou N=2 : $P(t_n|t_1... t_{n-1}) = P(t_n|t_{n-1})$. Cette hypothèse n'est pas linguistiquement défendable parce qu'elle suppose que la désambiguïsation morphosyntaxique est d'une portée de dépendance limitée entre deux phénomènes morphosyntaxiques (Paroubek & Rajman, 2000). Si N=1, notre probabilité P($t_{1,n}$) peut être simplifiée et exprimée par la formule suivante :

$P(t_{1,n}) = P(t_1)P(t_2| t_1)....P(t_n| t_{n-1})$

La probabilité $P(t_{1,n}|w_{1,n})$ d'association d'une séquence étiquettes à une séquence d'unités lexicales donnée se décompose donc en la « probabilité lexicale » P($w_i|t_i$) qui correspond à la nature de l'unité lexicale et à la « probabilité transitoire » P(t_i) qui correspond aux transitions entre les étiquettes.

$$
\begin{aligned}
P(w_{1,n}|t_{1,n})P(t_{1,n}) &= \prod_{i=1}^{n} P(t_i|w_i) \times P(t_n|t_{1,n-1}) \times P(t_{n-1}|t_{1,n-2}) \times \cdots \times P(t_2|t_1) \\
&= \prod_{i=1}^{n} P(w_i|t_i) \times P(t_n|t_{n-1}) \times P(t_n - 1|t_{n-2}) \times \cdots \times P(t_2|t_1) \\
&= \prod_{i=1}^{n} [P(w_i|t_i) \times P(t_i|t_{i-1})]
\end{aligned}
$$

Nous concluons que les méthodes de désambiguïsation probabiliste cherchent à trouver la séquence d'étiquettes qui maximise la probabilité lexicale et la probabilité transitoire.

$$\arg\max_{t_{1,n}} P(t_{1,n}|w_{1,n}) = \arg\max_{t_{1,n}} \prod_{i=1}^{n} [P(w_i|t_i) \times P(t_i|t_{i-1})]$$

Ce qui différencie un modèle d'un autre est la manière de calculer la probabilité lexicale et la probabilité transitoire. En utilisant un modèle de Markov caché (HMM) par exemple, on cherche à calculer l'estimation maximale en observant le comportement des unités lexicales dans le corpus d'apprentissage. En effet, durant la phase d'annotation morphosyntaxique d'une séquence d'unités lexicales, on cherche à calculer la probabilité de toutes les séquences d'étiquettes possibles en utilisant la méthode du « maximum de vraisemblance » par exemple.

$P(w_i|t_i) = f(w_i|t_i) / f(t_i)$ avec f(w,t) le nombre d'occurrence d'un mot w ayant l'étiquette t et f(t) est le nombre d'occurrence de l'étiquette t.

$P(t_i|t_{i-1}\ldots\ t_{i-k}) = f(t_{i-k}\ldots t_i)/f(t_{i-k}\ldots\ t_{i-1})$ avec $f(t_m\ldots t_n)$ est le nombre d'occurrence d'une étiquette dans la séquence $t_m\ldots t_n$.

Un algorithme doit donc trouver la séquence d'étiquettes de probabilités conditionnelles maximales en calculant la probabilité de toutes les séquences d'étiquettes pouvant être associées à une séquence d'unités lexicales. À cause de l'importante ambiguïté morphosyntaxique des langues naturelles, un tel algorithme devient rapidement d'une complexité exponentielle et la recherche de la séquence d'étiquettes optimale est difficilement envisageable. En effet, l'utilisation d'un modèle de Markov caché implique une correspondance formelle entre les états cachés et les étiquettes morphosyntaxiques possibles et entre les symboles observables et les unités lexicales. Ces correspondances permettent une récupération de l'ensemble des résultats produits. Ensuite, il est possible en utilisant des techniques de programmation dynamique de mettre en place l'algorithme de Viterbi (1967) qui présente une solution efficace, de complexité linéaire, pour la résolution du problème de recherche de la séquence d'étiquettes de probabilité conditionnelle maximale.

L'inconvénient des méthodes probabilistes de désambiguïsation réside dans la diminution significative de performance lorsqu'il s'agit d'étiqueter un texte d'un domaine ou d'un genre relativement éloigné du corpus d'apprentissage. En effet, comme nous avons vu, les probabilités du modèle sont estimées grâce aux fréquences observées à partir d'un corpus d'apprentissage. Mais, le corpus d'apprentissage ne permet pas d'avoir toutes les estimations des unités lexicales d'une langue puisqu'il y a des unités lexicales à annoter qui ne figurent pas dans le corpus d'apprentissage. Par conséquent, l'estimation de ces unités lexicales sera égale à zéro. Ce problème commun à toutes les méthodes probabilistes appelé « *problème des*

données clairsemées »[37] (*Sparse Data Problem*) ou « *régularisation* » peut être traité en utilisant des méthodes dites de « lissage »[38] (*smoothing method*) qui cherche à améliorer l'estimation de probabilité des mots peu fréquents telle que la méthode « *lissage de Laplace* »[39] et la méthode « *lissage add delta* »[40] qui permettent de redistribuer la probabilité afin de maximiser la valeur des unités lexicales peu fréquentes (Manning & Schutze, 1999).

3.4.2. Les méthodes à base de règles

Les méthodes à base de règles permettent de désambiguïser les unités lexicales dans leur contexte d'énonciation en utilisant des règles écrites à la main. Ces règles formalisent des contraintes d'occurrences en utilisant les unités à annoter et l'ensemble des traits qui composent la description morphosyntaxique tel que la catégorie grammaticale, le genre et le nombre pour les noms et les adjectifs comme elles peuvent aussi utiliser d'autres types de traits tels que des traits sémantiques (Paroubek & Rajman, 2000).

Les premiers systèmes à base de règles (Klein & Simpson, 1963) posaient de réels problèmes liés au nombre important des règles à décrire par un expert. On parle généralement de milliers de règles (Samuelsson & Voutilainen, 1997) à accumuler, à maintenir dont il faut définir l'ordre d'application. Une tâche jugée « humainement très exigeante » en l'absence de formalismes et d'environnement de développement facilitant l'élaboration et la gestion de règles linguistiques.

Cette tâche est devenue possible grâce à l'utilisation des plateformes linguistiques telles que NooJ qui offre un environnement de développement graphique, des analyseurs permettant l'application d'un ensemble de règles, et une structure d'annotation de texte (TAS) qui permet de gérer les différentes annotations produites. Selon l'approche NooJ, les règles de désambiguïsation peuvent être écrites sous forme d'une grammaire hors contexte de type 2 de la hiérarchie de Chomsky. Ces grammaires dites de « levée d'ambiguïté » sont des

---

[37] En apprentissage automatique, on appelle *problème des données clairsemées* ou *régularisation* tout processus qui ajoute des informations à un modèle, généralement en procédant à une distribution a priori sur les paramètres du modèle, afin d'éviter le surapprentissage.

[38] Une méthode de lissage applique une ou plusieurs techniques qui permettent de réduire les irrégularités et les singularités des paramètres du modèle afin d'améliorer les estimations non pertinentes qui constituent du bruit.

[39] Le lissage de Laplace consiste à maximiser l'estimation de probabilité des unités lexicales en ajoutant 1 au numérateur et en renormalisant le dénominateur.

[40] Lissage add delta est une généralisation de lissage de Laplace qui consiste à ajouter au numérateur et au dénominateur un hyper-paramètre qui sera déterminé en fonction de la taille du corpus.

grammaires locales qui sont appliquées lors d'une analyse syntaxique locale. A titre d'exemple la figure suivante montre une grammaire locale pour la désambiguïsation de quelques mots grammaticaux fréquents et ambigus en français. Pour catégoriser les séquences reconnues, les grammaires locales ajoutent généralement des annotations dans la TAS mais avec notre exemple, nous montrons qu'elles peuvent supprimer des annotations qu'on juge incorrectes. En effet, les grammaires locales produisent des filtres qui sont utilisés pour supprimer de la TAS toutes les annotations incompatibles. Nous citons l'exemple du graphe de la figure 28 qui est composé de 9 sous-graphes permettant de lever l'ambiguïté automatiquement des mots grammaticaux « *a* », « *de* », « *en* », « *est* », etc.

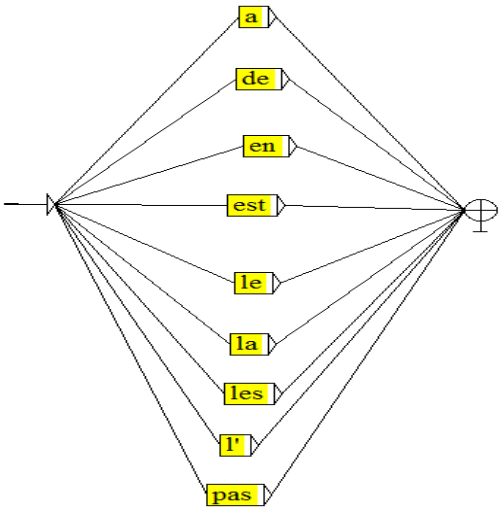


Figure 28. Grammaire locale pour la désambiguïsation des mots grammaticaux fréquents

Le sous-graphe suivant permet la désambiguïsation automatique de la forme « *a* » qui peut être un verbe ou un nom. En appliquant cette grammaire locale, un filtre sera produit et sera utilisé pour supprimer de la TAS toutes les annotations de « *a* » qui ne sont pas un verbe sauf dans le cas où la forme « *a* » est précédé par un déterminant « <DET> ».

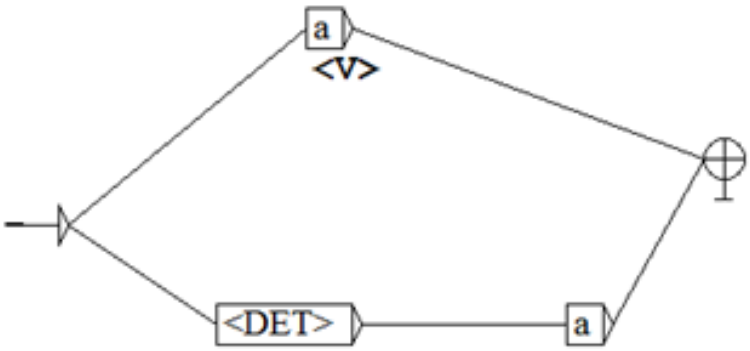


Figure 29. Grammaire locale pour la désambiguïsation de « a »

De même pour le sous-graphe suivant qui permet la désambiguïsation automatique de la forme « en ». Cette dernière peut être une préposition ou un pronom. Notre grammaire locale permet de supprimer de la TAS toutes les annotations de la forme « *en* » qui ne sont pas des prépositions sauf lorsque la forme « *en* » est suivie d’un verbe. Dans ce cas, c’est seulement l’annotation pronom « <PRO> » qui sera sauvegardé dans la TAS.

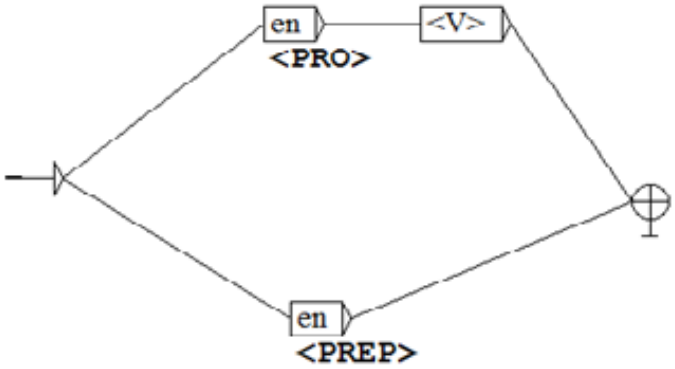


Figure 30. Grammaire locale pour la désambiguïsation de « *en* »

Au contraire des méthodes par apprentissage qui modélisent le contexte proche de l’unité lexicale à annoter, c'est-à-dire qui choisissent un voisinage de taille fixe, généralement des bigrammes ou trigrammes de l’étiquette considérée, les méthodes à base de règles, et plus particulièrement celles utilisant les grammaires hors contextes, permettent la modélisation des contextes avec un grand nombre de voisinages. Ce qui présente un avantage considérable pour désambigüiser certaines formes. Prenons l’exemple du graphe suivant qui permet de désambiguïser la forme « *la* » qui peut être un déterminant féminin singulier ou un pronom. La modélisation de séquences telles que « *je la lui ai donnée* » permet de désambigüiser la forme « *la* » et l’annoter comme étant un pronom.

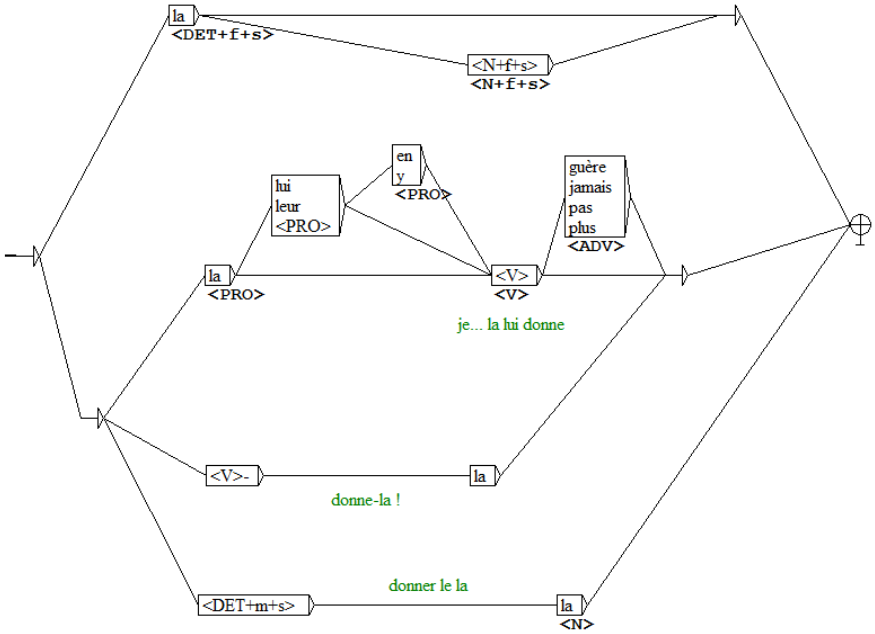


Figure 31. Grammaire locale pour la désambiguïsation de « la »

Concrètement, nous avons mis en place une chaîne de traitement pour l'étiquetage morphosyntaxique basée sur le module français de NooJ. Nous avons donc utilisé les ressources linguistiques fournies pour étiqueter un article du journal *Le Monde*. Nous avons donc appliqué les différentes phases de la méthode illustrées sur la figure suivante. Suite à l'analyse lexicale qui a permis la normalisation et la segmentation des unités lexicales, la reconnaissance des formes simples et composées et les valeurs numériques, nous avons appliqué la grammaire locale de la figure 28 qui a permis la désambiguïsation de neuf formes fréquentes et ambiguës. Finalement, nous avons obtenu un taux de réussite de 72,2%.

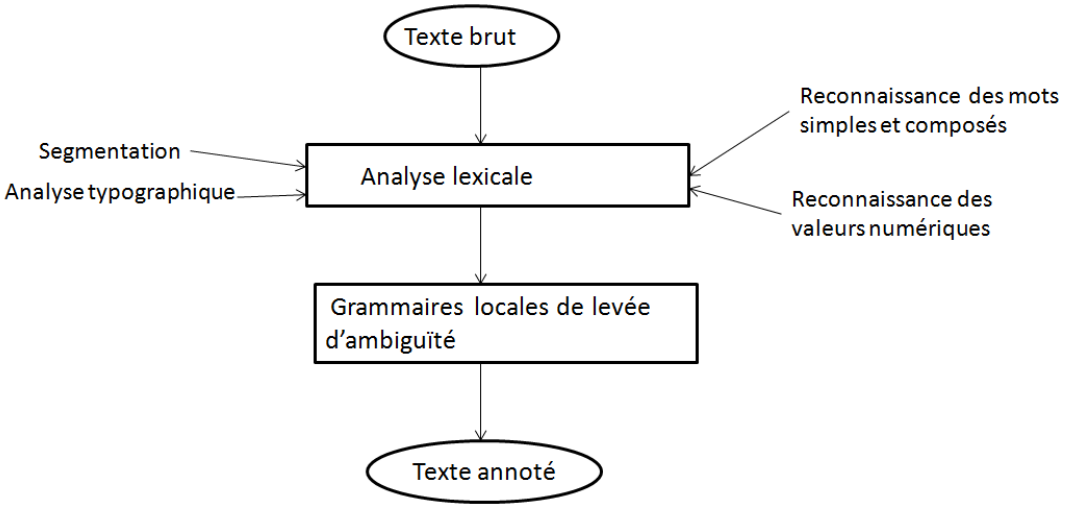


Figure 32. Méthode d'étiquetage morphosyntaxique basée sur l'approche NooJ

4. Reconnaissance des entités nommées

4.1. Historique

De 1987 à 1998, sept conférences appelées MUC[41] (*Message Understanding Conferences*) ont été organisées aux États-Unis financé par l'agence DARPA[42] *(Defense Advanced Research Projects Agency*), qui est une agence du département de défense aux Etats-Unis, chargée de la recherche et du développement des nouvelles technologies, pour encourager le développement des méthodes d'extraction d'information (Poibeau, 2003). Ces conférences sont en réalité des campagnes d'évaluation ; une sorte de compétition, qui rassemble plusieurs équipes du milieu industriel et académique et qui oppose différents systèmes d'extraction d'information. Il s'agit donc d'évaluer ces systèmes en utilisant des mesures métriques comme le rappel et la précision adoptés lors de MUC-2 et le F-Mesure introduit lors de MUC-4. La conférence MUC-6 (1993-1996) avait comme ambition de décomposer le système d'extraction d'information en plusieurs modules dont chacun cherchait à extraire un type d'information. C'est plus précisément en 1996 que la reconnaissance d'entités nommées a été définie comme une sous-tâche de l'extraction d'information en s'intéressant à certaines expressions référentielles et porteuses de sens.

4.2. Les typologies d'entités nommées

Lors de la conférence MUC-6, la tâche de la reconnaissance d'entités nommées se résumait à l'identification des noms de personne, de lieu et d'organisation (Grishman & Sundheim, 1996). Cette catégorisation a été élargie par la suite pour intégrer les expressions numériques et les expressions temporelles. Dès lors, les entités nommées correspondent à des expressions linguistiques organisées hiérarchiquement sur deux niveaux :

- *ENAMEX :* représente des expressions qui réfèrent à des personnes, des lieux et des organisations ;

[41] Les conférences MUC ont été lancées pour encourager le développement de nouvelles et de meilleures méthodes d'extraction d'informations. Elles ont permis l'élaboration de normes d'évaluation comme l'adoption de métriques : la précision et le rappel.

[42] DARPA (Defense Advanced Research Projects Agency) est une agence du département de la défense des États-Unis responsable du développement de technologies émergentes à utilisation militaire. En collaborant avec des partenaires universitaires, industriels et gouvernementaux, DARPA formule et exécute des projets de recherche et de développement afin de fournir des technologies dans divers domaines.

- *TIMEX :* représente l'ensemble des expressions temporelles qui sont constituées par des expressions de temps et par des dates ;
- *NUMEX :* représente l'ensemble des expressions numériques qui sont constituées par des quantités monétaires et par des pourcentages.

Cette typologie, qui a servi pour le traitement d'un corpus journalistique, a été appelée à s'élargir suite à un changement du corpus lors de MUC-7 afin d'ajouter une nouvelle catégorie qui permette d'annoter les noms des produits. Elle a été testée sur un corpus multilingue de données des accidents aériens (Grishman & Sundheim, 1996).

En analysant plusieurs langues différentes telles que l'espagnol et l'arabe, les campagnes *Automatic Content Extraction*[43] (ACE) ont proposé une typologie considérablement plus fine comportant 37 catégories réparties sur 7 types (Doddington et al. 2004) :

– *Facility : Airport, Building-Grounds, Path, Plant, Subarea-Facility ;*

– *Geo-Political : Continent, County-or-District, GPE-Cluster, Nation,*

Population-Center, Special, State-or-Province ;

– *Location : Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body ;*

– *Organization : Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports ;*

– *Person : Group, Indeterminate, Individual ;*

– *Vehicle : Air, Land, Subarea-Vehicle, Underspecified, Water ;*

– *Weapon : Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Underspecified.*

Plusieurs autres importantes campagnes ont été organisées parmi lesquelles les campagnes d'évaluation des corpus oraux en français appelée « Systèmes de Transcription d'Émissions Radiophoniques »[44] ESTER1 (2003-2005) (Le Meur et al., 2004) et ESTER 2 (2006-2008) (Galliano et al., 2009). La typologie mise en place durant les campagnes ESTER est de trois niveaux : le premier et deuxième niveau sont constitués des types et des catégories

[43] Automatic Content Extraction (ACE) est un programme de recherche pour le développement de technologies avancées d'extraction d'information lancé de 1999 à 2008, succédant à MUC et précédant à Text Analysis Conference.

[44] ESTER est une campagne d'évaluation scientifique francophone organisée depuis 2005 qui a été mis en place par l'AFCP (Association française de la communication parlée). Elle vise à mesurer plusieurs points des systèmes de transcription et d'étiquetage de corpus audio numériques.

définies lors de MUC-6 avec en plus un troisième niveau contenant des sous-catégories permettant de caractériser finement l'entité nommée en question.

– Personnes : humain réel ou fictif, animal réel ou fictif ;

– Fonctions : politique, militaire, administrative, religieuse, aristocratique;

– Lieux : géographique naturel, région administrative, axe de circulation, adresse (adresse postale, téléphone et fax, adresse électronique), construction humaine ;

– Organisations : politique, éducative, commerciale, non commerciale, média divertissement, géo-socio-administrative ;

– Production humaine : moyen de transport, récompense, œuvre artistique, production documentaire ;

– Date et heure : date (date absolue, date relative), heure ;

– Montant : âge, durée, température, longueur, surface et aire, volume, poids, vitesse, autre, valeur monétaire.

Dans la même thématique et dans la continuité des campagnes ESTER 1 & 2, citons la campagne d'évaluation ETAPE[45] (Évaluations en Traitement Automatique de la Parole) (2011-2012) durant laquelle le guide d'annotation Quaero[46] (Rosset et al. 2011b) a été utilisé pour assurer une classification fine des entités nommées en mettant en place une typologie comportant 8 catégories principales et 32 sous-catégories.

Pour répondre à des besoins applicatifs variés et spécifiques, diverses autres typologies ont été définies. A titre d'exemple, nous citons la typologie proposée par Tran (2006) utilisée par les systèmes de traduction automatique et par les systèmes de questions-réponses et la typologie utilisée par des systèmes de recherche d'information.

Bien que les nombreuses typologies définies soient d'une complexité variée, nous constatons que les catégories nom de personne, nom de lieux et nom d'organisation de type noms propres (ENAMEX) sont bien présentes dans toutes les typologies présentées. En effet, comme le montre le tableau des caractéristiques des différentes compagnes d'évaluation (Nouvel, 2012) du tableau 3, ces dernières contiennent des catégories et des sous-catégories qui dépendent principalement de la langue, de la nature du corpus à annoter et des besoins applicatifs visés.

[45] ETAPE (*Évaluations en Traitement Automatique de la Parole*) a pour objectif de mesurer les performances des technologies vocales appliquées à l'analyse des flux télévisés en langue française.

[46] Quaero est un programme de recherche réalisé en collaboration entre 32 partenaires publics et privés sur l'analyse automatique et l'enrichissement de contenus numériques, multimédias et multilingues. Il a permis le développement d'un guide d'annotation des entités nommées sur lequel la compagne ETAPE s'est appuyée.

Les entités nommées sont généralement associées aux feuilles de la typologie mais l'organisation existante entre les différents types, qui peut être de nature hiérarchique ou de nature ontologique, mise en place grâce à la formalisation des relations entre les différents types, permettent d'associer une entité nommée à plusieurs nœuds de la typologie.

Date	Campagne	Langue, modalité	Types	Métriques
1996	MUC-6	anglais écrit, rapports	pers, org, loc	f-mesure
1997	MUC-7	anglais écrit, journalistique	pers, org, loc, date, heure, montant, pourcent	f-mesure
1997	MET-1	espagnol, chinois et japonais, écrit journalistique	pers, org, loc, date, heure, montant, pourcent	f-mesure
1998	MET-2	chinois et japonais, écrit journalistique	pers, org, loc, date, heure, montant, pourcent	f-mesure
1999	IREX	japonais, écrit journalistique	pers, org, loc, artefact, date, heure, montant, pourcent	f-mesure
2002	CoNLL-2002	espagnol et flamand, écrit journalistique	pers, org, loc, misc	f-mesure
2003	CoNLL-2003	anglais et allemand, écrit journalistique	pers, org, loc, misc	f-mesure
2006	HAREM	portugais, écrit journalistique	pers, org, loc, temps, œuvre, événement, abstraction, chose, valeur, autre	pondération d'erreurs
2006	SIGHAN	chinois, écrit	pers, org, loc, entité géo-politique	f-mesure
2007	ACE07	anglais, arabe et chinois, écrit journalistique et conversationnel	pers, org, loc, bâtiments, entité géo-politique, armes, véhicules	pondération d'erreurs
2007	EVALITA 2007	italien, écrit journalistique	pers, org, loc, entité géo-politique	f-mesure
2008	ACE08	anglais et arabe, écrit journalistique et conversationnel	pers, org, loc, bâtiments, entité géo-politique	pondération d'erreurs
2009	ESTER2	français, oral journalistique	pers, org, loc, temps, montant, fonction, produit	SER
2011	EVALITA 2011	italien, oral journalistique	pers, org, loc, entité géo-politique	f-mesure
2012	ETAPE	français, oral journalistique et conversationnel	pers, org, loc, temps, montant, fonction, produit	SER

Tableau 3. Caractéristiques des principales campagnes d'évaluation

4.3. Définition des entités nommées

Les entités nommées correspondent à un ensemble d'expressions linguistiques diverses définies d'une manière pragmatique lors des conférences MUC tels que les noms de personnes, les noms des lieux, les noms organisations, les dates, etc. Elles sont considérées comme des informations pertinentes permettant une meilleure représentation, voir une compréhension des contenus textuels. Avec l'importante expansion des données langagières, leur reconnaissance devient indispensable afin de répondre à des besoins applicatifs variés qui visent non seulement la recherche et l'extraction d'informations mais aussi une compréhension automatique des textes. Nous constatons que l'ensemble des expressions

linguistiques correspondant aux entités nommées sont issues de considérations pragmatiques en énumérant une liste des objets mentaux du « monde » qui dépendent fortement de la nature de l'information langagière à traiter et des besoins applicatifs auxquels ils répondent. En effet, l'ensemble des entités nommées évolue d'une campagne d'évaluation à une autre en intégrant plusieurs changements sur la typologie et le processus d'annotation des phénomènes rencontrés. La typologie correspond à la détermination de la catégorie et relève d'une prise en compte de la réalité du monde tandis que l'annotation correspond à la détermination de conventions d'annotations et relève de la prise en compte du texte (Nouvel et al., 2015). La typologie et les directives d'annotation sont donc définies par chaque compagne d'évaluation créant ainsi une dynamique permettant la prise en compte des travaux antérieurs, des nouveaux phénomènes rencontrés, qui dépendent généralement du corpus et de la langue, et des applications pour lesquelles la REN est une composante principale.

A notre connaissance aucune réflexion théorique sur la nature linguistique des entités nommées n'a été effectuée lors de la conférence MUC-6 et aucune définition standard n'a été proposée. Les quelques tentatives en vue de définir les entités nommées sont peu nombreuses et les énoncés définitoires ont évolué ces dernières années pour inclure diverses notions linguistiques. Les formules définitoires les plus récentes servent des théories du sens qui cherchent à établir un lien entre le signe linguistique et le référent, et des théories du processus applicatif qui s'intéressent à modéliser des mécanismes de reconnaissance d'une information textuelle, comme un socle commun acquis à toutes les entités nommées (Nouvel et al., 2015). Un travail que nous considérons comme pionnier sur le sujet est celui d'Eherman (2008) dans lequel une étude approfondie des théories linguistiques a été effectuée et une définition des entités nommées a été proposée : « Étant donné un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » Eherman considère que le cadre applicatif constitue le contexte nécessaire pour la définition d'un ensemble de types d'information formant les entités nommées. De ce fait une telle définition permet de mettre en valeur la nature mouvante des entités nommées puisque ces dernières ne forment pas un ensemble stable de types d'informations mais plutôt un ensemble d'expressions linguistiques qui dépend fortement de la nature de l'information à traiter, ce qui correspond au « corpus », et des besoins applicatifs visés ce qui correspond au « modèle applicatif ».

Notons que les entités nommées sont constituées essentiellement des noms propres, de descriptions définies et d'expressions numériques et temporelles. Nous pouvons donc conclure qu'elles sont des expressions hétérogènes et de catégories syntaxiques diverses qui

figurent au sein des syntagmes nominaux. Selon Nouvel et al. (2015) le fonctionnement référentiel est l'aspect commun, stable et constant à toutes les entités nommées. Ils supposent l'existence d'un mécanisme de désignation stable permettant l'établissement d'un lien entre l'expression linguistique (entité nommée) et un référent des mondes possibles. Ce mécanisme repose autour de deux notions clés, la référence à un particulier et l'autonomie référentielle. Par conséquent, une entité nommée est une expression linguistique qui a la capacité de renvoyer à un référent unique (référence à un particulier) et qui peut, par ses seules ressources évoquer un référent (autonomie référentielle). Nouvel (2012) propose une formule définitoire qui met en valeur le processus applicatif : les entités nommées « désignent des objets mentaux de manière stable, à partir desquels il est attendu qu'une représentation logique opère ».

4.4. Problèmes liés à la reconnaissance des entités nommées

Les entités nommées forment un ensemble non exhaustif d'expressions linguistiques qui ont la particularité de désigner d'une manière rigide des objets mentaux. En effet, plusieurs expressions linguistiques peuvent satisfaire cette définition et la typologie des entités nommées peut être élargie pour intégrer diverses expressions selon les besoins applicatifs. Par exemple, dans un contexte industriel, il est possible d'ajouter de nouvelles catégories d'entités nommées afin d'extraire des informations qui font référence à des produits, des marques et des vendeurs. Ceci nous amène au problème de la détermination des catégories qui a été abordé par Eherman (2008). De plus, chaque type d'entité nommée est de nature ouverte, c'est à dire que la liste des expressions linguistiques qui correspondent à un type d'entité nommée n'est pas finie, à cause de l'apparition de nouvelles expressions liée à des facteurs extralinguistiques.

Linguistiquement, les entités nommées sont des entités « polysémiques ». En effet, plusieurs référents peuvent être associés à une même entité, on parle d'une pluralité de référents. De ce fait, quelques phénomènes linguistiques posent des difficultés pour la connaissance des entités nommées :

- **Homonymie :** Une même entité nommée peut désigner plusieurs référents.
- **Synonymie :** Plusieurs entités nommées peuvent désigner un même référent.
- **Métonymie :** Une entité nommée peut désigner un référent par un autre.

La caractérisation des phénomènes linguistiques de polysémie des entités nommées permet d'étudier les différents cas ambiguïtés. En effet, la désambiguïsation est une phase indispensable dans le processus de reconnaissance des entités nommées qui consiste à

déterminer le référent qui correspond à l'entité nommée en question. L'élaboration de cette tâche nécessite généralement le développement de ressources linguistiques diverses permettant l'analyse de la structure et du contexte d'apparition d'une expression linguistique.

4.5. Approches pour la reconnaissance des entités nommées REN

La tâche de reconnaissance des entités nommées consiste à identifier les séquences d'unité atomique du vocabulaire (ALU) qui correspondent aux entités nommées et leur assigner la catégorie correspondante d'une typologie prédéfinie. Différentes méthodes ont été mises en place afin d'effectuer cette tâche en utilisant diverses techniques que nous classons en deux approches à savoir : approche orientée données et approche orientée connaissances. L'approche orientée données repose sur des algorithmes d'apprentissage automatique qui permettent, à partir d'un corpus annoté manuellement ou semi-automatiquement, de construire un modèle capable d'identifier les séquences d'ALU qui correspondent à des entités nommées et de les classer dans des catégories de la typologie construite généralement automatiquement à l'aide des différentes annotations utilisées dans le corpus de référence. Tandis que l'approche orientée connaissances s'appuie sur diverses ressources linguistiques principalement des lexiques et des règles de réécritures élaborées par un expert afin d'effectuer plusieurs niveaux d'analyse linguistique généralement de l'analyse morphologique à l'analyse sémantique dans le but de repérer l'entité nommée au sein d'un texte et de la désambiguïser en lui associant une catégorie d'une typologie prédéfinie manuellement. Il existe des méthodes dites « hybrides » qui combinent des techniques issues des deux approches (Mikheev et al., 1998). En effet, ces méthodes cherchent le bon compromis entre les résultats du modèle résultant d'un processus d'apprentissage automatique et les résultats d'un processus d'application des ressources linguistiques construites manuellement par un expert. Nous ne dressons pas un panorama de toutes les méthodes de reconnaissance d'entités nommées, nous renvoyons pour cela aux travaux de (Friburger, 2002) traitant le sujet en profondeur. Cependant, nous présentons les méthodes les plus utilisées.

4.4.1 Approche orientée données

Les méthodes orientées données consistent, à partir d'un corpus de référence annoté manuellement ou semi-automatiquement et en utilisant un algorithme d'apprentissage automatique, à construire automatiquement un modèle probabiliste capable d'extraire les entités nommées. Les algorithmes utilisés cherchent à trouver la distribution de probabilité $P(y|x)$, où y est l'ensemble des entités nommées qui sont assignées à x, et x est une séquence

d'unités lexicales. En effet, le calcul d'une telle distribution n'est possible que si les séquences d'ALU, qui correspondent aux entités nommées, ont été étiquetées a priori. En premier lieu, une méthode de détermination des frontières des entités nommées doit être appliquée. Cette méthode consiste donc à estimer si un ensemble d'ALU contigus forment une entité nommée. Or, comme le montre l'exemple illustré par la figure ci-dessous, en utilisant le format *Begin-Inside-Outside (BIO)*[47], le problème revient à classer chaque ALU dans une des trois catégories *Begin*, *Inside* ou *Outside* dont chacune est associée à une catégorie de la typologie d'entités nommées.

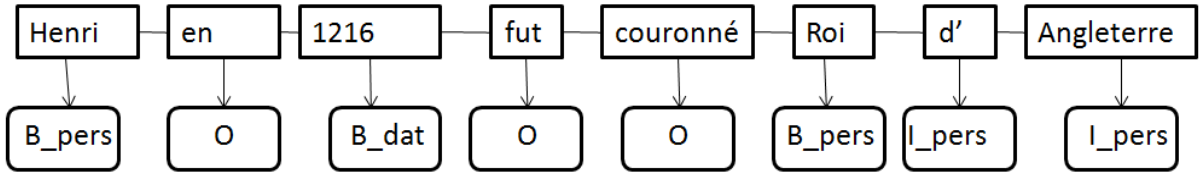


Figure 33. Annotation d'une phrase en utilisant le format BIO

Plusieurs algorithmes ont été utilisés pour réaliser une estimation des probabilités des classes permettant une telle annotation comme les modèles bayésiens (Roth et al., 2002), les machines à vecteurs de support (SVM) (Isozaki et Kazawa 2002), et l'entropie maximale (EM). Comme le montre la figure ci-dessous, diverses informations linguistiques, qui sont considérées comme des indices internes tels que les informations morphologiques, la classe grammaticale, le lemme et les traits sémantiques sont généralement pris en compte afin de garantir une efficacité maximale.

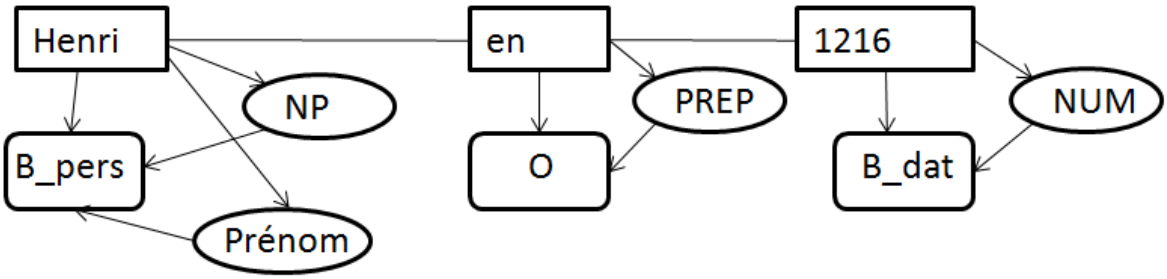


Figure 34. L'utilisation des différentes informations linguistiques associées à l'ALU

Suite à cette phase de détermination des frontières d'entités nommées, plusieurs annotations de format BIO peuvent être associées à une ALU. Afin de lever l'ambiguïté, des

[47] BIO (*Begin-Inside-Outside*) est une annotation ayant trois valeurs à savoir *Begin*, *Inside* et *Outside*. Elle est généralement utilisée pour annoter des séquences d'unités comme les entités nommées en marquant respectivement le début de la séquence, l'appartenance à la séquence et la non-appartenance à la séquence.

modèles séquentiels peuvent être appliqués tels que le modèle de Markov à états cachés (HMM).

Comme nous avons vu pour la désambiguïsation de l'étiquetage morphosyntaxique (section 3.4.1.), le modèle markovien va chercher à trouver la séquence d'étiquettes qui maximise la probabilité locale et la probabilité transitoire en se basant sur des observations du corpus de référence.

$$\arg_{t1,n} \max P(t_{1,n}|w_{1,n}) = \arg_{t1,n} \max(P(w_{1,n}|t_{1,n})P(t_{1,n}))$$

Ensuite, les différentes étapes vues à la section 3.4.1 seront appliquées à savoir la mise en place de l'algorithme de Viterbi en utilisant des techniques de la programmation dynamique et l'application des méthodes de lissage pour résoudre le « problème des données clairsemées » appelé aussi « régularisation ».

La méthode, que nous avons décrite jusque-là, emploie deux modèles probabilistes. L'un effectue un étiquetage a priori en tenant compte des informations locales et l'autre désambiguïse les ALU en prenant en considération la dépendance avec les classes voisines. Or, il existe des modèles qui cherchent à assembler les deux modèles en question. A titre d'exemple le modèle de Markov à entropie maximale (MEMM, Maximum Entropy Markov Models) (McCallum et al., 2000) qui est un modèle graphique pour l'étiquetage des séquences combinant les caractéristiques des modèles de Markov cachés (HMM) et des modèles d'entropie maximale (MaxEnt). Citons également le modèle des champs aléatoires conditionnels (CRF) (Lafferty et al.,2001) qui a démontré une grande efficacité pour la reconnaissance d'entités nommées et pour l'étiquetage morphosyntaxique. En effet, comme le montre le schéma de la figure ci-dessous, les CRF sont considérés comme des modèles Markoviens graphiques non-orientés qui permettent de prendre en compte la séquentialité c'est-à-dire la dépendance des variables « voisines » tout en utilisant la spécificité des différents traits qui sont considérés comme des connaissances redéfinies sous formes de propriétés (*features*).

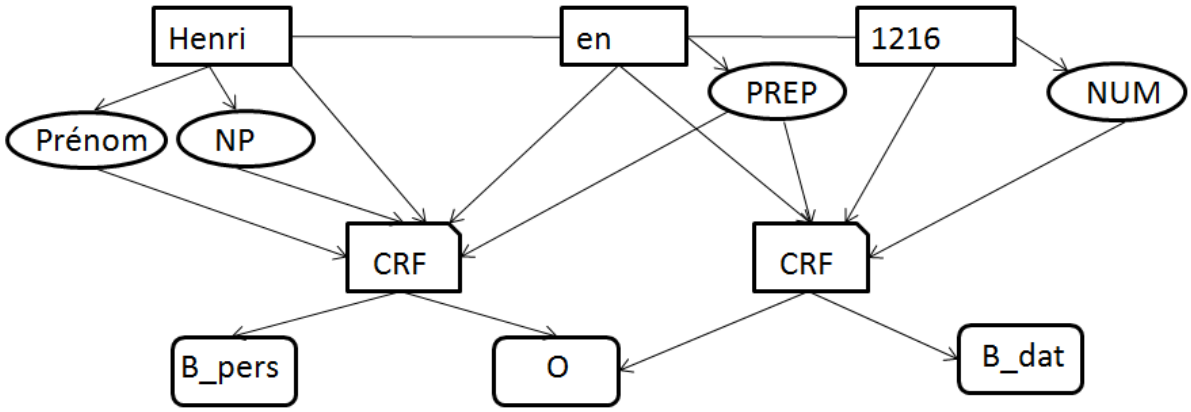


Figure 35. Annotation avec les CRF

4.4.1 Méthodes orientées connaissances

Les méthodes orientées connaissances ont remarquablement évolué et elles sont considérées comme les plus efficaces pour la reconnaissance des entités nommées (Galliano et al. 2009). Comme le montre la figure ci-dessous, ces méthodes effectuent plusieurs niveaux d'analyses linguistiques, de l'analyse morphologique à l'analyse sémantique, des entités et de leurs contextes. En plus des ressources linguistiques utilisées pour l'analyse lexicale et l'analyse morphosyntaxique, ces méthodes nécessitent le développement de ressources spécifiques de natures diverses telles que des règles d'analyse morphologique et des lexiques, qui correspondent aux indices internes et qui permettent l'identification des entités nommées, des lexiques pour recenser des ALU qui correspondent aux indices externes tels que les mots déclencheurs et des lexiques qui correspondent à des listes exhaustives des entités nommées. Toutes ces ressources sont ensuite utilisées par des modèles syntaxiques et sémantiques construits manuellement ou semi-automatiquement par des experts (Coates-Stephens, 1993).

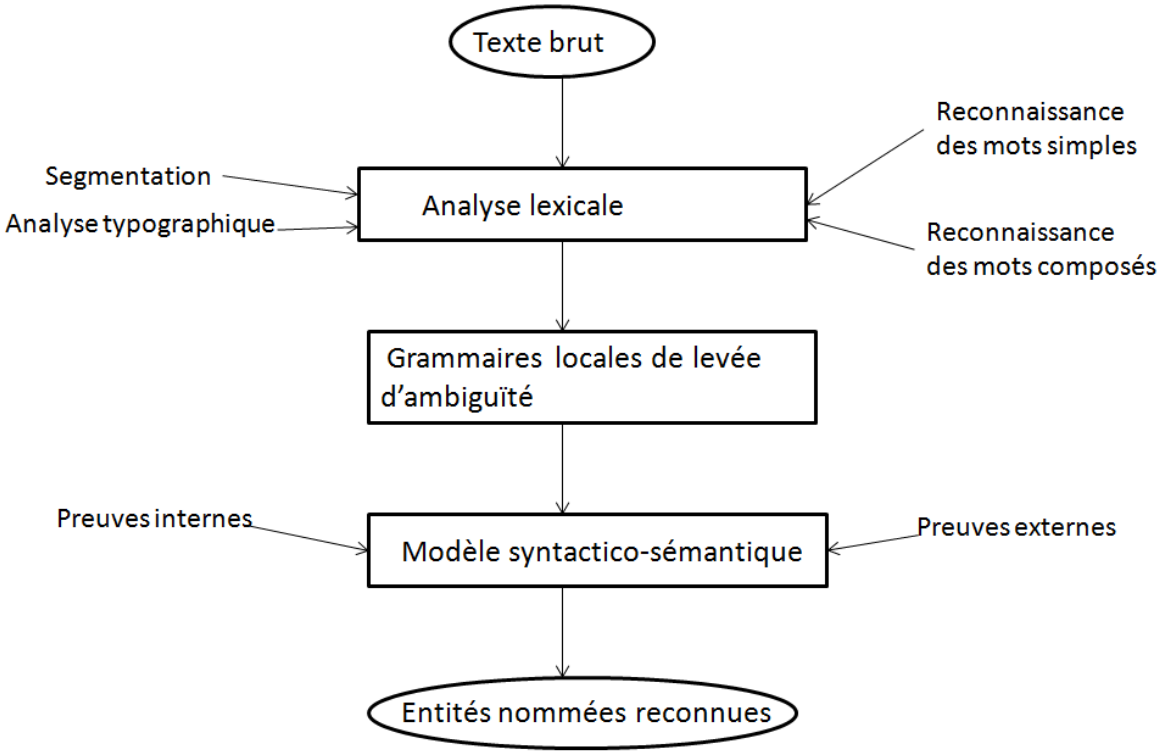


Figure 36. Processus des méthodes symboliques pour la reconnaissance d'entités nommées

Ces modèles peuvent être conçus à l'aide des plateformes linguistiques telle la plateforme NooJ. En effet, l'approche NooJ permet le développement de systèmes de localisation des phénomènes locaux dans le texte et d'attester leur appartenance à des classes grammaticales (Silberztein, 2015). Un tel système peut être mis en place grâce d'une part aux formalismes définis, à savoir les dictionnaires électroniques qui permettent une description exhaustive, riche et précise du vocabulaire d'une langue, et les grammaires locales, qui formalisent un ensemble de règles non-contextuelles qui correspondent aux grammaires hors contextes de type 2 de la hiérarchie de Chomsky. En effet, grâce à l'environnement offert par NooJ, il est possible de développer ces grammaires locales sous forme de graphes et de les appliquer afin d'annoter les séquences d'ALU en ajoutant des informations à la structure d'annotation du texte TAS. Ces informations indiquent l'appartenance des séquences d'ALU annotées à une catégorie de la typologie d'entités nommées. La figure ci-dessous montre un exemple d'une grammaire locale permettant la reconnaissance des noms de personnes réalisée pour un corpus journalistique en langue arabe (Mesfar, 2008). Elle est constituée d'un nœud initial, un nœud final et des nœuds intermédiaires. Ces derniers peuvent contenir des symboles grammaticaux ou des appels à des sous-graphes. Le graphe peut contenir également des étiquettes sous les nœuds permettant ainsi la production des annotations. Cette grammaire est donc constituée d'un ensemble de règles de réécritures permettant l'identification du nom de

personne grâce au sous-graphe « NomPersonne » et la description des contextes qui servent à catégoriser les noms de personnes dans la sous-catégorie « politique ». En effet, pour chaque séquence d'ALU repérée par ce graphe une annotation sera produite et ajoutée à la TAS indiquant le type, la catégorie et la sous-catégorie de l'entité nommée <ENAMEX+PERS+POLIT>.

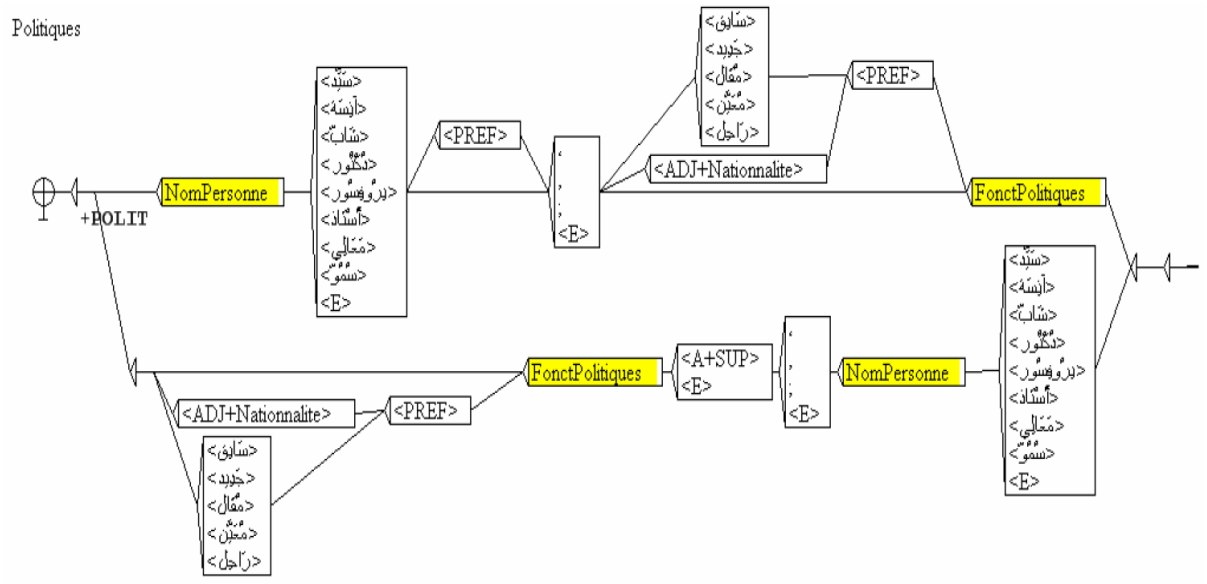


Figure 37. Grammaire locale pour la reconnaissance des noms de personnes

5. Évaluation des systèmes d'annotations automatiques

Un processus d'annotation consiste donc à attribuer à chaque ALU ou une séquence d'ALU une catégorie. L'évaluation de la performance d'un tel système revient à mesurer le nombre d'ALU pertinemment catégorisées parmi l'ensemble des ALUs catégorisées. Etant donné un corpus dit d'*évaluation* ou de *test,* qui devrait essentiellement être différent du corpus de référence utilisé pour la mise en œuvre d'un modèle symbolique ou d'un modèle par apprentissage, l'évaluation d'un système d'annotation automatique est effectuée via une comparaison entre les annotations produites par le système d'annotation automatique et les annotations réalisées par un expert manuellement ou semi-automatiquement.

Il existe plusieurs métriques d'évaluation des systèmes d'annotations automatiques tels que la mesure des taux d'annotations correctes (*accurracy,* en anglais)[48] qui définit le rapport entre le nombre d'ALU annotées convenablement et le nombre total d'ALU (Paroubek &

[48] La mesure des taux des annotations correctes (Accurracy) est une simple mesure statistique qui permet d'estimer les annotations correctes et les erreurs produites par un système d'annotation en calculant le rapport entre les résultats corrects retrouvés et le nombre total des unités du corpus de test.

Rajman, 2000) et les mesures en termes SER (*Slot Error Rate*) qui permet d'évaluer la qualité d'annotation en pondérant les erreurs et qui ont démontré leur efficacité pour l'évaluation des systèmes de reconnaissance d'entités nommées (Makhoul et al. 1999). Le modèle développé par le projet *Cranfield*[49] (Cleverdon et al, 1966) pour l'évaluation des performances des méthodes d'extraction d'information est le plus utilisé pour les systèmes d'annotation automatique des corpus. Les métriques utilisées dans ce modèle sont associées au bruit et au silence documentaire : la précision, le rappel et la F-mesure.

- Rappel (R) est défini par le nombre d'annotations pertinentes retrouvées par rapport au nombre total des annotations pertinentes. Cette métrique, qui s'oppose au silence, vise à mesurer la capacité du système à restituer l'ensemble des annotations pertinentes (Baccini et al, 2010).

R= Nombre d'annotations pertinentes retrouvées/Nombre total des annotations pertinentes

- Précision (P) est définie par le nombre d'annotations pertinentes retrouvées par rapport au nombre total des annotations retrouvées. Elle permet de mesurer la capacité de système à ne restituer que des annotations pertinentes (Baccini et al, 2010). Toutes les annotations non-pertinentes retrouvées constituent le bruit de système.

P = Nombre d'annotations pertinentes retrouvées/Nombre total des annotations retrouvées

- F-Mesure nommée aussi F-score, elle permet de calculer une moyenne combinant la précision et le rappel qui favorise les systèmes qui ont des valeurs de rappel et précision homogènes.

F-Mesure = 2PR/(P+R)

Nous adoptons ces métriques d'évaluation (Rappel, Précision et F-Mesure) qui se limitent à mesurer la pertinence des résultats produits sans prendre en considération d'autres paramètres tels que le temps machine, qui estime le temps d'exécution de système, la robustesse, qui permet d'évaluer le comportement du système pour traiter des données rares, imparfaites ou mal formées et enfin la finesse des catégories utilisées que ce soit celles qui forment le jeu d'étiquettes pour l'étiquetage morphosyntaxique ou celles qui constituent la typologie pour la reconnaissance des entités nommées. Ces paramètres peuvent être d'une

[49] Le projet Cranfield est lancé en 1960. Il a pour objectif de construire des collections de test avec lesquelles il est possible d'évaluer un système. La démarche expérimentale suivie dans le cadre du projet Cranfield a permis de définir des mesures métriques pour l'évaluation d'un système.

importance capitale selon les besoins applicatifs auxquels les systèmes d'annotation automatiques répondent.

6. Conclusion

Parmi les différents types d'informations linguistiques exposés, nous nous sommes intéressés aux systèmes d'annotations automatiques qui enrichissent les données textuelles par des informations grammaticales et par des informations sémantiques. Les informations grammaticales plus précisément les classes grammaticales sont produites grâce à des étiqueteurs morphosyntaxiques. L'étiquetage morphosyntaxique est une tâche indispensable pour les différents besoins applicatifs en TAL. Elle repose sur un formalisme « jeu d'étiquettes » permettant la description des propriétés d'une unité lexicale dans son contexte d'énonciation. Différentes méthodes symboliques ou par apprentissage ont été mises en place pour la réalisation de cette tâche. Ces méthodes ont démontré une grande efficacité en affichant un taux de performance qui dépasse les 95% pour des langues standardisées tels que le français et l'anglais.

Les systèmes d'annotation sémantiques, plus précisément les systèmes de reconnaissance d'entités nommées, ont des performances moins spectaculaires que les systèmes d'étiquetage morphosyntaxique. La définition des entités nommées fait toujours débat. Nous nous sommes référé à la définition proposée par Nouvel (2015) qui repose sur les caractéristiques référentielles des entités nommées. Ensuite, nous avons présenté divers aspects problématiques essentiellement linguistiques liée à la reconnaissance d'entités nommées. La mise en place d'un système de reconnaissance d'entités nommées consiste à les identifier et à leur associer un type. Par conséquent, il faut établir une typologie au sein de laquelle les différentes catégories sont organisées hiérarchiquement ou ontologiquement. La complexité des typologies dépend principalement de la langue, de la nature du corpus à annoter et des besoins applicatifs visés. Enfin, nous avons exposé les deux approches permettant la mise en place d'un système de reconnaissance d'entités nommées à savoir l'approche orientée données et l'approche orientée connaissances. Ces méthodes affichent un taux de performance qui varie entre 75% et 95% pour des corpus constitués principalement des textes journalistiques. Enfin, nous procéderons dans les chapitres à venir à appliquer certaines de ces méthodes à notre corpus en moyen français et nous évaluerons leurs performances en utilisant les métriques proposées dans le cadre du projet Cranfield à savoir le rappel, la précision et la F-mesure.

Chapitre 3

Le moyen français : spécificités et état de l'art

1. Introduction

Les historiens travaillent depuis toujours sur des textes en se focalisant essentiellement sur leurs déchiffrements et leurs éditions pour ensuite étudier leurs contenus en prenant en considération certains aspects stylistiques et/ou sociolinguistiques (Genet, 2010). A l'heure actuelle, certains historiens font de plus en plus appel à des méthodes linguistiques afin de tirer les informations qu'apportent les textes. De fait, l'historien commence récemment à s'intéresser à la linguistique saussurienne et aux nouvelles méthodes numériques d'analyse de données afin de faire parler ces données dans le but, affirmé par Michel Foucault (1966), de « redire ce qui n'a jamais été prononcé ». Il s'agit en effet essentiellement de respecter la source tout en essayant de tirer parti « de ce qu'elle transmet » (Genet, 2010) d'où le besoin actuel de développer des corpus numériques et des outils d'analyse pour les langues anciennes. Beaucoup de projets ayant différents objectifs essentiellement historiques et linguistiques ont été élaborés donnant naissance à des ressources linguistiques diverses et à des nombreux outils d'analyse.

Dans ce chapitre, nous commençons par aborder le contexte historique et la situation de la langue française à l'époque du moyen français. Ensuite, nous exposons des projets TAL qui ont traité des textes en moyen français. Finalement, ayant participé à la constitution du corpus « MEDITEXT », nous avons choisi d'utiliser ce corpus comme corpus de référence pour ce travail et nos travaux à venir.

2. Contexte historique et situation de la langue française

Le moyen français correspond à une variété de langue française ayant existé au moyen âge. C'est parmi les langues d'oïl celle qui est devenue de fait la langue officielle du royaume. À quelle époque a-t-elle existé ? Quels sont les principaux événements qui ont favorisé son expansion ? Et dans quelles circonstances a-t-elle été utilisée et a-t-elle évolué ?

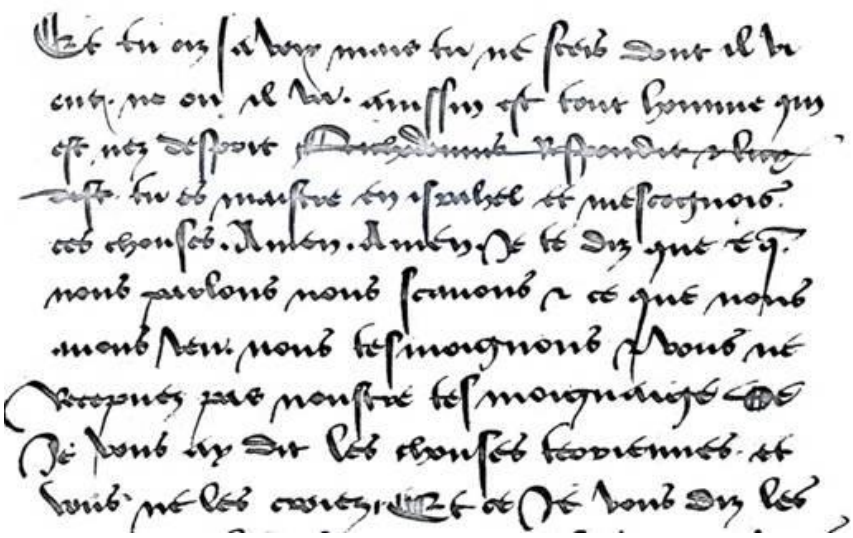


Figure 38. Écriture cursive du XV[e] siècle à Roanne

2.1. Cadre chronologique

Le terme « moyen français » a été employé au 19ème siècle par le philologue Arsène Darmesteter (Darmesteter et al., 1895) afin de distinguer une variété du français qui existait à la fin du moyen âge, de l'« ancien français ». Une étude approfondie sur le sujet a été effectuée en 1913 par le linguiste et philologue Ferdinand Brunot traitant et décrivant le « moyen français » dans le cadre du fameux projet d'écriture de l' « Histoire de la langue française » composée de 7 tomes. Travailler sur le moyen français suppose une capacité à identifier les textes écrits en moyen français. Cependant, vu les frontières très proches entre l'ancien français et le moyen français, et entre le moyen français et le Français Classique, cette tâche n'est pas triviale. Il faut cependant au préalable fixer le cadre chronologique dans lequel ces textes ont été écrits. Or, la période qui couvre le moyen français fait toujours débat entre les historiens des langues. Brunot définit la période de moyen français comme suit : « L'âge de moyen français est l'âge où la veille langue se détruit, où la langue moderne se forme. Il s'ouvre peu après l'avènement des Valois, et ne se ferme qu'après celui des Bourbons. » Tout en sachant que l'avènement des Valois date de 1328 et que le début des Bourbons est aux environs de 1589, cet intervalle de temps nous semble défini implicitement et borné par deux dates bien éloignées. Pierre Guiraud dans son livre consacré au moyen français en 1963 relie la période de moyen français à deux événements significatifs, le commencement de la guerre de Cent Ans (1337) et la fin des guerres de Religion (1598). Johan Huizinga (1967) dans son livre « Le déclin de moyen âge » retient une période relativement proche de celle de Guiraud. En effet, il associe l'âge du moyen français à une période qui commence au début du XIVème siècle et finit au XVIème siècle à la fin de la

Renaissance. D'autres travaux plus récents, même s'ils optent pour une date différente pour la fin de l'âge du moyen français, sont en accord avec Guiraud pour le faire débuter au XIVème. Nous faisons référence aux travaux de C. Marchello-Nizia qui a fait ce même choix dans son livre intitulé « La langue française aux XIVème et XVème siècles » et au « Dictionnaire du moyen français (DMF) » (Souvay, 2004) réalisé par l'unité CNRS ATILF à l'université du Lorraine à partir du corpus « Frantext moyen français » constitué de 219 textes de la période 1330-1500. C'est presque la même période – début du XIVème à la fin du XVème – qui a été choisie pour la constitution des deux corpus en moyen français, l'un dans le cadre du projet Modéliser le changement : les voies du français (MCVF) réalisé à l'université d'Ottawa (Martineau, 2009), l'autre dans le cadre du projet la base de Français Médiéval (BFM) réalisé à l'École Normale Supérieure de Lettres et Sciences Humaines de Lyon (Guillot et al., 2007). Pour nos travaux, nous disposons du corpus « MEDITEXT », que nous présenterons plus loin, constitué principalement de textes politiques en moyen français couvrant la période de la fin du XIIIème jusqu'au fin du XVème siècle. En effet, nous considérons que le premier mouvement de traduction déclenché par le roi Philipe IV le Bel (1285-1314) marque le commencement du moyen français. Cette opération de traduction a été développée et renforcée par les rois qui lui succèdent Jean II et Charles V. En effet, les premiers textes traduits en français à la fin du XIIIème siècle font partie des textes largement lus et utilisés aux XIVème et XVème siècles. Nous citons à titre d'exemple « Li livre du gouvernement des rois et des princes » par Henri de Gauchy qui est une traduction du « *De regimine principum* » de Gilles de Rome; « Le Livre de Boece de consilacion » par Jean de Meun qui est une traduction de la « *Consolatio Philosophiae* » de Boèce ; Le « Miroir des dames » par un Cordelier qui est une traduction du « *Speculum dominarum* » de Durand de Champagne et le « *Passaige de la Terre Sainte* » par un anonyme qui est une traduction du « *Liber sancti passagii christocolarum … Terre Sancte* », dit le *Directorium*, de Galvano di Leanto.

2.2. Contexte historique

La période du moyen français était une période agitée de l'histoire de France. En effet, suite au décès sans descendant direct du Charles IV en 1328, Edouard III roi d'Angleterre et petit-fils de Philippe Le Bel conteste le titre de roi de France à Philippe VI de Valois qui a été sacré roi de France en 1328, fondant la dynastie des Valois. Dès lors les rois des deux royaumes se disputent la couronne jusqu'en 1453. C'est la guerre de Cent Ans (1337-1453) pendant laquelle les français perdirent Calais, suite à une lourde défaite à Crécy en 1346. S'ajoute à tout cela une révolte dans la capitale menée par Etienne Marcel qui était vouée à

l'échec. Vers 1355, le fils d'Edouard III, surnommé le Prince Noir, ravage le Midi et bat le roi de France Jean II le Bon à Poitiers. Les conséquences sociales et politiques de cette première phase de la guerre de Cent Ans entrainent les deux royaumes dans une crise des monarchies. L'économie a été très affectée par l'impact de la guerre. A la même période plus précisément de 1346 à 1352, une épidémie de peste noire frappe l'Europe faisant des milliers de victimes. Rapidement, la guerre et la peste inscrivent le royaume dans un contexte de déclin démographique. La population française en 1450 est à peine la moitié de ce qu'elle était avant 1348. A cause de l'impôt et de la guerre, les campagnes connaissent des mouvements migratoires vers les villes. Victime de cette situation, la seigneurie essaie de restaurer le servage mais se retrouve face à une forte résistance paysanne qui a considérablement affaibli le pouvoir seigneurial. L'impôt, les tentatives seigneuriales pour restaurer le servage et l'hostilité qui ne cesse de s'accroître à l'égard des gens de guerre et la noblesse, font éclater des révoltes populaires violentes. Ferdinand Brunot a décrit cette période ainsi : « *Vers le milieu du XIVe siècle, les pires fléaux, l'invasion, la guerre civile, la peste désolent à la fois la France, qui tombe dans un état effroyable d'anarchie et de misère.* ».

Cette crise économique et sociale a impliqué un développement considérable de la société et des cultures commerciales. L'effondrement démographique a favorisé une meilleure alimentation ; les seigneurs et les paysans se sont orientés vers l'élevage ; les productions techniques métallurgiques et chimiques se sont développées ; les villes ont pris une importance croissante et ont vu de grandes campagnes de construction de fortification et enfin, les mines ont connu une importante extension.

Même après que Charles V, avec l'aide d'un chef de guerre Bertrand du Guesclin, ait battu le Prince Noir, il voit arriver les révoltes populaires à Paris (1380), Nîmes (1378), Montpellier (1379) et Béziers (1381) causées par la guerre, la peste et la famine. Il abolit les fouages et renforce l'ouvrage militaire afin de redresser le royaume. A sa mort, il laisse son héritage à son fils Charles VI âgé de douze ans et entouré de ses oncles qui ont besoin d'argent pour calmer la Flandre révoltée et conquérir le royaume de Naples. Les révoltes populaires pour la suppression de l'impôt et la redistribution des terres et des pensions continuent. A cause de cette crise économique et sociale, la noblesse en perte de vitesse, s'engage dans l'entreprise de la guerre et prend une place importante surtout avec la généralisation de l'armée contractuelle qui conduit à l'organisation d'une armée professionnelle. Cette situation a favorisé le système féodal. Les princes et les grands nobles disposent de clientèles militaires et de leur hôtel, qui est une institution qui gère le pouvoir et

le prestige. Le pouvoir des princes s'accroit au détriment du pouvoir royal et mène à l'établissement à leur profit de principautés autonomes et indépendantes.

Philippe le Hardi, duc de Bourgogne, reprend la situation en main en restaurant l'autorité de roi de France à Paris, à Rouen et en Flandre. De plus, il arrange le mariage de Charles VI avec Isabeau de Bavière pour conforter son alliance avec les Wittelsbach contre les Luxembourg. Puis, Charles VI lance des réformes avec une baisse de la fiscalité et la reconstitution du gouvernement. Mais ce dernier en 1392 a été emporté par une folie furieuse, laissant le royaume plonger dans une guerre civile entre les Bourguignons et les Armagnacs suite à l'opposition entre les ducs d'Orléans et de Bourgogne marquée par l'assassinat du frère du roi, le duc Louis d'Orléans, par Jean sans Peur.

Henri V profite de cette situation critique du royaume pour reprendre la guerre de Cent Ans afin de gagner les provinces promises au traité de Calais. Il envahit la France et bat l'armée française à Azincourt. Il prend donc Calais et entame la conquête de la Normandie qui se termine en 1419 par la capitulation de Rouen. Désireux de devenir roi de France, Henri V fait signer le traité de Troyes à Charles VI, traité qui prévoit de déshériter le dauphin Charles [VII] et de reconnaitre que la couronne de France revient à Henri V époux de sa fille Catherine de Valois et à ses héritiers. Mais Henri V meurt en août 1422, deux mois avant Charles VI. Le frère d'Henri V, duc de Bedford, s'installe à Paris pour exercer la régence de France pour le compte de son neveu Henri VI âgé de neuf mois. C'est la double monarchie, Bedford conforte sa place en remportant les victoires de Cravant (1423) et de Verneuil (1425).

Alors que l'armée de Charles VII est impuissante face aux Anglais pendant la « bataille des harengs », Jeanne d'Arc, une jeune paysanne lorraine, se donne pour mission de faire couronner le dauphin âgé de 22 ans. Elle rejoint l'armée de secours, délivre Orléans et écrase l'arrière-garde anglaise à Patay. Elle a joué un rôle majeur pour rassembler l'opinion publique en faveur du sacre de Charles VII qui a été célébré le 18 juillet 1429. En effet, dans la situation d'instabilité politique, économique et sociale du royaume, Jeanne d'Arc représente la volonté populaire de changement, ce « *sentiment national qui s'attache à un certain nombre de figurations, à une symbolique royale qui s'est lentement élaborée pendant des siècles* » (Marchello-Nizia, 2005). Cette période difficile a donc permis les changements importants qui ont conduit à la restauration de la puissance de la monarchie et à l'unification du royaume. Elle est ainsi considérée comme une période de préparation et de transition tant pour le système politique et social que pour l'esprit public. Le pouvoir royal, à travers les différents rois qui se succèdent, se développe au détriment des pouvoirs seigneuriaux et ecclésiastiques.

L'administration royale centralisée se renforce et l'appareil judiciaire ainsi que la justice royale connaissent un immense développement.

En 1450, Charles VII reprend l'Aquitaine après la bataille de Castillon, et en 1453, la Normandie après la bataille de Formigny. Et enfin, le traité de Picquigny est signé en 1475 annonçant la fin de la guerre de Cent Ans. La fin de cette période se conclut avec Louis XI (1461-1483) qui a unifié le royaume en lui rattachant plusieurs provinces et a mis en place un pouvoir royal centralisé et stable. Ferdinand Brunot a résumé cette période dans un esprit typique des élites du conservatrices du XIXème siècle ainsi : « *l'esprit public change, un nouvel idéal social, moral, intellectuel, commence à naître, déjà très net pour quelques-uns. Aussi sont-ce le XIVe siècle, et ceux qui le suivent qui pourraient avec raison être appelés des siècles de moyen âge ; intermédiaires entre les temps féodaux qui finissent et les temps modernes qui commencent, ils sont à la fois un temps de décadence et un temps de préparation* ». On pourrait aussi qualifier le XIVème siècle de siècle révolutionnaire à tous points de vue.

2.3. Situation de la langue française

La situation linguistique de la France au Moyen Âge est souvent qualifiée de complexe, à cause de l'existence de nombreuses langues dont le français central, les dialectes régionaux, le latin, l'occitan et le franco-provençal, comme le montre la figure 39.

Une distinction s'impose entre l'oral et l'écrit régional pour comprendre la situation linguistique au nord de la France. Cette distinction bien présente dans plusieurs travaux (Remacle, 1948, Delbouille, 1939, Gossen, 1968, Marchello-Nizia, 2005) a été décrite par C. Marchello-Nizia (2005) : « *La scripta régionale* [est une] *langue écrite plus ou moins fortement colorée de traits dialectaux, mais restant lisible cependant dans tout le domaine de langue française ; Et le dialecte parlé, le parler local tel qu'on devait le pratiquer parallèlement, auquel nous ne pouvons avoir accès, mais dont on peut supposer qu'il possédait des caractères nettement plus marqués que la scripta correspondante, puisqu'on sait que la communication entre locuteurs de provinces différentes était fort difficile.* ».

En d'autres termes, bien que les langues régionales parlées au Nord soient suffisamment différentes pour que l'intercompréhension entre elles soient difficile, les *scriptae* régionales étaient compréhensibles par tous les habitants des régions de langue d'oïl. En effet, cette distinction met en valeur la différence entre *scriptae* régionales unifiées et les parlers locaux différenciés (Lusignan, 2012). Par conséquent, les textes en *scriptae* régionales favorisent la communication en mettant plus l'accent sur des traits communs que sur les traits locaux qui

affirment la singularité locale. Malgré cette uniformisation des parlers en écrit, il est important de signaler que ce sont les documents écrits en *scriptae* régionales qui constituent les preuves d'existence des parlers régionaux.

Ce phénomène lié à une évolution historique et géographique de la langue française au Moyen Âge est bien connu par les linguistes qui étudient les « contacts de langues ». Ils l'appellent « inférence linguistique », considérant le français central de Paris comme une « langue véhiculaire » servant de moyen de communication entre des populations de parlers maternels différents. De ce fait, la langue écrite était différente de toutes les langues parlées au nord de la France et elle contenait donc des traits communs et des normes conventionnelles la rendant compréhensible par tous les habitants du Nord (Delbouille, 1939). Le français central de Paris est une langue dont on constate la présence et l'influence des traits dans les textes des différentes provinces du Nord. Ces traits s'affichent donc comme un moyen pour l'uniformisation de l'écriture bien qu'elle ne concerne qu'une faible partie de la population, les lettrés. Selon C. Marchello-Nizia (2005), ce fond des traits communs existant entre les différentes *scriptae* régionales du nord s'est développé à partir de la langue de la capitale Paris, appelée « francien » qui a existé depuis le XIème siècle jusqu'à la fin du XIIIème siècle. Si les traits du francien se sont étendus progressivement dans les régions de langue d'oïl depuis le début du XIIIème siècle, cela est dû au rôle qu'a joué Paris comme capitale incontestable du royaume depuis Philippe Auguste. En effet, comme le montre Outi Merisalo (1988), la chancellerie a joué un rôle majeur dans le développement de l'écriture qui était une affaire de professionnels au Moyen Âge et qui a conduit à la stabilisation de la langue vernaculaire écrite dans les villes du Nord dès le XIIIème siècle. Serge Lusignan (2012) en conclut que celui qui écrit en français à cette époque est l'héritier de traditions d'écriture qui varient en espace et en temps ; par conséquent, cette standardisation du français médiéval est faite dans la différence. Et il ajoute que le français était en contact au Nord avec le latin et le picard. Toujours selon Lusignan (2012), la *scripta* picarde, qui correspond au ressort de la nation picarde de l'Université de Paris, s'est imposée dans le système scolaire au point que les écoles qui apprenaient à lire et écrire en français pour les enfants utilisaient la *scripta* picarde. Cette découverte a été rendue possible grâce à une méthode d'histoire sociolinguistique considérant selon Lucien Febvre (1952) que « la langue étant le fait social par excellence ». Elle est capitale et change notre vision de l'histoire de la langue française au Moyen Âge qui s'était développée jusque-là comme une expansion du français de Paris et comme le résultat de son acceptation progressive comme la langue standard. De plus, « l'écrit vernaculaire

s'inspira du latin depuis l'alphabet, le tracé des lettres et le système d'abréviation, jusqu'à la mise en page des chartes autant que des codices » (Lusignan, 2012).

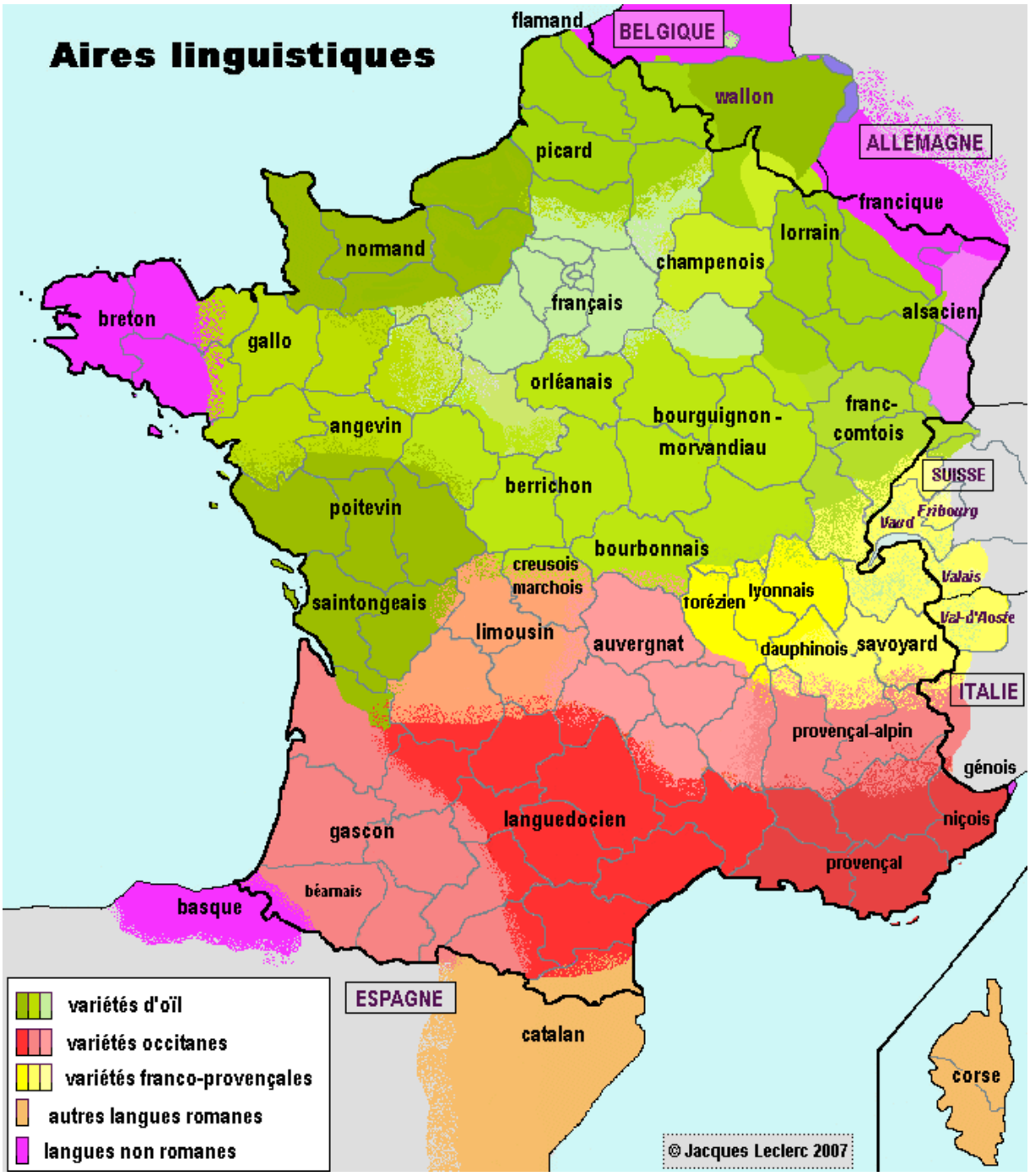


Figure 39. Cartographie des dialectes en France au XIVème siècle

L'expansion progressive du français central au XIVème siècle est souvent expliquée par des raisons politiques. Une analyse approfondie des actes royaux du XIVème siècle en France effectuée par Serge Lusignan a permis la mise en place d'une périodisation qui se rapporte à la stratégie quasi-officielle de la royauté en fonction du contexte politique. Comme le montre la figure ci-dessous, il distingue trois stratégies langagières différentes de 1315 à 1360 selon

le roi ou le prince, le contexte, la géographie et la période. En effet, jusqu'à 1330, le latin est à la fois la langue liturgique de l'église catholique et la langue de l'élite utilisée dans l'administration, la justice et l'éducation. Il est utilisé pour les actes dans les régions de langue d'oïl avec environ 77% des actes et dans les régions de langue d'oc avec environ 95%. Sous Charles IV (1322-1328) par exemple, le latin représente environ 90% des chartes. Dans cette période, l'utilisation du français se limite aux nobles de la région de langue d'oïl. Selon Hong (2017), sur les 57 actes français qui ont été envoyés par la royauté, 54 sont expédiés par Louis X (26) et Philippe V (28) pour réaffirmer les privilèges des nobles et des villes et convoquer les assemblées. Cependant les actes destinés au clergé des régions de langue d'oïl sont écrits en latin comme ceux destinés aux gens des régions de langue d'oc.

Le règne de Philippe VI est lié à l'essor de l'utilisation du français et un emploi différencié de la langue selon les régions surtout dans les actes juridiques et les actes des grandes chancelleries de 1330 à 1350. Durant cette période, Philippe VI a procédé pour la première fois à l'instauration du français comme langue royale. En effet, le latin restait la langue des juridictions ecclésiastiques, mais ces dernières ont vu, leurs pouvoirs se restreindre au profit des juridictions royales dont le personnel a utilisé, dès le XIIIème siècle, un latin compréhensible et relativement proche de la langue vulgaire. Par conséquent le passage vers le français au Nord a été facile à réaliser.

C'est dans les régions de langues d'oïl que le français a commencé à supplanter le latin. Les actes français ont donc atteint les 72% sous Philippe VI dans les régions de langue d'oïl. En plus des nobles et du peuple, le français est employé dans les actes destinés aux clercs dans les régions de langue d'oïl. Il est par conséquent la langue la plus employée dans les actes destinés au Nord. En revanche, le latin garde son hégémonie dans les régions de langue d'oc qui ont elles aussi reçu des actes français à destination des nobles et des clercs. La stratégie langagière de Philippe VI est ainsi basée sur l'aspect géographique, il utilise donc le français surtout pour les régions de langue d'oïl comme signe de reconnaissance de la royauté. Pour les pays de langue d'oc, le français était considéré comme la langue d'expression politique que partageaient la royauté et la société politique.

La période de 1350 à 1360 se caractérise par une oscillation entre le français et le latin. Avant la campagne de Poitiers en 1356 le français représente environ 84% des actes royaux. Le règne de Jean II se distingue par un retour à l'utilisation du latin comme langue de l'administration royale. La période suivante 1356-1360 est la période du règne de Charles V qui est marquée par l'essor du français avec 49% des actes royaux. Pour les régions de langue d'oc, le latin est préféré pour exprimer la rémission et la reconnaissance. En outre, le français

est beaucoup plus employé dans les actes relevant des affaires de l'État (guerre et impôt) y compris pour les pays de langue d'oc (Hong, 2017).

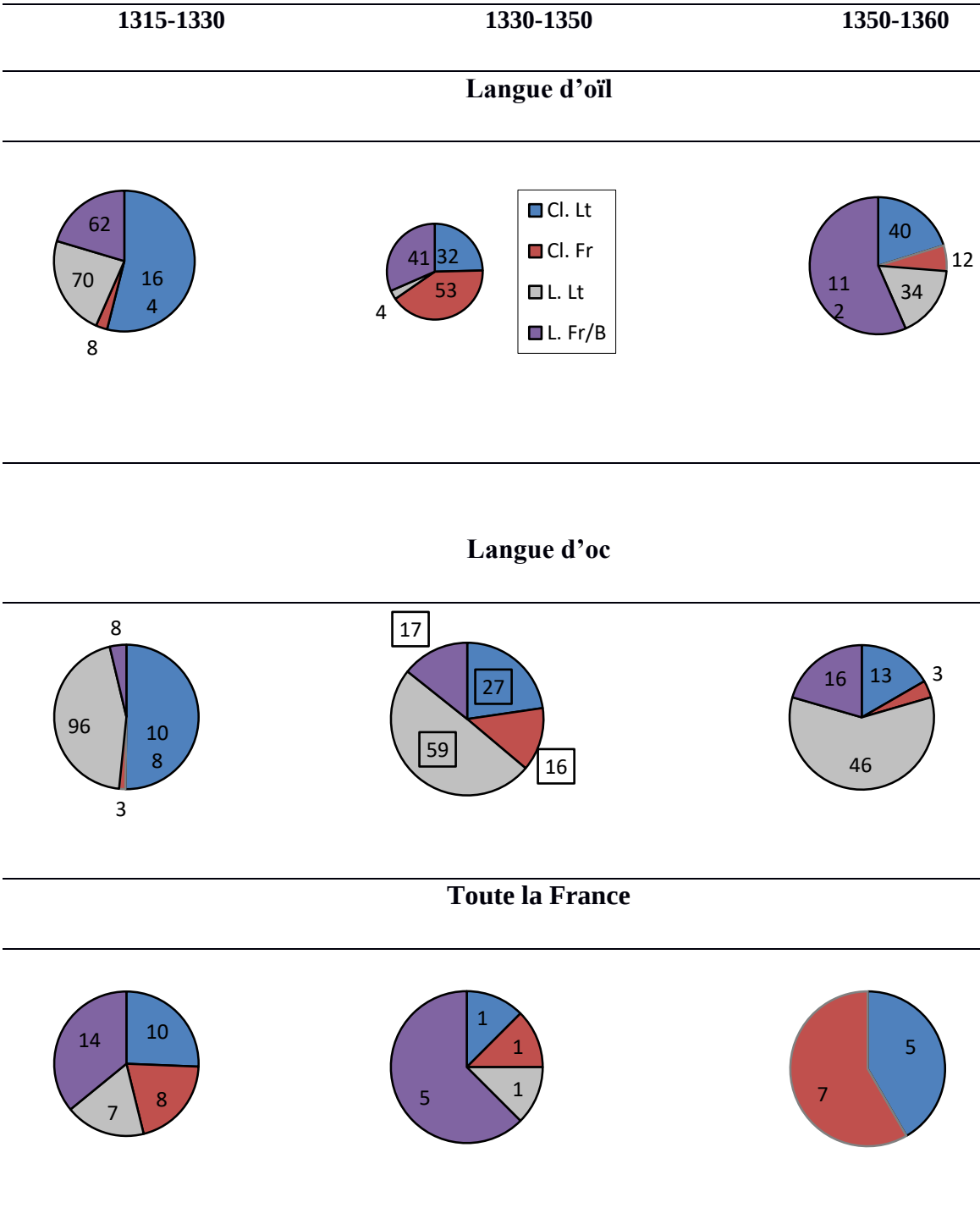


Figure 40. Langues des actes selon le statut social et la région des destinataires (1315 à 1360)[50]

[50] Diagramme extrait du (Hong, 2017)

En effet, des changements politiques majeurs, à savoir l'indépendance du pouvoir du souverain par rapport à l'église et le renforcement comme le développement d'une administration et d'un système judiciaire solides et centralisés, ont consisté à développer le pouvoir monarchique en restreignant les pouvoirs seigneuriaux et ecclésiastiques, et ils ont favorisé l'utilisation du français au détriment du latin.

Dans le domaine franco-provençal, la situation linguistique était bien différente de celle du Nord. Les langues écrites régionales ou les *scriptae* locales étaient bien éloignées du français. Par conséquent, deux types de *scriptae* locales ont été utilisés pour la rédaction des documents officiels. Des *scriptae* largement influencées par le franco-provençal étaient utilisées pour la rédaction des documents destinés à la population locale. Et des *scriptae* moins marquées par les traits dialectaux du franco-provençal contenaient quelques éléments de la langue française commune afin d'être compris par la population nordique et l'administration royale.

Dans le domaine occitan, la situation linguistique est encore plus complexe. Le français était considéré comme une langue étrangère qui n'a été utilisé dans les documents officiels que dans la seconde moitié du XVème siècle, après la fin de la guerre de Cent Ans. Pendant longtemps en effet le pouvoir royal a utilisé le latin et les langues locales pour la rédaction des documents officiels au Sud. La progression du français a été lente et il a fallu attendre l'ordonnance de Villers-Cotterêts prise par le roi François Ier imposant l'usage du français dans les actes officiels et de justice en 1539 pour que celui-ci se substitue à la langue d'oc.

En Angleterre, le français est la langue dominante de l'administration royale depuis la fin du XIIIème siècle jusqu'au début du XVème siècle. Le français est même longtemps considéré comme la langue de la *common law* jusqu'à l'époque moderne.

3. Spécificité du moyen français

Certains linguistes considèrent que le moyen français est une langue marquée par l'irrégularité et l'incohérence (Luce et al., 1858). En effet, comme toute langue non-standardisée ou non-uniformisée, le moyen français est caractérisé par l'instabilité des phénomènes linguistiques et des traits régionnaux. Il est ainsi considéré par Luce et al. (1858), Brunot (1913) et Guiraud (1963) comme une langue de transition ou comme une « période de genèse du français moderne » (Guiraud, 1963) voire « une étape intermédiaire entre l'ancien

français et le français moderne » (Brunot, 1913). En effet, la langue est dynamique par définition contenant des phénomènes qui évoluent et d'autres qui restent stables. Cependant, ces hypothèses, qui mettent en valeur l'aspect transitoire du moyen français, justifient leur validité par la coexistence de différents systèmes linguistiques.

Les études sur le moyen français visent à rechercher les régularités de son fonctionnement en produisant la description la plus cohérente et la plus complète dans plusieurs champs de la linguistique comme la morphologie, la morphosyntaxe, la distinction entre graphèmes et phonèmes, la structuration des différents types de syntagmes et l'ordre des mots des propositions.

Le moyen français est une langue en pleine évolution dont l'orthographe, le système flexionnel et la syntaxe ne sont pas stables. Il se singularise principalement par l'absence d'orthographe normalisée. En effet, nous constatons une importante variation orthographique. Par exemple, l'unité lexicale « *seigneur* » possède en français moderne deux formes : la forme masculine singulière « *seigneur* » et la forme masculine plurielle « *seigneurs* ». En moyen français, comme l'illustre le graphe de la production morphologique de la figure 41, une trentaine de variantes orthographiques attestées ont été associées à l'entrée « seigneur » tels « *seigour* », « *seignur* », « *seingeur* » et « *sengor* ».

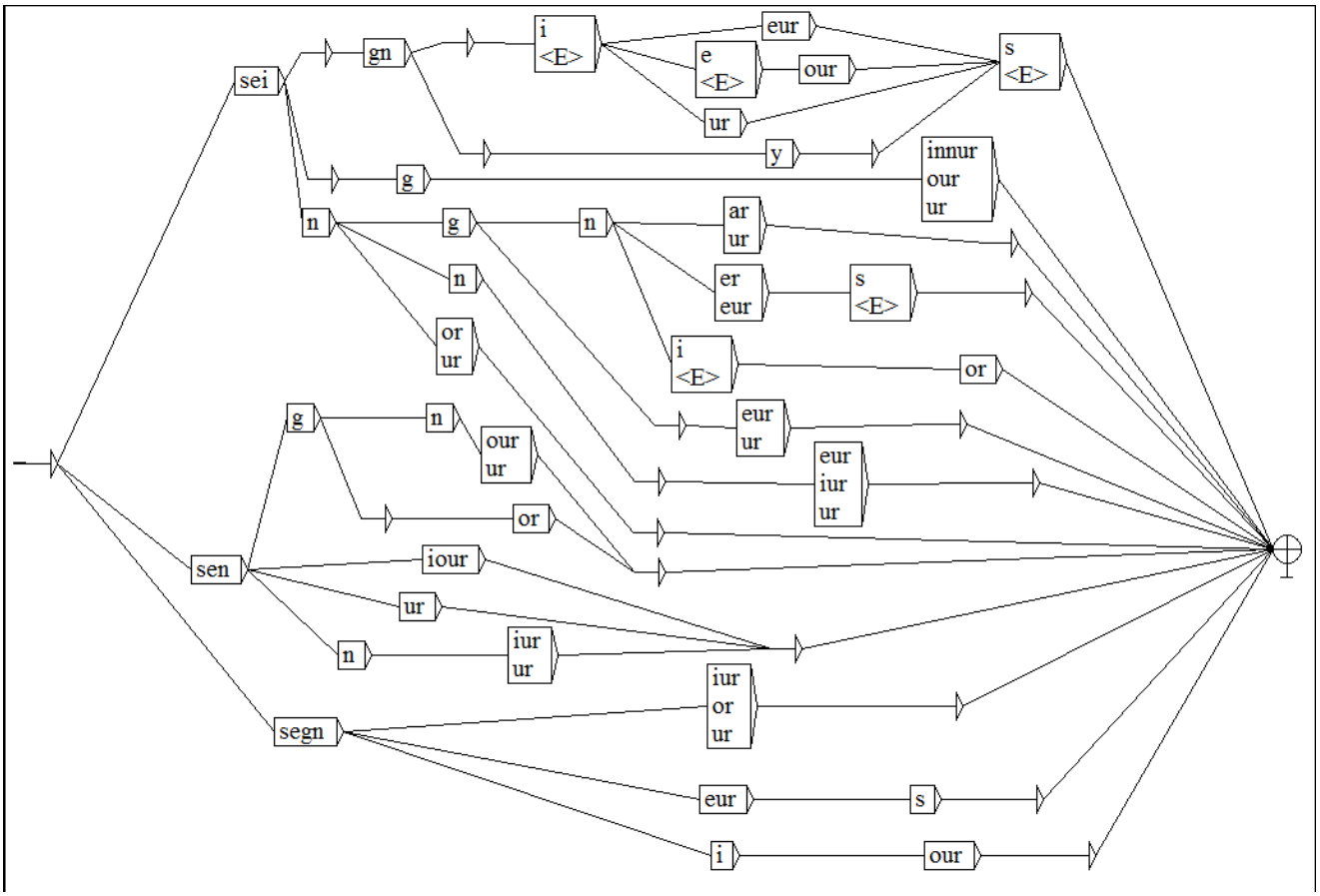


Figure 41. Graphe de production morphologique des variantes de « seigneur »

Cette liste des variantes produites par la grammaire de production morphologique de l'entrée lexicale « *seigneur* » ou encore la liste de variantes de l'entrée lexicale « *roi* » présentée ci-dessous n'est pas exhaustive. Elle correspond seulement aux variantes attestées par notre corpus. En d'autres termes, il se peut que d'autres variantes aient été employées dans des textes que nous n'avons pas analysés. En effet, comme le souligne Souvay & Pierrel (2009), « on se rend bien compte qu'il est difficile voire impossible d'établir une liste exhaustive des formes possibles pour une entrée lexicale ».

roi,NC+SENS=1
rai,roi,NC+SENS=1
re,roi,NC+SENS=1
ree,roi,NC+SENS=1
rei,roi,NC+SENS=1
reis,roi,NC+SENS=1
rey,roi,NC+SENS=1
roe,roi,NC+SENS=1
roi,roi,NC+SENS=1
roie,roi,NC+SENS=1
rois,roi,NC+SENS=1
roix,roi,NC+SENS=1
roy,roi,NC+SENS=1
roys,roi,NC+SENS=1
royz,roi,NC+SENS=1

Comme le français était la langue habituelle du pouvoir royal au XIVème siècle, l'administration préparait avec soin les textes adressés aux habitants des différentes régions du royaume qui n'étaient pas forcément familiarisés avec le français. De fait, les auteurs et les copistes ont développé des pratiques, qui ont fait évoluer considérablement le système orthographique, afin d'éviter des confusions entre les formes, voire entre les lettres. Par exemple, ils rajoutaient des consonnes non prononcées : *briefue, debuoir, enuieulx, moult, peult, subiect* afin de séparer les I et U consonnes des I et U voyelles, ils introduisaient la lettre H à l'initiale des formes entre la lettre U et la lettre V afin de distinguer des formes telles *huis* de *vis* et *huitre* de *vitre* et ils utilisaient des graphies latines tels que *compter* à côté de *conter* pour distinguer des homonymes, ou *corps* au lieu de *cors* et *sepmaine* au lieu de *semaine* pour se rapprocher de l'étymologie latine.

Le système orthographique a connu plusieurs changements par rapport à l'ancien français comme la réduction progressive de l'emploi des diphtongues dont l'utilisation devient limitée à quelques formes comme *cœur*, *fleur* et *œil,* l'emploi ornemental de certaines lettres : V pour U en début de mot, Y pour I en fin de mot, et l'utilisation de lettres en abréviations de deux autres lettres Z pour TS et X pour US. De plus, le latin et les différentes langues régionales ont largement influencé l'orthographe française. Par exemple, la forme *chiel* est utilisée pour *ciel* dans les textes picards et la forme *bastoun* pour *baston* dans les textes anglo-normands.

Le système flexionnel était en pleine évolution mais on constate une stabilisation de certains phénomènes. Par exemple, l'emploi de suffixes réguliers pour conjuguer les verbes et l'emploi courant de S, X ou Z pour le pluriel des substantifs tels que *ciels, cielz* et *cielx* pour *ciel*.

Comme le français est une langue parlée, plusieurs variantes sont dues au système phonétique. En effet, les phonèmes peuvent changer de graphies selon le contexte sonore ou selon leur position dans le mot. Aussi, ils sont marqués par une réduction des hiatus par exemple OU au lieu de AOU et OI remplace ËOI. Finalement et afin de s'approcher de la prononciation des locuteurs, nous constatons l'introduction de la cédille pour distinguer la lettre C prononcée K de celle C prononcée S et l'introduction des accents tels « *à* », « *â* », « *ê* », « *ô* » pour distinguer d'autres prononciations.

Le moyen français se caractérise par la variabilité tant géographique que chronologique de son lexique. Certaines formes sont utilisées plus que d'autres selon la période et la région. Le lexique est ainsi très influencé par les langues régionales et le latin. Au XIVème siècle, les lettrés, très fidèles au latin, utilisaient massivement des emprunts et des graphies latines dont on constate la présence pratiquement dans tous les textes de notre corpus. En effet, le mouvement de traduction des œuvres latines et grecques, l'apparition des nouvelles formes littéraires et des nouveaux concepts liés au contexte social et politique de l'époque ont permis l'introduction de nouveaux mots au français souvent calqués sur le latin. Si le néologisme fondé sur le latin est le plus courant, il existe d'autres phénomènes comme ceux qui se basent sur la morphologie dérivationnelle en ajoutant des préfixes et des suffixes à des radicaux existants ou comme la « relatinisation » (Gougenheim, 1959) qui consiste à rapprocher un terme de son étymologie latine c'est-à-dire sa signification la plus courante en latin en ignorant son sens en ancien français. Pour conclure, plusieurs mots français ont disparu du lexique, surtout les termes ayant des traits régionaux, laissant place à des graphies latines.

Le moyen français est marqué par une nette évolution de la syntaxe. En effet, l'ordre des mots commence à se stabiliser en sujet suivi d'un verbe, lui-même suivi d'un complément. En outre, les deux cas de la déclinaison existants en ancien français commencent parallèlement à disparaître et l'emploi des prépositions et des conjonctions est en nette augmentation au fil de temps rendant les phrases de plus en plus longues et complexes. Notons que certaines structures syntaxiques sont plus utilisées que d'autres selon la période et la région.

4. Des projets TAL traitant le moyen français

Dans divers domaines, de nombreux projets de recherche ont été lancés pour constituer et analyser divers corpus en langue française contemporaine produisant des ressources linguistiques et de nombreux outils permettant l'analyse des différents niveaux linguistique de la langue. Les projets concernant les langues anciennes sont moins nombreux notamment pour la langue française à la période médiévale. Le moyen français, en particulier, est souvent traité

dans le cadre des projets qui cherchent à étudier l'évolution du français au Moyen Âge. Par la suite, nous présentons les principaux projets qui ont conduit à constituer des corpus numériques en moyen français ainsi que les outils permettant leur analyse.

4.1. Le Nouveau Corpus d'Amsterdam

« Le Nouveau Corpus d'Amsterdam » (Prévost & Stein, 2012) est un corpus annoté par une description morphosyntaxique et un lemme. Il est le résultat d'une collaboration entre plusieurs chercheurs ayant divers compétences et intérêts. Son origine remonte aux travaux d'Anthonij Dees en 1971 sur l'évolution des démonstratifs en Ancien et moyen français au cours desquels il a constitué un corpus de chartes pour une analyse quantitative d'un grand volume de textes. Dans le but de généraliser l'approche quantitative de Dees, son corpus a été utilisé pour la constitution d'un corpus de référence pour les travaux linguistiques en ancien français. C'est ainsi qu'en utilisant les critères de sélection formulé par Carolus-Barré en 1964, Pieter van Reenen a retenu plusieurs textes recueillis par Dees, plus précisément les chartes originales datées et localisées. Ensuite, pour compléter le corpus, des chartes publiées ont été ajoutées avec comme date limite 1300. Fort de son succès, ce corpus a constitué la base de l'Atlas des formes et des constructions des chartes françaises du 13e siècle (Dees et al. 1980). Un nombre considérable de textes littéraires qui ont été analysés par Dees ont, eux, donné naissance à l'Atlas des formes linguistiques des textes littéraires de l'ancien français (Dees et al. 1987), et ont été ajoutés au corpus des chartes afin d'avoir un lexique plus riche.

Jusque-là, le corpus était constitué de textes bruts. Pour annoter les différents textes du corpus, diverses ressources lexicales et textuelles illustrées dans le tableau suivant ont été unifiées afin de construire un dictionnaire des formes contenant un jeu d'étiquettes composées de 50 descriptions morphosyntaxiques et des lemmes sélectionnés à partir de différentes ressources lexicales. Chaque entrée du dictionnaire est donc composée d'une forme, d'une partie du discours et d'un lemme.

Ressource	code	formes graphiques	informations
Tobler/Lommatzsch	T	58727	lemme, variantes
graphies verbales (Martin)	M	71265	catégorie, lemme
graphies du Godefroy	G	115498	catégorie, lemme
mots grammaticaux	S	3999	catégorie, lemme
lemmatisation manuelle	Z	4456	catégorie, lemme
index	I	41377	catégorie, lemme
chartes de l'Aube	C	4260	catégorie, lemme
Amsterdam Corpus	A	133894	catégorie

Tableau 4. Les ressources lexicales utilisées pour la construction du dictionnaire

Le dictionnaire des formes construites a été utilisé pour annoter manuellement un corpus en utilisant 255 étiquettes composées d'une partie de discours et des traits morphologiques. Ensuite, un étiqueteur probabiliste à savoir « *TreeTagger* » a été testé afin d'annoter morphosyntaxiquement les corpus. Enfin, un processus de balisage automatique de corpus annoté a été mis en place.

Comme montre la fiche technique du tableau 5, le corpus contient environ 300 textes et extraits de textes, 3,3 millions de mots, annotés morphologiquement et lemmatisés.

titre / référence bibliographique :	*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-D. Gleßgen
lieu et date de publication :	Stuttgart : Institut für Linguistik/Romanistik première version publiée en février 2006
site de téléchargement :	http://www.uni-stuttgart.de/lingrom/stein/corpus/ (accès réservé aux chercheurs licenciés)
format	texte avec balisage XML
annotation	parties du discours, catégories flexionnelles, lemmes
époque	environ 1150–1350
taille	299 textes et extraits de textes, 3.184.834 mots
genre	textes littéraires, dont 57 en prose (contenant 29% des mots)

Tableau 5. Fiche technique du Nouveau Corpus d'Amsterdam

4.2. Base de Français Médiéval (BFM)

Suite aux travaux sur la syntaxe de l'ancien français et du moyen français de C. Marchello-Nizia dans les années 1980, le besoin d'un corpus de textes représentatif de la période médiévale se fait sentir. Résultat, en 1989, la Base de Français Médiéval (BFM) a été créée comme un complément du Dictionnaire de moyen français (DMF). Elle contenait principalement des textes numérisés à partir d'éditions de référence antérieures à 1320.

D'autres textes postérieurs en moyen français ont été ajoutés par la suite. Actuellement, la base est constituée de 153 textes en Ancien et en moyen français.

Cette base, qui couvre une importante aire géographique et une longue période allant du IXe siècle à la fin du XVe siècle, est utilisée par nombreux chercheurs comme un corpus représentatif de l'histoire du français. En effet, ce corpus a été constitué en respectant deux aspects indispensables de la représentativité à savoir l'aspect quantitatif, en numérisant des volumes importants des données langagières et l'aspect qualitatif, en prenant en considération la diversité des genres textuels et les différents types de textes tant en prose qu'en vers. Ce corpus a été enrichi par l'ajout de nouveaux textes tout en garantissant sa diversité.

Depuis sa création, la BFM a comme principal objectif d'offrir aux chercheurs des ressources linguistiques et des outils permettant de faciliter la recherche sur le français et son histoire. À cet effet, un étiqueteur morphosyntaxique probabiliste « *Treetagger* » a été entrainé sur un corpus d'apprentissage, qui est considéré comme un échantillon représentatif de la base, afin d'annoter les différents types et genres des textes existants. Dans ce contexte, un jeu d'étiquettes a été élaboré « Cattex 2009 » afin de définir pour chaque forme une description morphosyntaxique constituée d'une partie de discours et d'un type. Ce formalisme a donc été utilisé pour l'annotation automatique de tous les textes de la base. Ensuite, une vérification manuelle s'est imposée pour garantir une qualité maximale des annotations morphosyntaxiques.

De plus, des outils ont été développés pour la mise en place d'une méthode permettant de structurer les textes de la base en suivant les recommandations de la TEI comme montre la figure ci-dessous. Un ensemble de balises de la TEI jugées pertinentes a donc été choisi comme standard d'édition électronique afin d'assurer l'interopérabilité des données.

```
<lb n="24"/>
</s>
<q>
<s>
<w type="PONpga" xml:id="w106_003827">«%%</w>
<w type="NOMcom" xml:id="w106_003828">Dame</w>
<w type="NOMcom" xml:id="w106_003829">merveilles</w>
<w type="VERcjg" xml:id="w106_003830">sont</w>
<w type="VERppe" xml:id="w106_003831">avenues</w>
<w type="ADVgen" xml:id="w106_003832">laienz</w>
<w type="PONfrt" xml:id="w106_003833">.%,%.</w>
</s>
<s>
<w type="PONfbl" xml:id="w106_003834">-%%</w>
<w type="ADVint" xml:id="w106_003835">#come<ex>n</ex>t</w>
<w type="PONfbl" xml:id="w106_003836">,%%</w>
<lb n="25"/>
<w type="VERcjg" xml:id="w106_003837">fait</w>
<w type="PROper" xml:id="w106_003838">ele</w>
<w type="PONfbl" xml:id="w106_003839">,%%</w>
<w type="VERcjg" xml:id="w106_003840">di</w>
<w type="PROper" xml:id="w106_003841">le</w>
<w type="PROper" xml:id="w106_003842">moi</w>
<w type="PONfrt" xml:id="w106_003843">.%,%.</w>
</s>
```

Figure 42. « Queste del Saint Graal » structurée et annotée par les outils de BFM

Enfin, comme l'illustre la figure suivante, cette structuration a pu être exploitée par des logiciels tels que les portails d'édition et des logiciels de textométrie comme TXM afin d'avoir une meilleure représentation, lecture et exploration de ces textes.

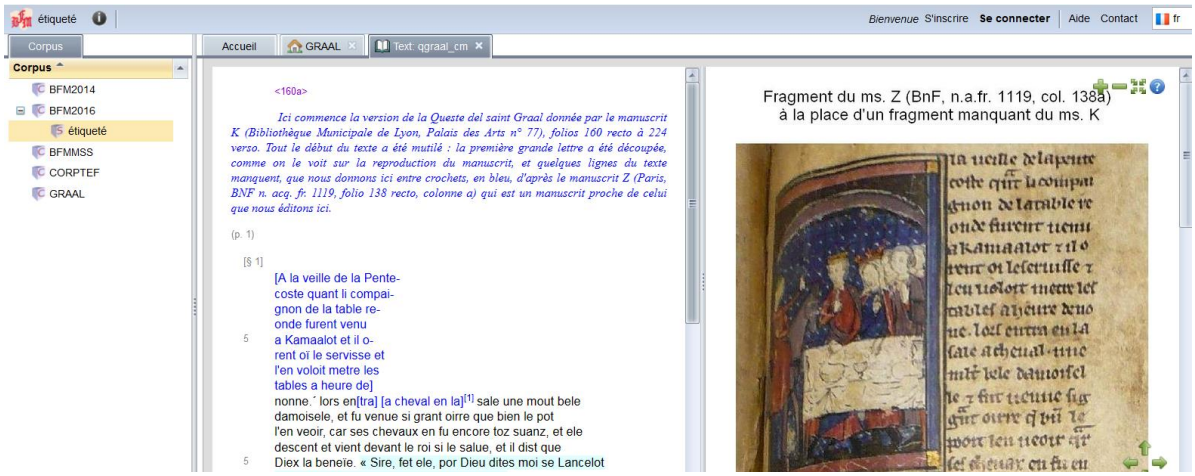


Figure 43. L'édition numérique « Queste del Saint Graal » utilisant le portail de BFM

4.3. Modéliser le changement : les voies du français (MCVF)

Ce projet de recherche, porté par l'université d'Ottawa et dirigé par France Martineau, a comme ambition d'étudier l'évolution du français en partant du constat que « le français s'est toujours développé dans la diversité ». Ainsi les études réalisées cherchent à explorer la diversité des formes du français à travers l'histoire afin de construire un modèle théorique

renouvelé du changement linguistique à partir du moment où le français s'est implanté de façon durable au Canada vers les XVIIème-XVIIIème siècles tout en remontant jusqu'à ses origines au Moyen Âge plus précisément au IXème siècle.

De ce fait, un corpus numérique a été constitué qui couvre plusieurs périodes de l'histoire du français à savoir l'ancien français, le moyen français, le français du XVIème siècle et le français classique. En plus des sources classiques utilisées pour l'étude de l'histoire du français telles que les grands textes littéraires et les discours à propos de la langue, ce corpus a l'avantage d'héberger des textes de la vie sociale tels des archives judiciaires et notariales, des traités et des récits de voyage. Dans ce contexte, le projet contient un volet d'ingénierie du langage incluant plusieurs équipes afin de développer divers outils permettant de structurer les textes d'une façon uniforme en utilisant les recommandations TEI, de les enrichir par des annotations morphologiques et syntaxiques et d'interroger et d'explorer le corpus.

1 2 3 4 5 6 7 8 9 Suivant Total 270 fiches

Auteur	Titre	Date	Genre	Région de l'auteur
	Mémoire sur le projet de constitution	1785-02-28	Document administratif	
Adjudant général des milices	Réponse du bureau de l'adjudant général des mlices à la lettre de P. Careau	1845	Correspondance	
Allain	Lettre d'affaires courantes du prêtre Jean-Louis Allain à Charles Loeder	1838	Correspondance	
Anburey	Journal de voyage de Thomas Anburey en Amérique	1793	Journal / Mémoire personnel	
Angleterre	Lettre d'Élisabeth d'Angleterre	1579-03-09	Correspondance	Angleterre
Anonyme	Abrégé de la vie de Monsieur Olier	1847	Biographie	
Anonyme	Acte de Douai de 1204, et quelques autres textes non datés de la fin du XIIe siècle		Document administratif	
Anonyme	Annonce d'une loterie	1783		
Anonyme	Aucassin et Nicolette		Nouvelle / Fabliau	
Anonyme	Cahier de chansons	1824	Folklore	
Anonyme	Cahier de chansons religieuses en langue autochtone	1847	Folklore	
Anonyme	Cent Nouvelles Nouvelles		Nouvelle / Fabliau	

Figure 44. Exemple de textes du corpus MCVF

4.4. Syntactic Reference Corpus of Medieval French (SRCMF)

Le principal objectif du projet « Syntactic Reference Corpus of Medieval French » *(SRCMF)* consiste à produire un corpus du Français Médiéval enrichi par des annotations syntaxiques. Ce projet est le résultat d'une collaboration entre plusieurs équipes de recherche en France et en Allemagne coordonnée par le Dr S. Prévost (Laboratoire de recherche Lattice, Paris) et par le Professeur A. Stein (Université de Stuttgart). Comme l'illustre la figure 45, des textes de *BFM* et du « Nouveau Corpus d'Amsterdam » ont été annotés afin de produire un corpus de référence annoté syntaxiquement pour le Français Médiéval.

Title	**Date**	**Words**
Serments de Strasbourg	842	117
Sequence de Sainte Eulalie	881	191
Vie de Saint Alexis	vers 1050	4 832
Passion de Clermont	2nde moitié 10e	2 749
Vie Saint Léger	2nde moitié 10e	1 398
Chanson de Roland	vers 1100	29 338
Lapidaire en prose	milieu 12e	4 799
Tristan de Beroul	entre 1165 et 1200	27 257
Yvain de Chretien de Troyes	1177-81	42 103
Quatre Livres des Rois	fin 12e	40 000
Aucassin et Nicolete	fin 12e/déb. 13e	10 009
La Conqueste de Constantinople de R. de Clari	après 1205	33 994
Queste del Saint Graal	vers 1220	40 000
Miracles de G. de Coinci	1218-1227	25 000
Roman de la Rose de J.de Meun	1269-1278	20 000

Figure 45. Liste des textes annotés dans le cadre du projet *SRCMF*

Il s'agit d'élaborer un modèle linguistique à partir des données textuelles dans le but d'étudier certains phénoménes linguistiques. Le modèle syntaxique adopté par le projet est basé sur la dépendance. Certains choix et décisions ont été pris, à savoir :

- Les relations syntaxiques sont centrées sur les formes de verbes fléchis qui en constituent les principaux nœuds.
- Tous les éléments de la proposition, y compris les sujets, dépendent du verbe central.
- Les critères morphosyntaxiques sont prioritaires sur les critères sémantiques, bien que des critères sémantiques soient utilisés dans certains cas.

Un logiciel *NotaBene* (Mazziotta, 2012) a été donc développé dans le cadre du projet afin de faciliter l'annotation syntaxique manuelle de corpus. Ce dernier va servir pour la mise en place d'un analyseur syntaxique automatique pour le Français Médiéval.

Un processus d'annotation a été donc élaboré en utilisant *NotaBene* qui est composé de 4 étapes :

1. Chaque texte est annoté indépendamment par deux experts en utilisant *NotaBene* (figure 46).
2. Le même logiciel est utilisé pour comparer l'annotation des deux experts.
3. Après l'élimination des erreurs évidentes, les différences restantes sont soit soumises à une discussion générale sur le forum du projet, soit directement transmises aux superviseurs du projet pour une décision de leur part.
4. Les choix et les décisions pris en compte ainsi que les discussions autour des différents problèmes confrontés permettront le développement et l'amélioration du modèle linguistique.

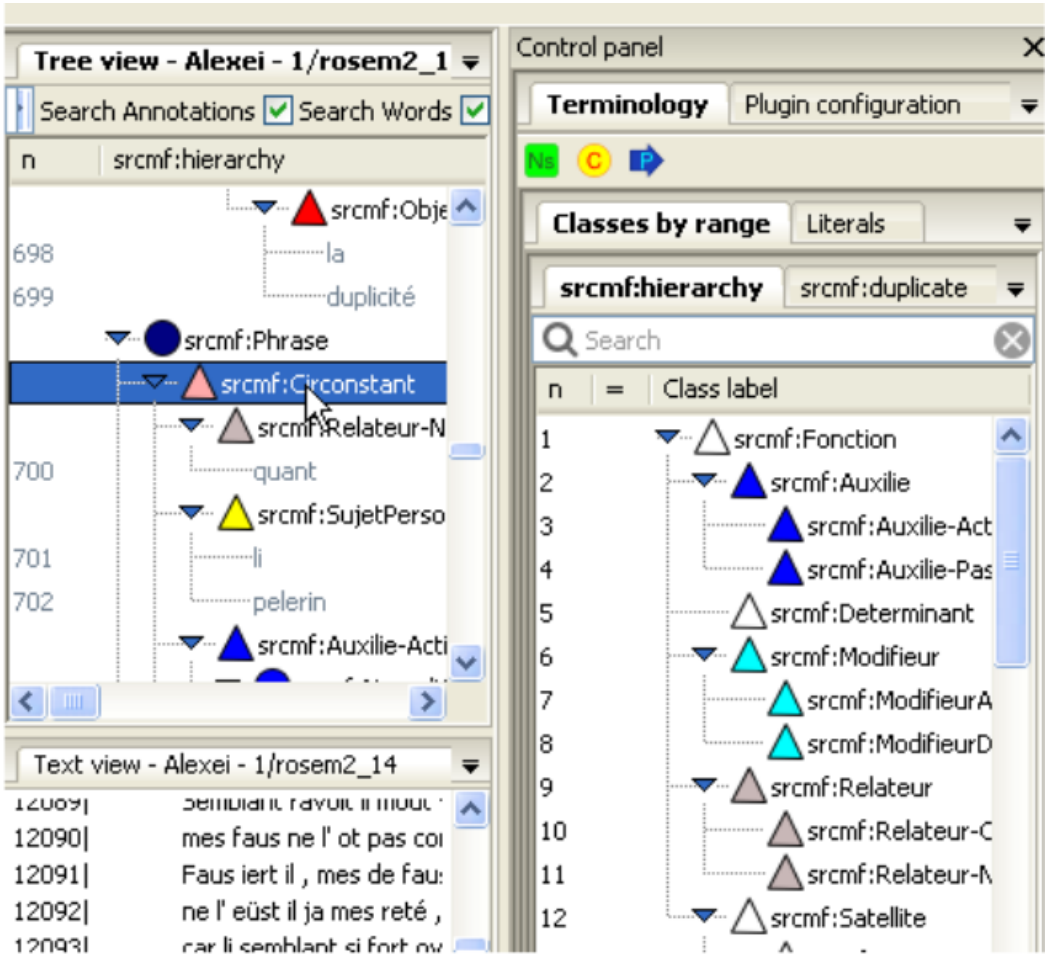


Figure 46. L'environnement « *NotaBene* » d'annotation de *SRCMF*

Techniquement, les annotations syntaxiques produites par *NotaBene* sont exprimées avec le formalisme *RDF* et elles sont stockées dans des fichiers distincts de celui des textes de corpus qui sont structurés en utilisant les recommandations XML-TEI.

Aucun outil n'a été développé dans le cadre de projet *SRCMF* pour permettre l'exploration de corpus. Les membres du projet ont choisi d'exploiter les différentes annotations syntaxiques réalisées en utilisant des outils standards pour lesquels des formats d'exports sont nécessaires. A titre d'exemple, un script a été réalisé permettant la conversion des annotations de *SRCMF* en format *CoNLL* (Carreras & Màrquez, 2005) qui est un standard défini par *Conference on Computational Natural Language Learning.* Aussi, les annotations de *SRCMF* peuvent être exportées en format de *TigerXML* (Lezius et al., 2002) afin de pouvoir interroger le corpus en utilisant le logiciel *TigerSearch* (König et al., 2003). Ce dernier a été intégré au sein de logiciel *TXM* afin de permettre une meilleure exploration et analyse du corpus.

5. Constitution de corpus « MEDITEXT »

La constitution de corpus « Meditext » a été initié au début des années 1980 pour des fins historiques par deux historiens Jean-Philippe Genet et Claude Gauvard du laboratoire des médiévistes occidental de Paris (LAMOP). Confrontés à un processus de numérisation et de production des documents électroniques à la fois lourd et lent, la tâche d'élaboration et d'exploration du corpus était freinée et n'avait pas pu suivre la rapide évolution technologique matérielle et logicielle. En effet, la complexité et la performance du matériel informatique n'étaient pas spectaculaires avec une capacité de stockage limitée et une vitesse de traitement relativement lente. Cependant, nous considérons que le volume de données textuelles construites est important malgré la faible capacité de stockage des supports physiques. Il s'agit donc de mettre en format électronique principalement des textes politiques en moyen français dont certains sont rares et précieux.

En 2010, le programme européen « *Signs and States* » lancé par Jean-Philippe Genet a donné un nouvel élan à « MEDITEXT » dans le but de définir les développements culturels et politiques de la société et d'étudier l'ensemble des transformations qu'ils impliquent du XIVe au XVe siècle. En effet, pendant quatre ans, les textes de « MEDITEXT » ont été rassemblés par les membres du projet dans le but de construire une bibliothèque numérique capable d'opérer une sélection de textes provenant de régions et de périodes différentes. Cette opération devant permettre une extraction d'un vocabulaire le plus varié possible afin d'analyser de manière plus pertinente et plus efficace les textes médiévaux. Par conséquent,

toute une phase de prétraitement a été effectuée afin de nettoyer et de convertir les textes au format TXT en utilisant l'encodage de caractères universel UNICODE. Ensuite, une application en ligne a été mise en place afin de permettre aux historiens de constituer des corpus annotés en trois langues à savoir moyen français, Moyen Anglais et latin. Cette application qui s'appelle PALM « Plateforme d'analyse linguistique médiévale » est considérée comme un véritable réseau entre les médiévistes pour le développement des textes numériques et le partage des textes entre eux. En effet, PALM contient des outils permettant la mise en place d'une méthode de travail pour structurer les textes en XML-TEI et les décrire avec un ensemble de métadonnées afin de conserver pour chaque texte son origine, sa date de création, son auteur, son genre, ainsi que les renseignements sur son édition.

FrP26AL14Fr15	
Titre :	Ordonnance de Charles VI [28 déc. 1388]
vers/prose :	prose
Champs :	Juridique
Type de texte :	Acte ou lettre
Texte lemmatisé :	Non
Periode :	2éme moitié du XIVe siècle
Langue principale :	Français
Pays d'origine :	France
Editions :	Ordonnances des rois de France de la troisième race, vol. 7, Charles VI (1383-1394), D.-Fr. Secousse, Paris, 1745, p. 768.
Saisi par :	Naomi Kanaoka
Dates :	28 décembre 1388
Genres :	Ordonnances
Notes :	423 mots Ancien code: ORD10

Figure 47. Exemple des métadonnées décrivant une ordonnance de Charles VI

Enfin, PALM contient un onglet « Bibliothèque » qui présente la partie du corpus de « MEDITEXT » ouverte au public. Les outils développés ont permis à la base textuelle d'être cumulative et ils ont servi pour enrichir le corpus avec des centaines de textes supplémentaires.

Les textes en français, qui nous intéressent dans le cadre de ce projet, viennent principalement du nord de la France et de l'Angleterre et leurs dates vont du début de XIVème siècle jusqu'au début de XVIème siècle, nous distinguons :

- Des textes ayant trait à des événements politiques identifiés (discours ; lettres ; traités ; poèmes ; sermons ; chroniques) ;

- Des textes consacrés de manière générale au bon et au mauvais gouvernement. Le corpus rassemble également des textes gouvernementaux (proclamations, ordonnances)
- Des textes adressés au roi par ses sujets (cahiers de doléances ; requêtes ; lettres de rémission).

ANGLETERRE	Nombre de fichiers	FRANCE	Nombre de fichiers
Textes politiques	24	Actes et Lettres	48
Discours politiques	90	Textes historiques	28
Sermons politiques	26	Sermons	100
		Sermons politiques	8
Discours des *speakers*	30	Traités et discours en prose	16
Poésie politique	16	Poèmes politiques, moraux ou religieux	18
Vie de Saints	9	Théâtre	20
Total	195 fichiers	Total	220 fichiers

Tableau 6. Un aperçu des textes politiques en français appartenant à « MEDITEXT»

6. Conclusion

Nous considérons que l'âge du moyen français est la période allant de la fin du XIIIème siècle à la fin du XVème siècle. Le moyen français est une langue en pleine évolution dont l'orthographe, le système flexionnel et la syntaxe ne sont pas stables. Il se singularise principalement par l'absence d'orthographe normalisée et par la variabilité tant géographique que chronologique de son lexique.

Divers projets sur l'évolution de la langue française au moyen âge ont été lancés permettant la constitution de corpus numériques en moyen français et le développement d'outils d'annotation automatique principalement des étiqueteurs morphosyntaxiques probabilistes. Le programme européen « *Signs and States* » est l'un des rares projets qui étudie le moyen français à des fins historiques afin de faire parler les données. Il a permis de reprendre la constitution du corpus numérisé « MEDITEXT », qui a été initié dans les années 1980, contenant principalement des textes politiques. Une application en ligne PALM, qui a comme objectif la gestion et le développement d'une version plus évoluée, structurée et annotée de « MEDITEXT » à travers des interfaces adaptées, a été mise en place. Dans ce contexte et dans la continuité du programme SAS, cette étude a pour objectif de décrire le moyen français en concevant des méthodes basées sur une approche descriptive. Ces

méthodes permettent le traitement de plusieurs niveaux d'analyse linguistique afin de générer une annotation automatique fine et précise qui peut être grammaticale en utilisant une méthode d'étiquetage morphosyntaxique ou sémantique en utilisant une méthode de reconnaissance des entités nommées.

Deuxième Partie

Analyse automatique du moyen français

Chapitre 4

Construction du dictionnaire en moyen français

1. Introduction

Les applications en TAL nécessitent l'emploi de ressources linguistiques qui sont des éléments capitaux pour toute analyse automatique. Les dictionnaires électroniques en particulier sont considérés comme une ressource linguistique indispensable dont le formalisme doit être pris en compte pour tout traitement des données textuelles. En effet, ils interviennent à tous les niveaux de l'analyse linguistique, depuis l'analyse lexicale jusqu'à l'analyse pragmatique. A titre d'exemple, l'utilisation du dictionnaire est indispensable pour l'étiquetage morphosyntaxique et la reconnaissance des entités nommées. On appelle donc dictionnaire électronique tout formalisme spécifiquement conçu pour décrire les différents éléments du vocabulaire d'une langue afin d'être exploité par des applications informatiques. Par conséquent, ce formalisme doit permettre l'unification d'une description exhaustive des éléments du vocabulaire et un processus d'exploitation fiable et robuste de ces descriptions.
Les dictionnaires électroniques varient selon l'objectif du projet et les finalités de traitement des données pour lesquels ils ont été conçus. Par conséquent, plusieurs formalismes ont été mis en place afin de répondre à des besoins applicatifs divers parmi lesquels le formalisme des dictionnaires électroniques *DELA*, le réseau lexical *WordNet* et le standard *Lexical Markup Framework* (LMF) qui sont les plus réputés et qui ont donné lieu à d'importantes ressources linguistiques en plusieurs langues. Depuis 2002, un nouveau formalisme des dictionnaires électroniques a été développé dans une optique de formalisation des langues selon l'approche NooJ qui a permis de dépasser les limites des dictionnaires *DELA* et d'assurer une meilleure précision, exhaustivité et intégration de tous les niveaux linguistiques (Silberztein, 2005).

Dans ce chapitre, nous dressons l'état de l'art des formalismes existants permettant la construction des dictionnaires électroniques en nous focalisant sur le formalisme adopté par NooJ. Ensuite, nous présentons le processus mis en œuvre pour la construction d'un dictionnaire électronique en moyen français. Enfin, nous exposons quelques réalisations : la formalisation des entrées de notre dictionnaire et le développement des grammaires morphologiques.

2. Différents types de dictionnaires électroniques

Dans cette partie, nous présentons les principaux formalismes de référence à savoir WordNet, LMF, DELA et NooJ qui ont permis la mise en place des dictionnaires électroniques d'une étonnante ampleur et ont été utilisés par diverses applications d'analyse automatique.

2.1. Le réseau lexical WordNet

WordNet (Kilgarriff, 2000) est une base de données lexicale de la langue anglaise conçue par des linguistes du laboratoire des sciences cognitives de l'université de Princeton[51]. En effet, c'est un formalisme qui forme des réseaux lexicaux en utilisant des ensembles de synonymes cognitifs appelés par *synsets*. Une *synset* est l'unité atomique du système. Elle décrit, à l'instar d'un dictionnaire éditorial, un concept, un sens distinct ou un usage particulier à travers un groupe de mot et fournit des exemples de quelques contextes d'utilisations. Comme l'illustre la figure suivante des différentes *synsets* de l'entrée lexicale polysémique « *book* », de telles descriptions permettent le regroupement des entrées lexicales qui partagent une ou plusieurs *synset*s communes d'une même catégorie grammaticale à savoir des noms, des verbes, des adjectifs et des adverbes.

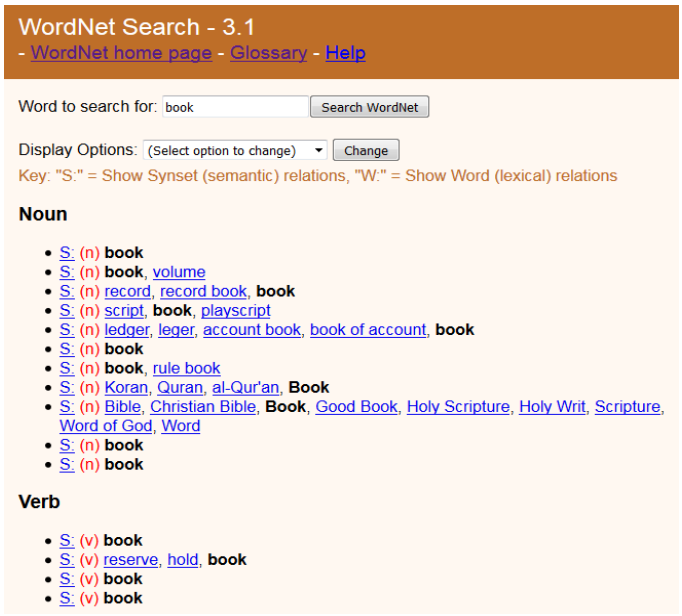


Figure 49. Les différents synsets de l'entrée « book »

[51] Wordnet est lancé par George A. Miller au milieu des années 1980. De plus de la version en ligne, la base est distribuée sous licence libre et la dernière version distribuée date d'avril 2013.

Les *synsets* sont liés entre eux par des relations conceptuelles et sémantiques qui permettent de répertorier, de classifier et de mettre en relation le contenu sémantique et lexical des éléments de la langue. Par conséquent, le réseau résultant des mots et des concepts forme une structure efficace qui peut être considérée à la fois comme un outil de navigation permettant à l'utilisateur de se déplacer dans les différents sens, usage et concepts d'une entrée lexicale, et comme une ressource linguistique supposée utile pour les applications en TAL.

WordNet est aussi artificiellement un thésaurus qui regroupe les mots en fonction de leurs significations. Mais il diffère de ce dernier par le fait qu'il :
- interconnecte les mots selon leurs différents sens. Par conséquent, les entrées lexicales qui se trouvent à proximité les uns des autres dans le réseau sont sémantiquement désambiguïsées.
- annote les relations sémantiques entre les mots, tandis que les groupements de mots dans un thésaurus ne suivent aucun motif explicite autre que la similitude de signification.

En plus de leur utilité pour classer et mettre en relation des entrées lexicales sous forme d'un réseau, d'un groupe ou d'un thésaurus, les *synsets* peuvent aussi représenter des concepts abstraits et formels qui peuvent être organisés ontologiquement. En effet, le système de classification relie les concepts entre eux via des relations conceptuelles-sémantiques permettant ainsi d'élaborer la structure d'une ontologie et de catégoriser le sens des mots au sein des classes de cette dernière. Par conséquent, une ontologie par domaine interrogeable peut être créée dans le but de répondre à certains besoins applicatifs.

Fort de son succès, ce formalisme, qui a été initialement conçu pour la langue anglaise, s'est généralisé pour décrire d'autres langues telles que les langues décrites dans le cadre du projet euroWordNet dont l'objectif est la description de plusieurs langues européennes comme le français, le hollandais, l'italien, l'espagnol, l'allemand, le tchèque et l'estonien.

Si le réseau lexical WordNet est un outil de navigation efficace et utile pour les linguistes, son utilisation pour des analyses automatiques n'est pas toujours fiable. En effet, les *synsets* utilisent des synonymes appelés par « *sister terms* » pour décrire le sens courant comme le cas de « *volume* » synonyme de « *book* » dans notre exemple de la figure 49. Or, ce synonyme n'est pas opérationnel puisqu'il manque l'information dans la base indiquant si les deux termes sont commutables ou non. De plus, les informations existantes dans les *synsets* pour la description des sens ou les informations sémantiques (hyperonyme, hyponyme, etc.) fournies par WordNet pour former des réseaux et/ou des ontologies manquent de structuration

au point qu'elles ont été jugées non suffisamment précises pour être utilisables par des applications d'analyse automatique (Silberztein, 2015).

2.2. Le standard Lexical Markup Framework (LMF)

Lexical Markup Framework (LFM) (Francopoulo et al., 2006) est le standard de l'Organisation internationale de normalisation ISO pour les lexiques du traitement automatique des langues (TAL) destinés aux usages éditoriaux et/ou applicatifs. Dans le but de développer des normes en TAL et des représentations lexicales standardisées, un modèle consensuel a été mis en place grâce à un métalangage défini dans un schéma XML à partir de l'ensemble des descriptions des dictionnaires électroniques les plus réputés. Ce modèle considéré comme une structure commune à tous les dictionnaires électroniques permet donc l'interopérabilité et l'échange des ressources lexicales. Dans une perspective de valorisation de la diversité culturelle et de la communication multilingue qui existent dans le monde numérique, le LMF offre des dispositifs génériques puissants cherchant à couvrir toutes les langues naturelles. Il permet ainsi le développement des ressources monolingues, bilingues et multilingues de l'écrit comme de l'oral en utilisant des structures simples ou complexes. Il supporte la description de plusieurs niveaux d'analyse linguistique à savoir la morphologie, la syntaxe, la sémantique et aux notations multilingues d'une façon extensible et modulaire. Une réflexion approfondie sur la morphologie a été menée afin d'assurer une représentation qui associe une entrée lexicale à un ensemble de paradigmes de flexion et de dérivation.

Comme le montre l'exemple ci-dessous, nous avons modélisé l'entrée lexicale « roi » en moyen français en utilisant le modèle de LMF pour décrire seulement quatre variantes orthographiques. Ce dernier se compose d'une partie noyau obligatoire et de modèles optionnels permettant de décrire le niveau morphologique, syntaxique et sémantique. En effet, des descripteurs linguistiques élémentaires sont modélisés afin de spécifier un ensemble d'informations dites « globales » comme la langue par exemple et d'autres descripteurs linguistiques élémentaires sont associés au lexique comme le jeu d'étiquettes qui peut être composé des parties du discours et d'un ensemble de propriétés linguistiques pour le français par exemple nous citons le genre et le nombre d'un nom commun. Les étiquettes grammaticales ainsi que les valeurs des propriétés prédéfinies sont donc normalisées selon les principes définis par la norme ISO 12620 (Ide & Romary, 2004) et sauvegardées dans un registre de catégories de données qui est consultable et éditable en ligne.

```
<LexicalResource dtdVersion="15">
  <GlobalInformation>
        <feat att="languageCoding" val="ISO 639-3"/>
  </GlobalInformation>
  <Lexicon>
   <feat att="language" val="fr"/>
   <LexicalEntry>
        <feat att="partOfSpeech" val="commonNoun"/>
         <Lemma>
              <feat att="writtenForm" val="roi"/>
        </Lemma>
        <WordForm>
              <feat att="writtenForm" val="roi"/>
              <feat att="grammaticalNumber" val="singular"/>
         </WordForm>
        <WordForm>
              <feat att="writtenForm" val="ree"/>
              <feat att="grammaticalNumber" val="singular"/>
         </WordForm>
         <WordForm>
             <feat att="writtenForm" val="reis"/>
              <feat att="grammaticalNumber" val="plural"/>
         </WordForm>
        <WordForm>
              <feat att="writtenForm" val="rois"/>
              <feat att="grammaticalNumber" val="plural"/>
         </WordForm>
    </LexicalEntry>
  </Lexicon>
</LexicalResource>
```

2.3. Le formalisme DELA

Le système de dictionnaires DELA (Courtois & Silberztein, 1990) a été développé dans les années 1980-2000 au Laboratoire d'automatique documentaire et linguistique LADL à l'Université Paris 7 et il est aujourd'hui considéré comme une référence pour la description du vocabulaire et de sa flexion[52].

Concrètement, c'est un mécanisme qui consiste à associer à chaque entrée lexicale un ensemble d'informations linguistiques sous forme d'une description morphosyntaxique. Ce formalisme est mis en place en utilisant deux types de dictionnaires à savoir le dictionnaire

[52] La version française du système DELA a connu ses évolutions majeures dans les années 80 et 90. Elle contient un lexique des mots simples DELAS, un lexique associé de transcription phonétique DELAP et un lexique des noms composés DELAC.

des mots simples DELAS et le dictionnaire des mots composés DELAC. Grâce au logiciel INTEX[53] (Silberztein, 1993a), une liste des formes simples fléchies DELASF est créé automatiquement à partir de DELAS et une liste des formes fléchies des mots composés DELACF est générée automatiquement à partir de DELAC. En effet, les entrées des dictionnaires DELAS et DELAC contiennent des descriptions morphosyntaxiques qui font appel à des grammaires rationnelles. Ces grammaires sont appliquées aux entrées lexicales correspondantes dans les dictionnaires DELA en utilisant les programmes de génération des formes fléchies d'INTEX afin de générer les dictionnaires des formes fléchies.

Prenant l'exemple de l'entrée du dictionnaire électronique DELAS « chanter » qui est décrite comme un verbe qui se conjugue sur le modèle de *aider* (V3), transitif (tr) et qui appartient au vocabulaire de base du français (z1)[54].

chanter,V3+tr+z1

INTEX génère à partir de cette entrée une liste de mots fléchis contenant les formes conjuguées du verbe « chanter » en utilisant la grammaire rationnelle V3 dont nous exposons un extrait :

chante,chanter.V+z1:P1s:P3s:S1s:S3s:Y2s
chantes,chanter.V+z1:P2s:S2s
chantez,chanter.V+z1:P2p:Y2p
chantent,chanter.V+z1:P3p:S3p
chantera,chanter.V+z1:F3s
chanterai,chanter.V+z1:F1s
chanteraient,chanter.V+z1:C3p
chanterais,chanter.V+z1:C1s:C2s
chanterait,chanter.V+z1:C3s
chanteras,chanter.V+z1:F2s
chanterez,chanter.V+z1:F2p
chanteriez,chanter.V+z1:C2p
chanterions,chanter.V+z1:C1p
chanterons,chanter.V+z1:F1p
chanteront,chanter.V+z1:F3p

Nous présentons également un exemple d'une entrée du dictionnaire DELAC qui décrit une « carte de géographie » comme un nom ayant la structure « Nom de Nom » (NDN) qui est

[53] Les dictionnaires DELA sont associés à des codes morphologiques décrivant la flexion et la dérivation. Intex permet de compiler ces dictionnaires dans le but d'obtenir des dictionnaires des formes fléchies.

[54] Une liste d'opérateurs peut être prédéfinie par le système Intex et utilisée dans les dictionnaires DELA, par exemple l'opérateur *z1* pour désigner le vocabulaire de base et l'opérateur *conc* pour désigner un objet concret.

utilisée pour mettre l'entrée au pluriel. Elle désigne un objet concret (Conc) et fait partie du vocabulaire de base (z1).
carte de géographie, N+NDN+Conc+z1

Cette entrée permet la génération des listes des mots composés fléchis contenant la forme singulière féminine et la forme plurielle féminine de l'entrée « carte de géographie » :
carte de géographie,.N+NDN+Conc+z1:fs
cartes de géographie,carte de géographie.N+NDN+Conc+z1:fp

INTEX permet l'interrogation de ces dictionnaires DELAS et DELAC à travers des expressions rationnelles qui permettent par exemple d'extraire la liste des noms pluriels qui ne se terminent pas par « s », « x » et « z » ou encore la liste des mots composés commençant par la forme « carte » afin d'obtenir les types de cartes recensées dans le dictionnaire, comme dans l'extrait suivant.

carte agronomique,.N+NA+Conc:fs
carte aéronautique,.N+NA+Conc:fs
carte bancaire,.N+NA+Conc+z1:fs
carte bathymétrique,.N+NA+Conc:fs
carte bleue,.N+NA+Conc+z1:fs
carte budgétaire,.N+NA:fs
carte communale,.N+NA+Conc:fs
carte céleste,.N+NA+Conc:fs
carte d'accréditation,.N+NDN+Conc:fs
carte d'acheteur,.N+NDN+Conc:fs
carte d'admission,.N+NDN+Conc:fs
carte d'anniversaire,.N+NDN+Conc:fs
carte d'embarquement,.N+NDN+Conc:fs
carte d'entrée,.N+NDN+Conc:fs
carte d'immatriculation,.N+NDN+Conc:fs
carte d'introduction,.N+NDN+Conc:fs
carte d'ordinateur,.N+NDN+Conc:fs
carte d'émulation,.N+NDN+Conc:fs
carte de boutons,.N+NDN+Conc:fs
carte de chemin de fer,.N+NPNPN:fs
carte de condoléances,.N+NDN+Conc:fs
carte de crédit,.N+NDN+Conc:fs
carte de débarquement,.N+NDN+Conc:fs
carte de fidélité,.N+NDN+Conc:fs
carte de géographie,.N+NDN+Conc:fs
carte de journaliste,.N+NDN+Conc:fs
carte de lecteur,.N+NDN+Conc:fs

De plus, INTEX peut utiliser les dictionnaires DELAS/DELAC d'une part pour annoter des textes en faisant un étiquetage a priori hors contexte, permettant d'associer les entrées de

ces deux listes aux différentes formes de textes, et d'autre part pour le développement et l'application des grammaires locales.

Il existe des dictionnaires DELAS et DELAC pour une dizaine de langues. Pour le français un dictionnaire DELAS a été développé contenant plus de 130000 entrées et un dictionnaire DELAC d'environ 30000 entrées[55].

2.4. Les dictionnaires électroniques de NooJ

Si les formalismes DELA et WordNet permettent de répondre à des besoins applicatifs spécifiques comme l'indexation des noms simples et composés, la lemmatisation et la création des ontologies, le système des dictionnaires NooJ a une couverture applicative plus large et plus diverse puisqu'il a été conçu pour des projets de formalisation de langues qui ont comme objectif de traiter tous les niveaux d'analyse linguistique. Le dictionnaire électronique de NooJ présente en effet des solutions pour remédier aux limites de DELA.

Le formalisme des dictionnaires NooJ permet de décrire l'ensemble du vocabulaire d'une langue au sein d'un même dictionnaire. La formalisation du vocabulaire ne nécessite donc aucun effort de distinction entre les types d'entrées lexicales qu'elles soient simples ou composées. Car c'est lors de la compilation du dictionnaire que le générateur des formes fléchies traite la distinction entre les types des entrées lexicales. En effet, le compilateur interprète les opérateurs de description du vocabulaire en fonction du type d'entrée lexicale. Par exemple l'opérateur *FLX* assigne des règles de flexion à une entrée lexicale, est interprété différemment selon le type de l'entrée lexicale simple ou composée. Prenons l'exemple des deux entrées lexicales ci-dessous, l'une est simple et l'autre est composée :

pomme,N+FLX=TABLE
année lumière,N+FLX= <P>TABLE

On constate que le mode de fonctionnement du compilateur est différent pour les deux entrées. Pour la première entrée, il applique les règles de flexion à la forme simple « pomme » en ajoutant un « s » pour le pluriel « pommes ». Pour la deuxième entrée lexicale, le compilateur segmente d'abord la forme composée « année lumière » et n'applique les règles de flexions qu'à la deuxième forme « lumière » en lui ajoutant le « s » du pluriel, « année lumières ».

[55] Depuis 2001, le dictionnaire n'a plus connu de mise à jour significative.

Le compilateur intègre donc les procédures de flexion des entrées lexicales simples et composées. En outre, ce système a la particularité d'unifier l'utilisation des règles morphologiques pour tous les éléments du vocabulaire. En d'autres termes et comme notre exemple ci-dessus le montre pour deux entrées lexicales, les grammaires morphologiques qu'elles soient de flexion ou de dérivation, une fois développées, peuvent être exploitées pour décrire les entrées lexicales simples et composées. Cette méthode de compilation du dictionnaire a donc permis de simplifier et d'améliorer le processus de traitement des données assuré par NooJ. De plus, on constate une meilleure optimisation des performances et des ressources par rapport au système DELA. En effet, environ 85% des opérateurs utilisés pour décrire les entrées lexicales simples sont utilisés et appliqués de la même manière pour les entrées lexicales composés, d'où l'intérêt de traiter toutes les entrées lexicales en même temps par un seul compilateur.

Au contraire du système DELA qui contient des milliers d'entrées qui ne sont pas des éléments du vocabulaire comme *aujourd* (constituant d'*aujourd'hui*) et *parce* (constituant de *parce que*), cette conception des dictionnaires NooJ présente l'avantage de permettre le développement d'un dictionnaire où chaque entrée lexicale correspond à une ALU, et où chaque ALU est décrit par une seule entrée lexicale. En effet, cette propriété assure une meilleure analyse du texte en diminuant considérablement le nombre d'ambigüités générées par le système.

Une des limites majeures des formalismes comme DELA et WordNet est qu'ils ne prennent pas en compte un important procédé de formation des mots : la dérivation. La dérivation permet en effet de former des éléments de vocabulaire par l'adjonction d'un ou plusieurs affixes à un morphème lexical appelé base. Par exemple, l'entrée lexicale « *local* » a comme formes dérivées *« local**iser »,*** *« localis**ation »*** et *« **dé**localisation »*. Le dictionnaire NooJ traite la dérivation qui, à l'instar de la flexion, peut être formalisée via un langage rationnel. En effet, NooJ avec son formalisme permet de générer les formes dérivées à partir d'une entrée lexicale. Par conséquent, les formes dérivées ne constituent pas des entrées lexicales indépendantes dans les dictionnaires mais elles sont associées à une ALU. Voici un exemple d'une entrée lexicale qui décrit la dérivation :

voler,V+tr+FLX=Aimer+DRV=Re+DRV=Able:Artiste

Cette entrée représente l'élément du vocabulaire « *voler* » qui est un verbe (V) se conjugue avec le modèle *Aimer (+FLX=Aimer),* qui accepte une préfixation en *re (+DRV=RE),* et qui se dérive en *volable* qui se fléchit elle-même sur le modèle *Artiste.*

Cette option est d'une grande utilité pour les applications de TAL. Par exemple, une simple requête NooJ telle que <voler> retrouve toutes les occurrences de l'ALU *voler*, y compris lorsque cette unité linguistique apparait sous les formes « *revoler* », « *revolons* » ou « *volable* ».

Le dictionnaire NooJ permet également d'associer à une entrée lexicale différentes variantes orthographiques. En effet, des opérateurs ont été définis afin de faciliter la mise en place de cette liaison entre variante et ALU tels le caractère spécial (-) qui représente l'alternance entre l'agglutination (exemple audiovisuel), le trait d'union (exemple audio-visuel) et l'espace (audio visuel) ; et le caractère spécial « = » qui représente l'alternance entre le trait d'union (exemple audio-visuel) et l'espace (exemple audio visuel)[56]. De plus, dans le cas où les règles morphologiques de flexion ou de dérivation et les opérateurs spéciaux ne permettent pas de décrire l'ensemble des variantes, le dictionnaire NooJ offre la possibilité de relier explicitement les variantes à l'entrée lexicale. Nous citons l'exemple de l'association du nom « tsar » à quatre variantes « csar », « czar » et « tzar ».

tsar,N+Hum+FLX=Crayon
csar,tsar,N+Hum+FLX=Crayon
czar,tsar,N+Hum+FLX=Crayon
tzar,tsar,N+Hum+FLX=Crayon

De ce fait, nous considérons que le système des dictionnaires NooJ constitue un formalisme puissant qui permet de décrire l'ensemble des variantes flexionnelles, dérivationnelles et orthographiques. En outre, ce formalisme permet l'ajout d'autres traits et informations linguistiques tels que les traits syntaxiques et sémantiques. En effet, il est possible de définir, autant que nécessaire, des nouveaux traits linguistiques pour diverses utilisations via un fichier de paramétrage appelé fichier de définition de propriétés « *_properties.def*». Le système intègre ces différents traits dans le processus d'annotation et d'interrogation sans que ce soit au détriment des informations linguistiques de base prédéfinies par le système.

3. Processus de construction d'un dictionnaire en moyen français

Dans cette section nous présentons le processus de la construction du dictionnaire électronique en moyen français qui vise à définir les entrées lexicales et à leur associer

[56] Ces opérateurs permettent seulement de décrire les variations orthographiques. Les mots en contractions ou en agglutinations sont traités lors de l'analyse lexicale.

l'ensemble des informations linguistiques à savoir un lemme, la nature grammaticale et des informations syntactico-sémantique[57], des règles morphologiques de flexion et de dérivation et des variantes orthographiques. Au contraire des langues standardisées où il est possible à partir d'un ensemble de ressources linguistiques d'élaborer une liste des lemmes qui constituent le vocabulaire standard de la langue (Mesfar, 2008), le moyen français se caractérise par l'absence d'un dictionnaire de référence qui définirait le vocabulaire en employant une orthographe standardisée. En effet, l'instabilité et la non-unicité de l'orthographe du moyen français, qui contient des variantes tant dialectales que stylistiques et des formes qui subissent des faits d'évolution morphologiques, rendent la tâche de construction du vocabulaire standard plus complexe.

Notre démarche, organisée en trois étapes principales figurées dans la figure 50, consiste à analyser et à traiter seulement les formes attestées ayant existé dans des textes en moyen français. De ce fait, nous avons commencé par déterminer le vocabulaire le plus fréquent du MEDITEXT et cherché à le lemmatiser en utilisant des ressources linguistiques externes. Puis, nous avons procédé à une annotation semi-automatique du corpus MEDITEXT afin de constituer le vocabulaire standard. Cette annotation a permis de valider, d'enrichir et d'améliorer la couverture du vocabulaire. A cette étape, les entrées du dictionnaire de standard établies sont constituées de l'entrée lexicale associée à sa partie du discours et à une liste de variantes. Finalement, ce dictionnaire a été formalisé en dictionnaire électronique NooJ. Par conséquent, des règles morphologiques de flexion et de dérivation ont été associées aux entrées lexicales.

[57] Les informations syntactico-sémantiques sont utilisées dans notre dictionnaire pour distinguer les homonymes par le trait « *SENS* » comme pour l'exemple roi,NC+SENS=1. Elles seront aussi utilisées pour spécifier les preuves internes des entités nommées afin de distinguer les noms, prénoms, noms de lieux, etc.

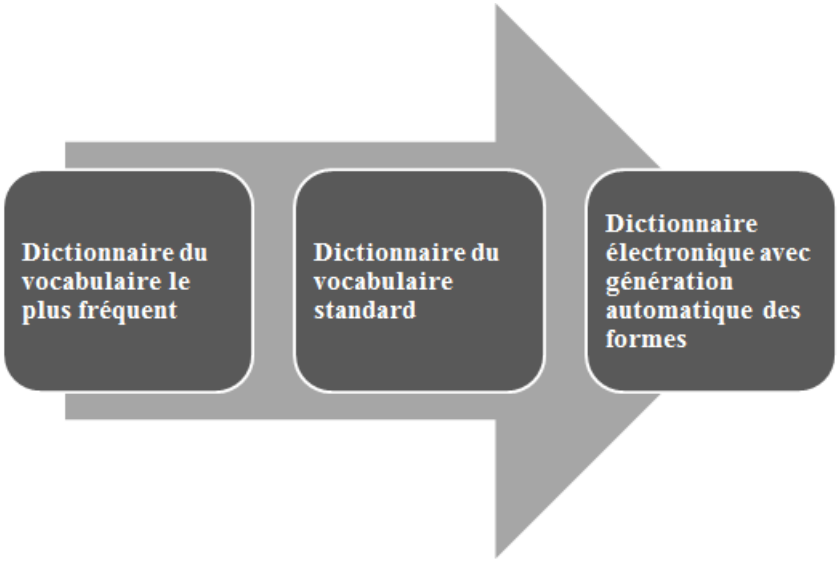


Figure 50. Processus de la construction du dictionnaire électronique en moyen français

3.1. Identification du vocabulaire le plus fréquent

A notre connaissance, les ressources linguistiques en moyen français ne sont pas nombreuses et pour la plupart elles sont hétérogènes ce qui a nécessité un travail de restitution. Dans cette section, nous présentons donc les ressources linguistiques disponibles et accessibles et la manière de les exploiter pour la constitution d'un dictionnaire du vocabulaire le plus fréquent à savoir *(i)* le Dictionnaire du moyen français (*DMF*) élaboré par Robert Martin et ces collaborateurs au laboratoire ATILF du CNRS ("Analyse et Traitement Informatique de la Langue Française"), *(ii)* l'*anglo-normand dictionary* (Bennett, 2007) réalisé par des équipes aux université Aberystwyth et l'université de Swansea dirigées par David Trotter et *(iii)* la base lexicale *Orthofonic* développée par deux chercheurs Fabrice Jejcic et Claire Fondet du LAMOP (Laboratoire des Médiévistes Occidentales de Paris) en 2011.

3.1.1 Dictionnaire du moyen français (DMF)

Le Dictionnaire du moyen français (DMF) est une application en ligne qui offre en utilisant un dictionnaire électronique du moyen français la fonction de consultation d'un dictionnaire éditorial traditionnel avec plusieurs options lexicographiques. A l'instar du Trésor de la langue française (TLF) pour le français moderne, il est considéré comme le dictionnaire de référence de la langue française aux XIVe et XVe siècle. En effet, un des principaux objectifs du DMF est « de mettre le dictionnaire de l'ancienne langue au niveau des dictionnaires modernes, tout particulièrement dans le traitement des mots sémantiquement complexes. » (Martin & Bazin-Tacchella, 2012).

Les textes du DMF couvrent la période qui va de 1330 à 1500 et ils sont composés des textes dits « intégraux » qui forment une sous-base de FRANTEXT (Bernard et al., 2002) constituée de 220 textes et des textes dits « partiels » qui contiennent 460 textes qui sont provisoirement réservés aux rédacteurs du DMF et qui ne sont pas accessibles par FRANTEXT.

Ce dictionnaire électronique a été structuré afin de permettre l'agrégation de plusieurs informations linguistiques et extralinguistiques avec la possibilité de créer des liens entre les diverses informations. Une des particularités de DMF est qu'il traite les homonymes en utilisant des règles de dégroupement qui sont fondées sur l'étymologie. Les homonymes sont donc dégroupés en entrées multiples dans le but de distinguer toutes les entrées lexicales de même forme mais d'étymologie distincte. Ainsi par exemple l'entrée lexicale *louer1* a été utilisée pour exprimer *laudare* et l'entrée lexicale *louer2* exprime *locare* (R. Martin, 2012).

En outre, le DMF est une application subdivisée en deux volets qui font la richesse du dictionnaire électronique : la consultation électronique et l'édition numérique.

- La consultation électronique consiste à donner à l'utilisateur la possibilité d'accéder au contenu du dictionnaire à travers diverses fonctionnalités de navigation à l'intérieur et en dehors de l'ouvrage. En effet, en plus d'un index alphabétique de toutes les entrées du dictionnaire et un outil permettant à l'utilisateur d'avoir un accès à des entrées bien déterminées par une opération de filtrage ou un système d'auto-complètion à partir de l'initial des entrées lexicale, un système d'interrogation appelé « LGerm »[58] a été mis en place permettant à partir d'une forme quelconque d'effectuer une lemmatisation hors-contexte afin de trouver la ou les entrées lexicales correspondantes comme le montre la figure 51 qui affiche un exemple du résultat d'interrogation du DMF par la forme « *roy* ».

[58] Le LGerm est un système qui permet, à partir d'une base de formes lemmatisées et un ensemble de règles graphémiques et morphologiques, de lemmatiser des mots en moyen français facilitant ainsi la consultation d'un dictionnaire,

Figure 51. Exemple d'interrogation du DMF par la forme « *roy* »

Une fois que les différents outils d'interrogation du dictionnaire ont donné accès au contenu de l'entrée recherchée, comme l'illustre la figure 52, plusieurs possibilités de navigation sont offertes permettant de consulter diverses informations qui sont reliées entre elles telles les différentes définitions accordées, les formes attestées liées à l'entrée lexicale, des lexiques auxquels l'entrée lexicale appartient, des exemples de contextes d'apparition ainsi que la base textuelle des sources à partir desquelles l'entrée lexicale a été construite.

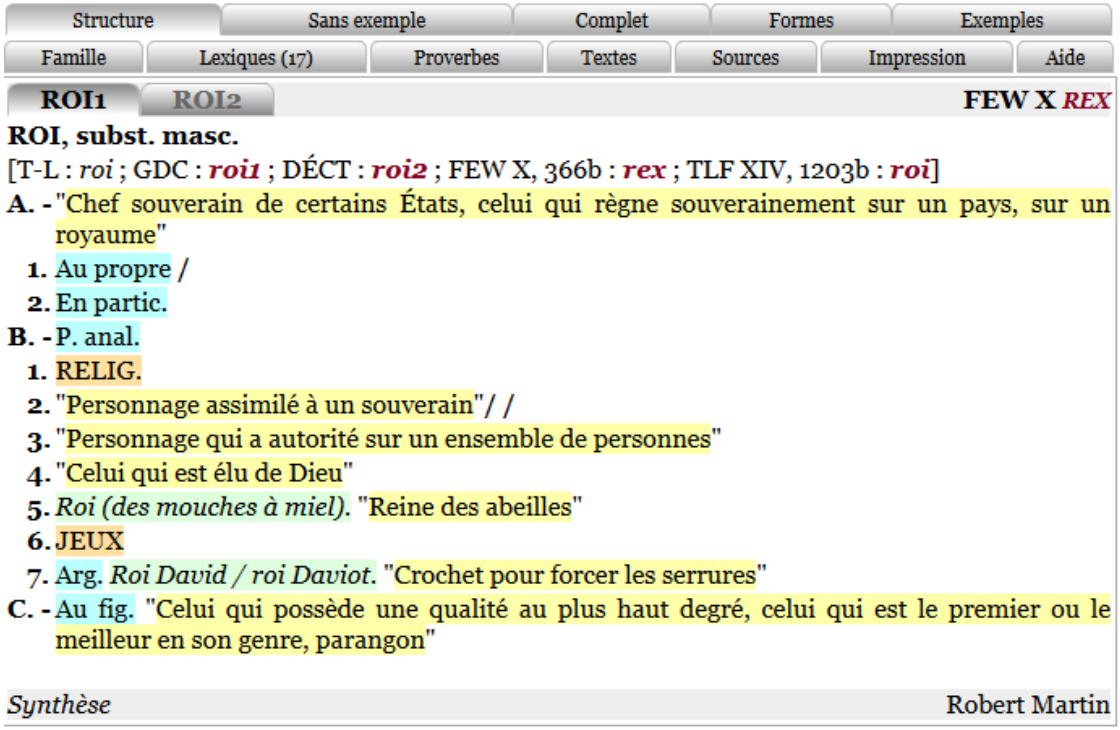


Figure 52. Consultation de l'entrée « roi1 » de DMF

- L'édition numérique est la mise en place de la méthodologie employée pour l'élaboration du DMF en utilisant le développement des outils informatiques. En effet, en plus des interfaces de saisie qui facilitent aux rédacteurs la gestion du dictionnaire électronique par ajout, modification et suppression des entrées lexicales, le système contient des outils d'aide à la rédaction tels un « correcteur lexicographique » et un « masque de saisie ». Ce dernier est une sorte d'éditeur en ligne qui permet, au fil de la rédaction, « d'indiquer le type d'information qu'il convient de fournir et d'afficher au fur et à mesure des balises à remplir ou les choix à faire parmi les balises possibles dans certains contextes » (R. Martin, 2012). Quant à l'utilisation du correcteur lexicographique, il joue le rôle de compilateur qui permet la signalisation des erreurs existantes dans les fichiers de saisie. Ces erreurs sont de quatre types : des erreurs de balisage, des erreurs matérielles, des erreurs de référence et des erreurs de renvoi.

3.1.2. *Anglo-Normand dictionary*

A l'instar du DMF, l'*anglo-normand dictionary* est une application en ligne qui permet l'exploration, la consultation et la gestion du dictionnaire électronique. La particularité du dictionnaire de l'anglo-normand est qu'il vise à fournir les outils de base pour rendre possible la compréhension du langage dans lequel les textes anglo-normands ont été écrits. En effet, il a été constitué à partir d'une base textuelle qui documente une période essentielle, complexe et longue de l'histoire britannique et qui couvre plusieurs secteurs et domaines de la vie au Moyen Âge à savoir littéraire, scientifique, médical, administratif, historique, religieux, national, international et local. Ce dictionnaire est donc considéré comme un outil indispensable pour utiliser et explorer ces textes afin de comprendre le fonctionnement de la société qui les a produits.

L'*anglo-normand dictionary* a été élaboré à partir d'une base textuelle numérisée et éditée comportant environ 76 textes ouverts au public. Il couvre la période allant de la deuxième moitié du XIIIème siècle à la fin du XVème siècle. Il est doté d'une documentation complète consultable en ligne qui décrit toutes les interfaces et leurs fonctionnalités. En effet, l'interface de consultation offre deux façons d'accéder aux entrées :

- Par un index de toutes les entrées lexicales classées par ordre alphabétique. Comme le montre en effet la figure 53, l'application offre une interface de navigation en affichant plusieurs possibilités d'accès aux entrées lexicales. En plus d'un accès direct à toutes les entrées du dictionnaire avec une facilité de navigation au travers des

onglets et des interfaces, l'index a été divisé en segments permettant ainsi d'avoir un accès direct au segment contenant l'entrée lexicale recherchée par l'utilisateur.

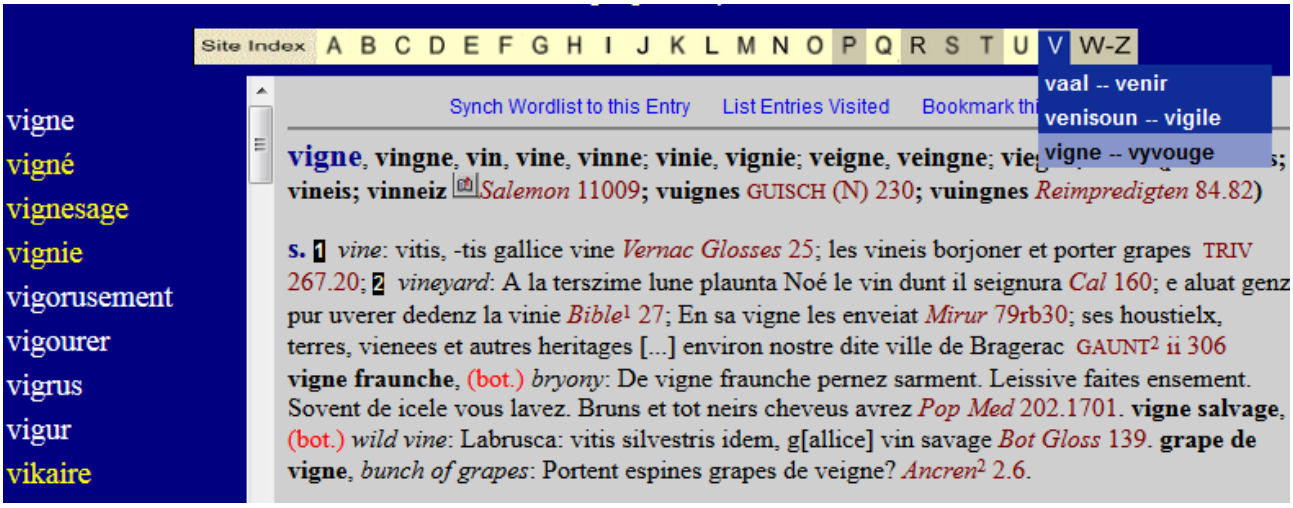


Figure 53. Accès à l'entrée lexicale « vigne » en parcourant l'index

- Par un formulaire de recherche du dictionnaire qui permet, à partir, d'une forme saisie par l'utilisateur d'interroger les variantes ainsi que les entrées lexicales auxquelles elles sont associées. Comme l'illustre la figure ci-dessous. Si une ou plusieurs entrées lexicales ou variantes correspondent à la requête lancée, le système offre toutes les entrées lexicales possibles sans utilisation des calculs qui permettent de filtrer ou de trier les résultats selon leurs pertinences. Un des avantages de ce système est qu'il offre la possibilité d'interroger via un concordancier la base textuelle du l'*anglo-normand dictionary*. Par conséquent, le système fournit les concordances affichant les différents contextes d'apparitions de la forme recherchée dans les textes de la base. Enfin, si l'utilisateur a une idée du sens général de la forme qu'il étudie, un outil « *Search Translations* » est mis en place afin de retrouver le synonyme ou l'équivalent en anglais de la forme étudiée.

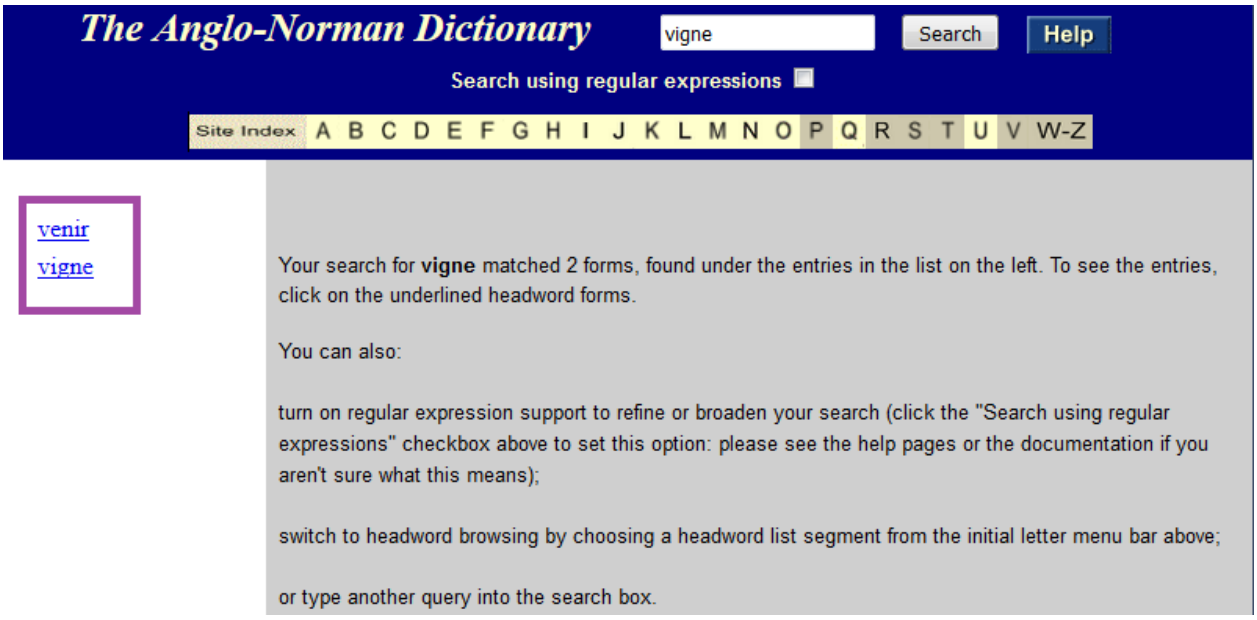


Figure 54. Accès à la forme lexicale « vigne » en utilisant le système d'interrogation

3.1.3. La base de données lexicales *OrthoFonic*

OrthoFonic est un projet de didacticiel pour l'apprentissage de l'orthographe française porté par (Jejcic et Fondet, 2011) qui ambitionnent par ce projet de renouveler l'enseignement du français courant en partant de formes vedettes et de formes fléchies classées selon un critère de fréquence. En effet, leur principal objectif est de mettre au point un système didacticiel d'apprentissage de l'écrit et de l'oral du français d'usage courant. Ce système se concrétise par une application en ligne qui exploite et gère un dictionnaire électronique du « français courant ». Faute de financement, il a fallu environ une trentaine d'années pour constituer ce dictionnaire et aucun développement applicatif n'a pu être effectué visant à la consultation et à la gestion dudit dictionnaire.

En fait, l'élaboration d'un tel dictionnaire du langage courant utile pour l'enseignement et pour les applications de TAL n'est pas une idée récente. Elle part du constat qu'il y a « une forme de "noyau" de la langue qu'il convient de privilégier dans l'enseignement ». Plus précisément, elle consiste à déterminer, en suivant une méthode scientifique de calcul statistique, le vocabulaire le plus utilisé de la langue française. Une des tentatives les plus intéressantes qui cherche à déterminer le vocabulaire courant du français date du début des années quatre-vingt et elle a donné lieu à un dictionnaire électronique appelé par Listes orthographiques de base (LOB) publiée en 1984 (*LOB*, 1984: 17). Bien que les résultats obtenus soient intéressants, le projet n'a pas pu être mené à terme pour analyser massivement et explorer un corpus volumineux qui couvre divers aspects chronologiques et géographiques

de la langue française à cause des difficultés techniques liées aux moyens matériels et logiciels disponibles dans les années 1980.

Grâce aux progrès technologiques réalisés ces trente dernières années, le projet a été relancé en 2006 en exploitant diverses ressources linguistiques. Ainsi la base de données lexicale d'*Orthofonic* est le résultat d'exploration de plusieurs bases de données lexicales construites à partir d'enquêtes ou de corpus informatisés. En plus des *Listes orthographiques de base* (*LOB*) qui constituent le noyau d'*Orthofonic,* les entrées de quatre autres bases lexicales ont été ajoutées : L'élaboration du français élémentaire[59] (Gougenheim et al. 1956, 1971), Le vocabulaire orthographique de base (Ters et al., 1968), L'échelle Dubois-Buyse[60] (Ters et al, 1969) et, enfin, le Dictionnaire des fréquences du Trésor de la langue française[61].

Une étude synthétique de ces différentes bases a été effectuée permettant d'en extraire un vocabulaire performant de haute fréquence d'usage. Elle a fait apparaître « une grande concentration du vocabulaire d'usage très courant : elles ont en commun la quasi-totalité des 1280 premiers mots du Trésor de la langue française dont la fréquence, rapportée au nombre d'occurrences, est supérieure à 5000 ». La base lexicale d'*Orthofonic* dans sa version finale contient deux fois plus de données et trois fois plus de champs lexicaux par rapport à la première version, avec la particularité de contenir les formes fléchies les plus fréquentes : la base compte 5500 entrées lexicales et 8500 formes.

Jejcic & Fondet (2011) affichent une évaluation et des statistiques suivantes sur l'efficacité de la base après avoir effectué des études de fréquence du vocabulaire :

- les 10 premiers mots couvrent près de 35 % de la fréquence d'usage;
- les 100 premiers mots couvrent près de 68 % de la fréquence d'usage;
- les 1000 premiers mots couvrent près de 89 % de la fréquence d'usage.

Dans sa configuration actuelle, la base lexicale *Orthofonic* comprend environ 8500 enregistrements et 16 champs par enregistrement. Ces champs couvrent plusieurs types d'annotation tels que des annotations orthographiques, des annotations phoniques et des

[59] Le projet *l'élaboration du français élémentaire,* pour lequel un « centre d'étude du Français élémentaire » a été créé dans les années 50, a été conduit sous la responsabilité scientifique de George Gougenheim. Il a comme but d'élaborer le vocabulaire de base le plus fréquemment utilisé dans le but de diffuser le français comme langage de communication international dans des circonstances historiques qualifiés de complexes (Coste, 2006).

[60] *L'échelle Dubois-Buyse* a été établie vers 1940 puis réactualisée par (Ters et al,1988). C'est une liste qui regroupe 3787 mots d'usage courant, vocables répartis sur 43 échelons, supposés connus de tout adulte francophone.

[61] *Dictionnaire des fréquences du Trésor de la langue françai*se est une liste contenant les formes les plus fréquentes du TLF.

annotations grammaticales. Comme le montre les trois tableaux ci-dessous, les annotations grammaticales constituent un jeu d'étiquettes contenant les parties du discours, le genre, le nombre, et indiquent les formes vedettes.

Code	Valeur	CJ	conjonction	NU	numéral
AJ	adjectif	IN	interjection	PN	pronom
AR	article	LO	locution	PS	préposition
AV	adverbe	NO	nom	VE	verbe

Tableau 7. Les étiquettes des « parties du discours »

Code	Valeur	Remarque:
F	féminin	l'absence des codes de notation du genre, M ou F, indique qu'il s'agit d'une forme épicène
M	masculin	

Tableau 8. Les étiquettes représentant le « genre »

Code	Valeur	Remarque:
S	singulier	l'absence des codes de notation du nombre, S ou P, indique qu'il s'agit d'une forme invariable
P	pluriel	

Tableau 9. Les étiquettes représentant la forme dite « vedette »

Pour conclure, l'utilisation de la base lexicale *Orthofonic* n'est pas limitée au français moderne. Elle nous a en effet été d'une grande utilité pour recenser et construire le vocabulaire standard de base du moyen français.

3.1.4. Constitution du dictionnaire du vocabulaire le plus fréquent

Comme toute langue non-standardisée, il n'existe pas un dictionnaire unique contenant des entrées lexicales auxquelles on peut associer toutes les formes attestées ayant existées dans des textes en moyen français. De ce fait, notre démarche consiste à exploiter les entrées lexicales définies par le peu de ressources linguistiques développées pour le moyen français à savoir le DMF et l'*Anglo Normand Dictionary*. Comme l'illustre la figure 55, l'élaboration d'une liste du vocabulaire le plus fréquemment utilisé en moyen français consiste à rassembler l'ensemble des formes de la base de données lexicales *Orthofonic* qui figurent dans MEDITEXT et l'ensemble du vocabulaire fréquents du MEDITEXT.

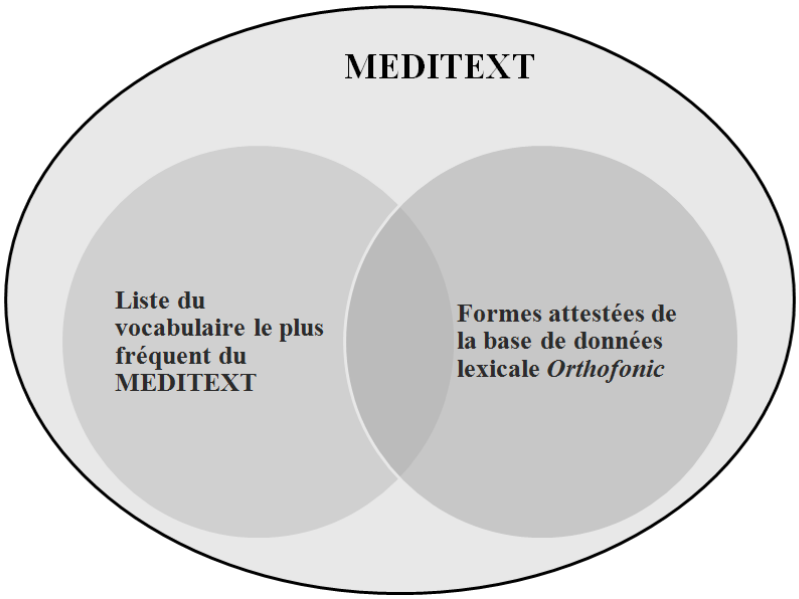


Figure 55. Constitution du vocabulaire le plus fréquent du moyen français

En effet, les formes de la base des données lexicales *Orthofonic* sont utiles pour l'analyse des formes conservées par le français moderne et des formes qui figurent dans des textes édités dont la graphie est modernisée. A partir de cette base, qui comptait au départ environ 8500 formes, nous avons automatiquement identifié une liste d'environ 5000 formes qui figurent dans MEDITEXT. Comme l'illustre la carte de section de la forme « livre » de la figure ci-dessous, cette liste des formes attestées a été vérifiée grâce à des méthodes d'analyse textométrique.

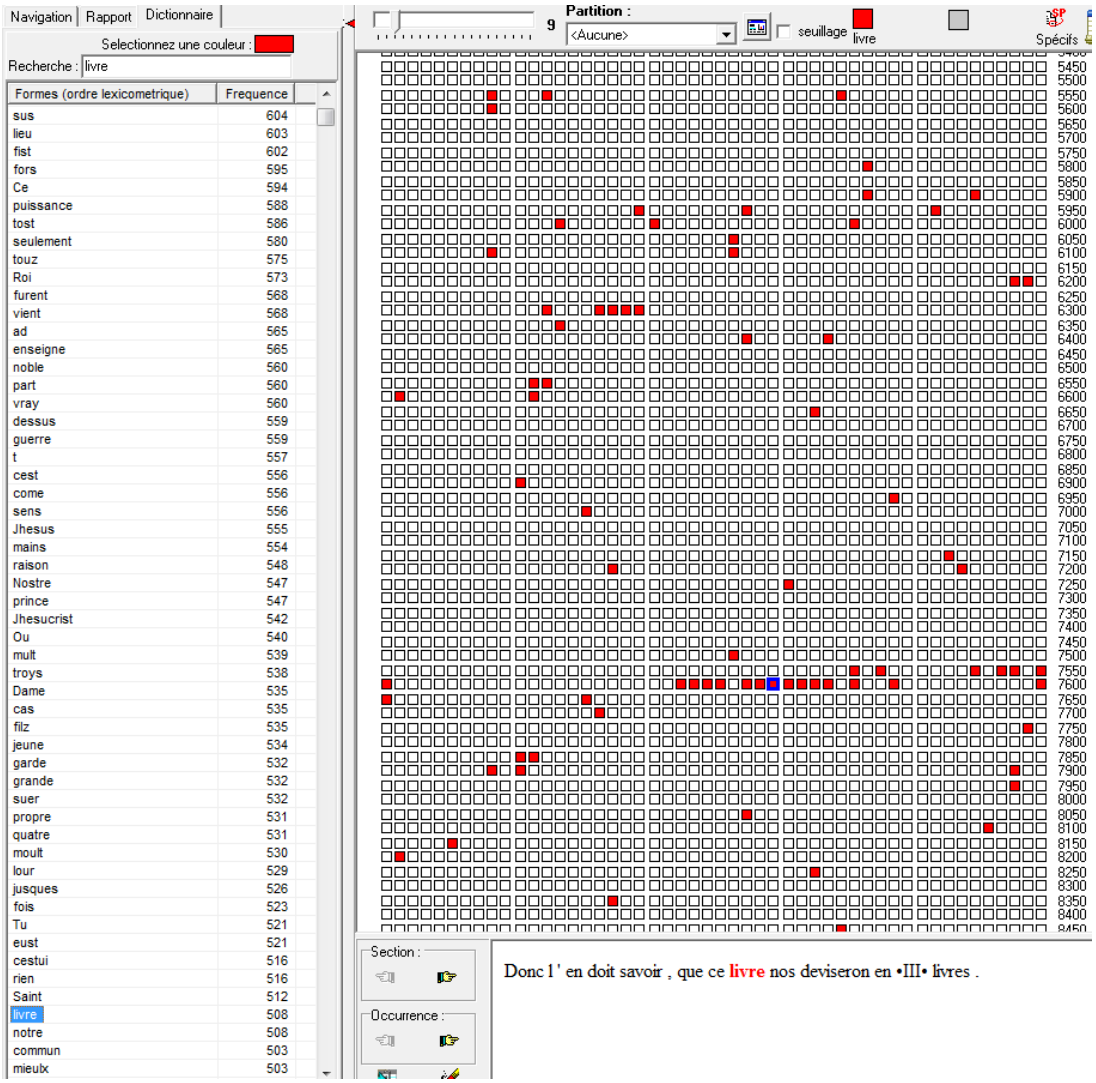


Figure 56. Carte de section de la forme « *livre* »

Cette liste initiale de 5000 formes a été enrichie et complétée par les 300 formes les plus fréquentes du MEDITEXT qui ne figurent pas dans *Orthofonic*. La liste finale contenant 5300 formes couvre environ 55,4% de notre corpus de test. Afin de construire un dictionnaire électronique du vocabulaire le plus fréquent, nous avons procédé empiriquement en interrogeant le DMF et l'*Anglo Normand Dictionary* pour associer à chacune des formes un lemme et une partie du discours.

Formes (ordre lexicometrique)	Frequence
et	67608
de	50263
en	30556
a	29342
la	28319
que	26433
le	20129
les	19946
est	17102
qui	16490
ne	16106
par	15263
il	14845
Et	11403
l	11165
ou	11130
se	9757
pour	9556
des	8685
ce	8549
du	7873
bien	7356
comme	6176
Dieu	6104
son	6055
si	5783
plus	5755
vous	5501
sont	5435
je	5307
n	5041
d	4942
sa	4854
pas	4794
dit	4790
qu	4520
grant	4414
au	4410
tout	4283
tu	4236
nous	4091

Figure 57. Extrait de la liste du vocabulaire le plus fréquent du corpus

L'interrogation automatique du DMF par les formes de notre liste consiste à extraire les annotations (partie du discours et lemme) correspondantes. Elle permet également d'enrichir la liste du vocabulaire le plus fréquent par des nouvelles variantes orthographiques. En effet, l'interrogation du DMF peut être effectué à partir d'une forme afin d'obtenir l'entrée lexicale correspondante ou à partir d'une entrée lexicale pour extraire les différentes variantes orthographiques attestées, comme le montre la figure 58.

Recherche d'une entrée

Formes de l'entrée RANÇON

Voir l'article RANÇON

16 formes différentes pour 38 occurrences

	occ.	BGV
raencon	2	
raençon	7	
raënçon	1	
raençons	4	
rainçon	1	
raïnçon	3	
rainson	1	
ramsons	1	
ranceon	1	
rancon	1	
rançon	5	
ranczon	1	
reançon	1	
renchon	1	
rençon	5	
renson	3	

Figure 58. Tableau des formes attestées du *DMF*

Comme l'illustre la figure ci-dessous, pour chaque forme de la liste, l'interrogation[62] du LGerm offre l'entrée lexicale du dictionnaire à laquelle un ensemble d'informations est associé comme le lemme et la partie du discours. Une fois les lemmes récupérés, notre système interroge les formes attestées du DMF afin de recenser les différentes formes associées à chaque entrée lexicale.

[62] Le LGerm offre à partir d'une forme le lemme associé dans le DMF. Or dans le cas où la forme n'est pas attestée, il applique un ensemble de règles morphologiques pour proposer des entrées lexicales potentielles. Nous avons fait le choix de n'utiliser que les formes attestées et nous avons interrogé l'*Anglo-Normand Dictionary* dans le cas où la forme n'apparaissait pas dans le DMF.

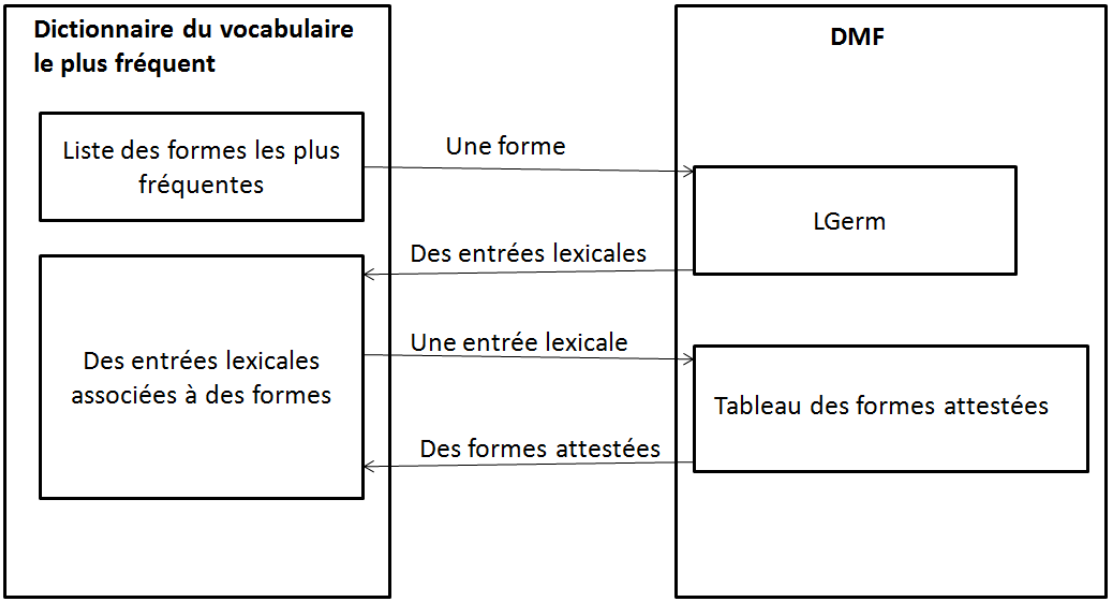


Figure 59. Interrogation automatique du DMF

Pour les formes qui n'existent pas dans le *DMF*, nous avons eu recours à une interrogation automatique de l'*Anglo-Normand Dictionary*. Comme l'illustre la figure 60, il est possible à partir d'une forme d'obtenir l'unité lexicale correspondante ainsi que les variantes orthographiques associées qui sont attestées dans des textes Anglo-Normands. En effet, ce dictionnaire répertorie des variantes orthographiques insulaires importantes pour l'analyse des textes de provenance anglaise. En conclusion, une première liste a été établie contenant les entrées lexicales les plus fréquentes associées aux différentes formes attestées existantes dans les textes du MEDITEXT, DMF et l'*Anglo-Normand Dictionary*.

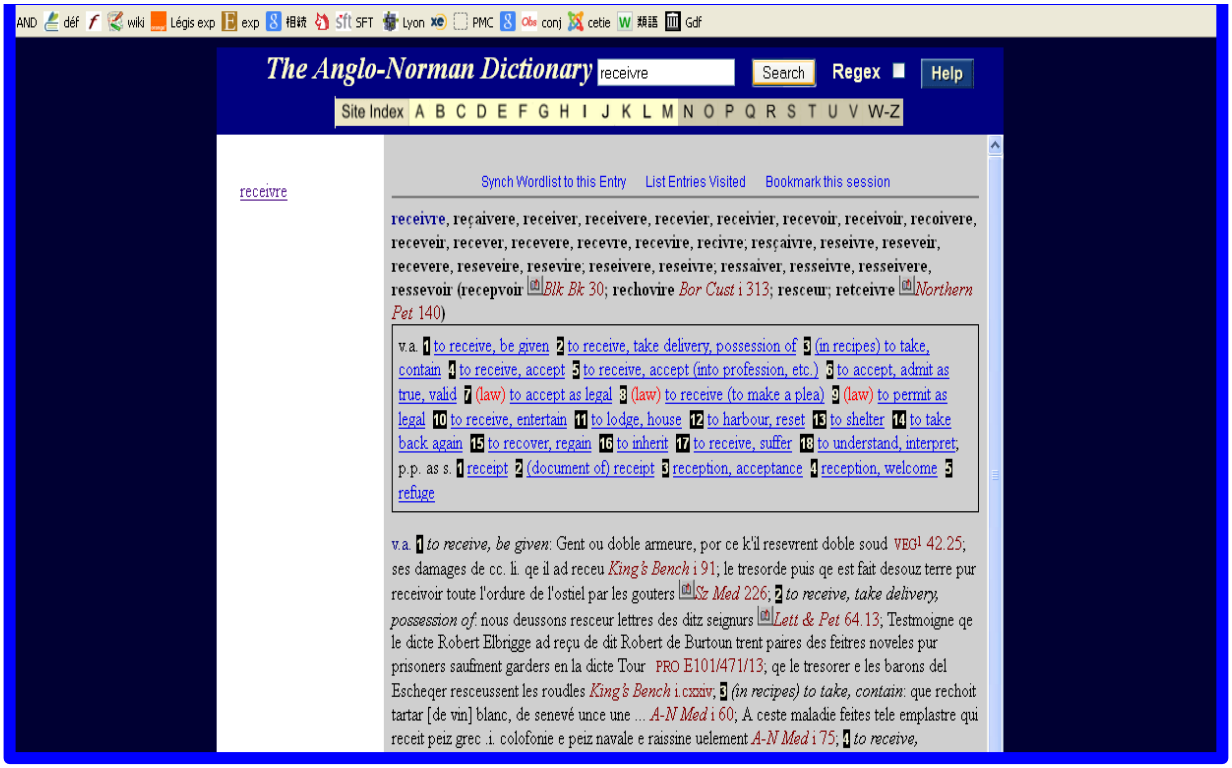


Figure 60. Interrogation manuelle de l'*Anglo-normand dictionary*

Enfin, une uniformisation des lemmes et des parties du discours s'avére nécessaire pour la construction d'un dictionnaire. En effet, les ressources linguistiques utilisées n'ont pas le même jeu d'étiquettes et le choix des lemmes de l'*Anglo-Normand Dictionary* devrait être effectué en conformité avec le système du DMF. Nous avons donc procédé à une validation manuelle du lexique constitué et nous avons converti semi-automatiquement les descriptions morphosyntaxiques du DMF et de l'*Anglo-normand dictionary* en un jeu d'étiquettes unique, présenté dans le tableau ci-dessous[63].

Abréviation	**Description**
ADV	Adverbe
A	Adjectif
Npropre	nom propre
Ncommun	nom commun

[63] Ce jeu d'étiquettes a été définit dans le cadre du projet « Signs and States »

PREP	préposition
PRO	Pronom
V	Verbe
CONJC	conjonction coordination
CONJS	conjonction de subordination
DET	déterminant
INTJ	interjection
NUMord	nombre ordinal
NUMcard	nombre cardinal

Tableau 10. Description du jeu d'étiquettes défini pour annoter « MEDITEXT »

3.2. Élaboration du dictionnaire du vocabulaire standard

La construction d'un dictionnaire de la langue consiste à décrire le vocabulaire standard de la langue et ne pas se contenter du vocabulaire le plus fréquent. De ce fait, à partir des glossaires, des index d'éditions et des informations encyclopédiques[64], nous avons recensé trois lexiques de l'époque médiévale contenant le vocabulaire juridique et institutionnel, le vocabulaire de la guerre et le vocabulaire religieux. Notre dictionnaire jusqu'alors ne comptait pas un nombre considérable de verbes. Nous avons donc eu recours au Dictionnaire électronique des formes fléchies (DELAF) du français moderne et au manuel d'André LANLY en 1977 « la Morphologie historique des Verbes français »[65], qui répertorient les verbes réguliers et irréguliers les plus usités en ancien français afin d'établir une liste contenant des verbes en français moderne et des verbes en ancien français. Au final une liste qui rassemble les lexiques thématiques, la liste des verbes du français moderne et la liste des verbes d'ancien français a été élaborée. Notre système d'interrogation automatique de la

[64] Il s'agit des sites sur la toile qui proposent des lexiques thématiques médiévaux. De plus, nous avons fait appel à Wikipedia pour vérifier et enrichir nos lexiques.

[65] Dans son livre « *la Morphologie historique des Verbes français* », Andry Lanly remontent aux origines des verbes irréguliers d'origines latin ou latinisés en expliquant leurs variations et en donnant les formes des conjugaisons régulières.

figure 59 a permis de vérifier l'existence de chacune des formes de cette liste et d'extraire les annotations associées. Comme pour le dictionnaire du vocabulaire le plus fréquent nous avons procédé, à ce stade, à une validation manuelle et à une étape d'unification des jeux d'étiquettes en utilisant le jeu d'étiquettes du tableau 10. Enfin, cette liste constituée contenait environ 6200 entrées lexicales dont 45% sont des verbes.

Concernant les noms propres, ces derniers ne font pas l'objet d'un recensement du DMF. De ce fait, un système[66] a été mis en place qui permet, à partir du glossaire ou d'index des éditions critiques, de recenser les noms propres et leurs variantes qui figurent dans MEDITEXT. Cette liste des noms propres attestés a été enrichie par un ensemble de noms propres recensés manuellement à partir des informations encyclopédiques classées selon des thématiques comme : les noms de ville, les noms de région, les noms de diocèse, les noms des institutions religieuses et les noms de dynastie comme les noms de roi, les noms de reine, les noms d'empereur romain et les noms mythologiques et bibliques. Ceci a permis l'obtention d'une liste d'environ 2600 noms propres. Finalement, le dictionnaire du vocabulaire le plus fréquent s'est ajouté à l'ensemble des entrées lexicales que nous avons pu extraire à partir des ressources linguistiques externes à MEDITEXT dans le but de constituer un dictionnaire du vocabulaire standard du moyen français. Ce dernier contenait au total environ 15.000 entrées lexicales qui couvrent environ 76.5% des formes de notre corpus.

Dans le but d'enrichir ce dictionnaire avec des formes issues du MEDITEXT, nous avons profité de la disponibilité des textes annotés semi-automatiquement en utilisant PALM dans le cadre du projet « *Signs and States* ». En effet, comme l'illustre la figure 61, à partir d'une interface adaptée que nous avons mise en place, une annotation semi-automatique[67] a été effectuée pour quelques textes. Linguistiquement et dans la mesure du possible, les annotations utilisées se composent des lemmes qui correspondent à ceux du DMF. Étant donné qu'il existe des textes d'origine anglaise, plusieurs lemmes n'ayant pas d'équivalent dans le DMF ont par contre été trouvés dans l*'Anglo-Normand Dictionary*. Dans les rares cas où il n'existe pas d'équivalents des formes ni dans l'*Anglo-Normand Dictionary* ni dans le DMF, sur la base des formes attestées, une des formes est définie comme lemme.

[66] Ce système permet d'extraire les noms propres à partir des index de quelques éditions critiques et de vérifier pour chaque nom propre s'il appartient à la liste des formes du MEDITEXT.

[67] Nous avons annoté automatiquement le corpus en appliquant un dictionnaire et nous avons développé l'interface d'annotation. Ensuite, un travail d'annotation semi-automatique a été assuré par Naomi Kanoaka linguiste experte en moyen français.

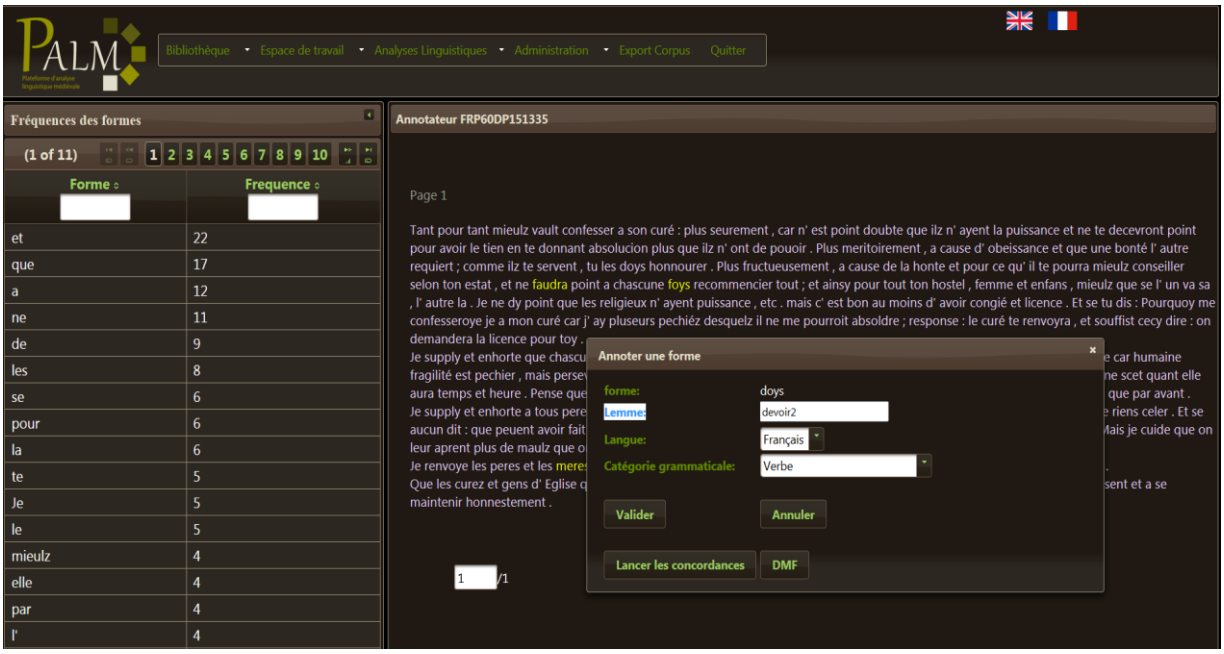


Figure 61. Interface PALM pour l'annotation semi-automatique des formes

Afin d'exploiter un ensemble des textes annotés manuellement couvrant environ 100.000 formes, un processus cyclique composé de trois étapes a été mis en place comme montre la figure 62 dans le but d'enrichir le dictionnaire du vocabulaire standard d'une manière itérative et incrémentale.

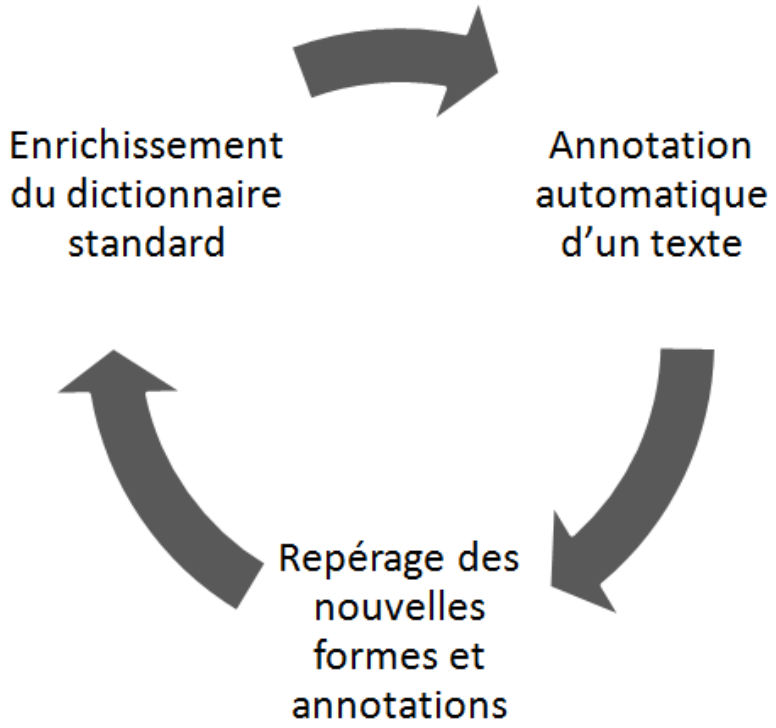


Figure 62. Cycle de l'enrichissement du dictionnaire en annotant « MEDITEXT »

1. L'annotation automatique : Il s'agit simplement d'appliquer notre dictionnaire électronique du vocabulaire standard en utilisant NooJ sur l'un des vingt textes annotés semi-automatiquement.
2. Repérage des nouvelles formes et annotations : En effectuant une simple comparaison entre les annotations générées automatiquement et les annotations effectuées semi-automatiquement, nous avons pu extraire les formes et les annotations capables d'enrichir notre dictionnaire.

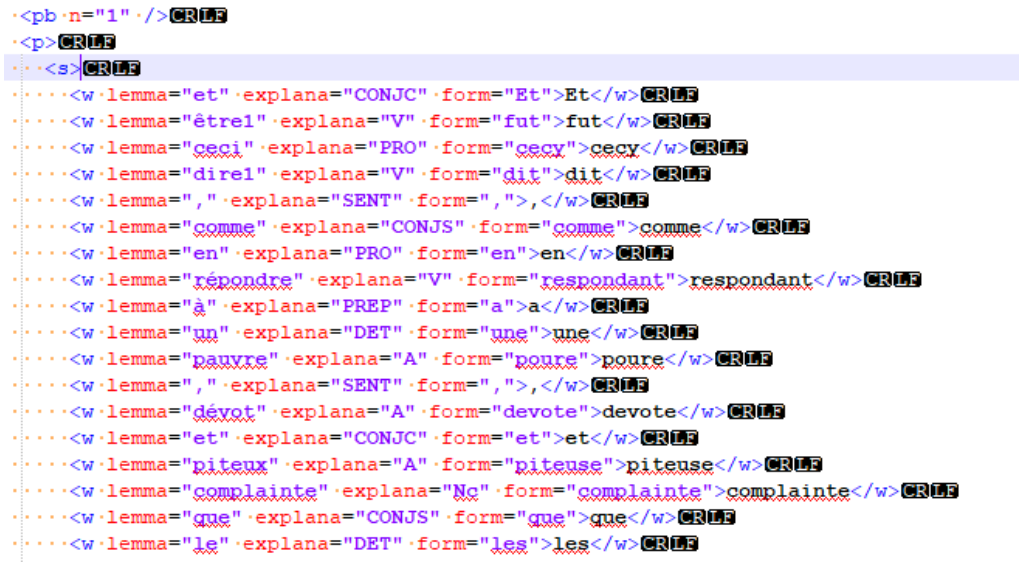


Figure 63. Structuration des annotations avec PALM

3. Enrichissement du dictionnaire standard : un algorithme a été développé pour l'alimentation automatique du dictionnaire standard du moyen français par des nouvelles formes et annotations qui ont été repérées à l'étape précédente.

Après avoir lancé ce processus sur les vingt fichiers annotés semi-automatiquement, notre dictionnaire standard du moyen français contenait environ 26.000 entrées lexicales qui couvrent 133.227 formes attestées dont 9.5% sont des noms propres et 41.3% sont des formes fléchies des verbes.

3.3. Dictionnaire du moyen français et génération automatique des formes

A cette étape, nous avons procédé à une description du système flexionnel et dérivationnel du moyen français dans le but de générer automatiquement, à partir de l'entrée lexicale, des formes non-attestées qui peuvent potentiellement exister. De ce fait, un système de classification a été mis en place pour attribuer des règles morphologiques à un ensemble d'entrées lexicales d'une même partie du discours. Ensuite, nous avons modélisé et regroupé

en machine à états finis les règles de flexions via des graphes NooJ. Finalement, nous avons associé les modèles morphologiques construits aux entrées lexicales correspondantes afin de compiler le dictionnaire et produire des formes fléchies.

3.3.1. Système de classification des entrées lexicales selon les paradigmes

Les entrées lexicales de notre dictionnaire standard se décomposent en trois informations essentielles à savoir le lemme, la partie du discours et l'ensemble des variantes associées. Voici un exemple de l'entrée lexicale « *roi* » de notre dictionnaire standard :

roi,NC+SENS=1
rai,roi,NC+SENS=1
re,roi,NC+SENS=1
ree,roi,NC+SENS=1
rei,roi,NC+SENS=1
reis,roi,NC+SENS=1
rey,roi,NC+SENS=1
roe,roi,NC+SENS=1
roi,roi,NC+SENS=1
roie,roi,NC+SENS=1
rois,roi,NC+SENS=1
roix,roi,NC+SENS=1
roiz,roi,NC+SENS=1
roy,roi,NC+SENS=1
roys,roi,NC+SENS=1
royz,roi,NC+SENS=1

L'analyse diachronique des formes grammaticales comme les pronoms et les articles, montre une réduction nette des formes irrégulières et étymologiques qui motive notre démarche. En effet, l'analyse des changements morphologiques des formes attestées de notre dictionnaire permis de valider des mécanismes de formation des variantes à partir d'une entrée lexicale. Suite à une classification automatique des entrées lexicales, selon la partie du discours et le mode de combinaison des morphèmes, les paradigmes morphologiques, qui peuvent être généralisés à des entrées lexicales d'une même classe, ont été identifiés.

Concrètement, en se basant sur de nombreuses études de morphologie du moyen français[68] et sur l'exploration de nos données qui sont notre dictionnaire standard et le corpus « MEDITEXT », nous avons modélisé de nombreux paradigmes flexionnels et dérivationnels dans un langage rationnel. Comme l'illustre la figure 64, ces paradigmes ont été utilisés pour

[68] Nous nous sommes appuyés sur plusieurs travaux comme (Marchello-Nizia, 2005), (Fouché, 1967), (Gossen C, 1968), (Lanly, 1977) et (Brunot, 1913).

un système de classification qui détermine des classes des entrées lexicales qui sont de même partie du discours et qui partagent en commun des paradigmes flexionnels. Notre système a permis non seulement une analyse confirmatoire pour déterminer la pertinence de ces paradigmes flexionnels mais aussi un regroupement des paradigmes qui s'appliquent ensemble à une même classe d'entrées lexicales.

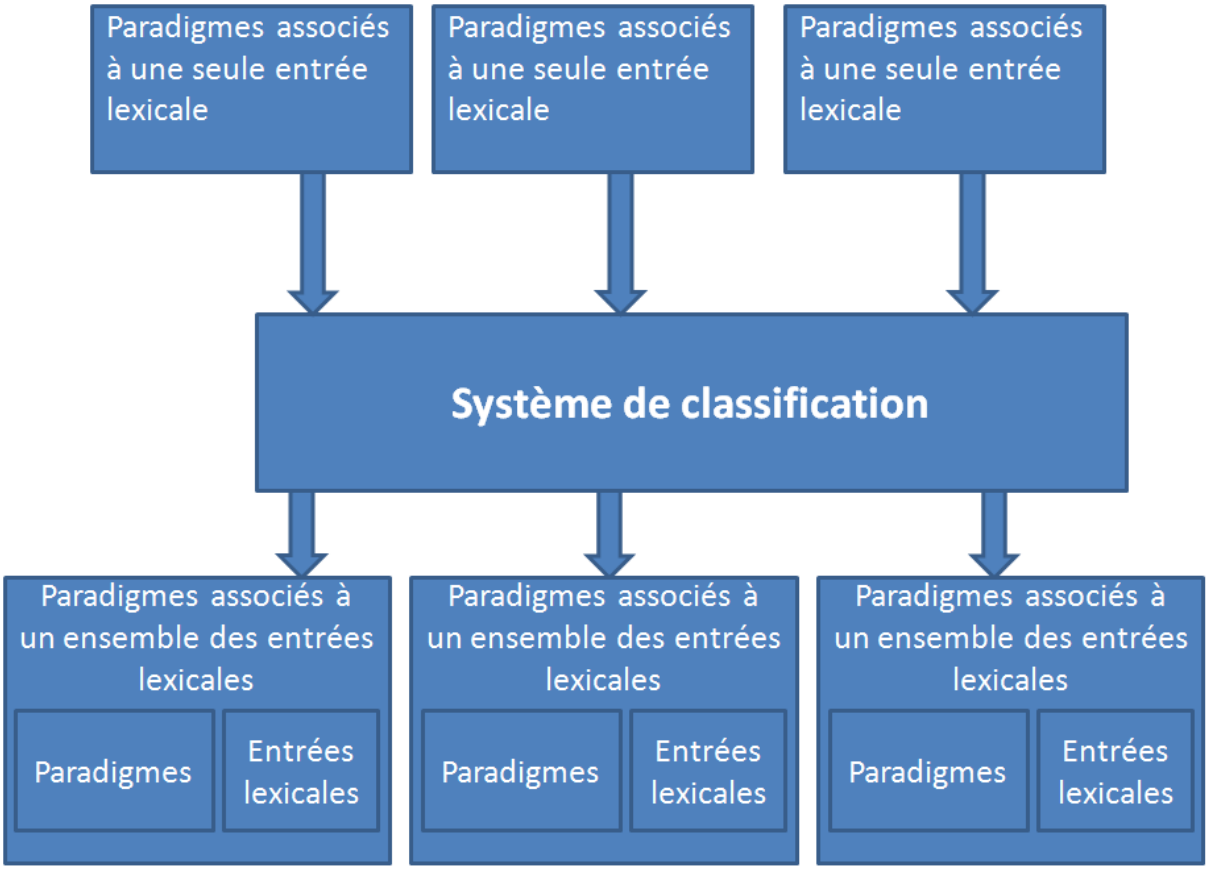


Figure 64. Schémas des entrées/sorties de notre système de classification

En suivant le formalisme des dictionnaires NooJ, le système flexionnel et dérivationnel a été décrit à l'aide des opérateurs prédéfinis[69]. Enfin, nous exposons cette étape de la mise en place de l'analyse morphologique en prenant comme exemple les règles de flexion des noms. La même démarche a été suivie pour décrire la morphologie des différentes parties du discours que sont les verbes, les adjectifs, les pronoms et les déterminants.

3.3.2 Formalisation des règles morphologique : Flexions des noms

En moyen français, les noms se particularisent par la disparition de l'une des caractéristiques essentielles de l'ancien français : la déclinaison. En effet, bien que notre

[69] Ces opérateurs sont décrits dans l'annexe B.

corpus présente encore des traces de déclinaison jusqu'au début du XVIème siècle, ces déclinaisons ne sont pas utilisées d'une façon régulière et s'estompent lorsqu'on progresse d'une période à une autre surtout pour les textes issus du nord de la France. De plus, nous constatons que le cas régime, qui est fréquemment utilisé car il assume de nombreux rôles syntaxiques, se généralise au détriment du cas sujet. Mais, il existe des formes du cas sujet qui ont pu persister et qui ont même réussi à s'imposer comme la forme standard. Les textes montrent aussi qu'il y a des formes des deux cas qui ont été gardées mais le plus souvent avec des sens ou fonctions différentes.

Cette régularité de la formation des noms communs apparue au XIVème siècle et qui a continué son évolution au XVème siècle, nous a donc permis de regrouper, grâce à notre système de classification, les règles morphologiques qui sont applicables à un ensemble d'entrées lexicales qui constituent une classe.

Ces règles morphologiques ont été donc modélisées en machines à états finis en utilisant les graphes NooJ. Pour les noms communs, nous avons déterminé 19 modèles morphologiques de flexion dont nous présentons quelques-uns.

Le premier graphe de la figure 65 présente les marques du pluriel « standard » fréquemment utilisées dans notre corpus pour fléchir les noms communs. Il permet de générer à partir d'un nom commun masculin singulier, la forme masculine singulier en ajoutant l'élément neutre ou la chaine vide <E> et les formes plurielles en ajoutant un des suffixes suivant « *s* », « *z* », « *t* », « *zt* » et « *x* ».

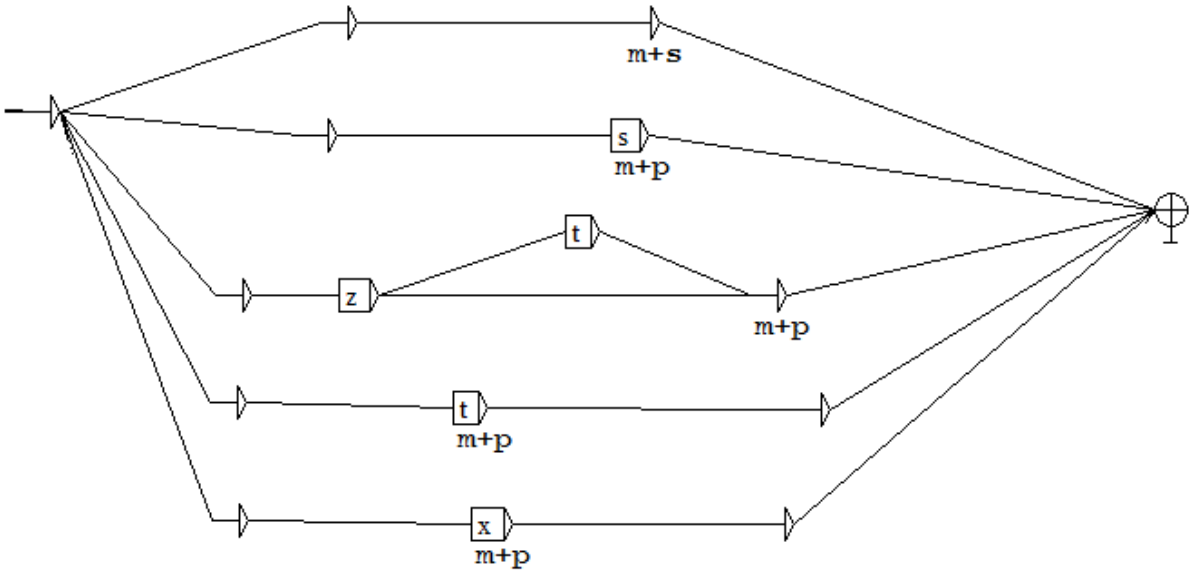


Figure 65. Graphe présentant principalement les marques du pluriel des noms communs

Ce graphe a été utilisé comme sous graphe appelé par « modèle_standard » qui a permis une factorisation des règles afin de décrire plus simplement et d'une manière plus lisible les

règles de flexions dans un graphe principal. Comme le montre la figure 66, un graphe principal a été développé permettant la description de certains noms communs qui se terminent par « u » afin de décrire les formes ayant un « l » ou une transformation de « u » final en « l » avant les marques du pluriel.

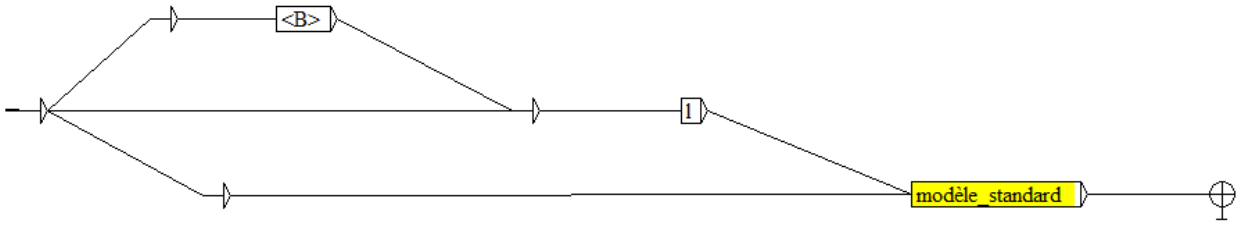


Figure 66. Factorisation des règles pour décrire des noms communs qui se terminent par « u »

D'autres modèles de flexion plus complexes ont été déterminés. Nous montrons dans ce qui suit un exemple d'un graphe décrivant la flexion des noms communs qui se terminent par « eur » comme « jureur » ou « honneur ». Notre système a permis le regroupement des règles pour cette classe des noms communs. Nous citons :

- la règle de flexion « <B3>ours » qui permet de fléchir « accuseur » en « accusours ».
- la règle de flexion « <B3>oirx » qui transforme « accuseur » en « accusoirx ».

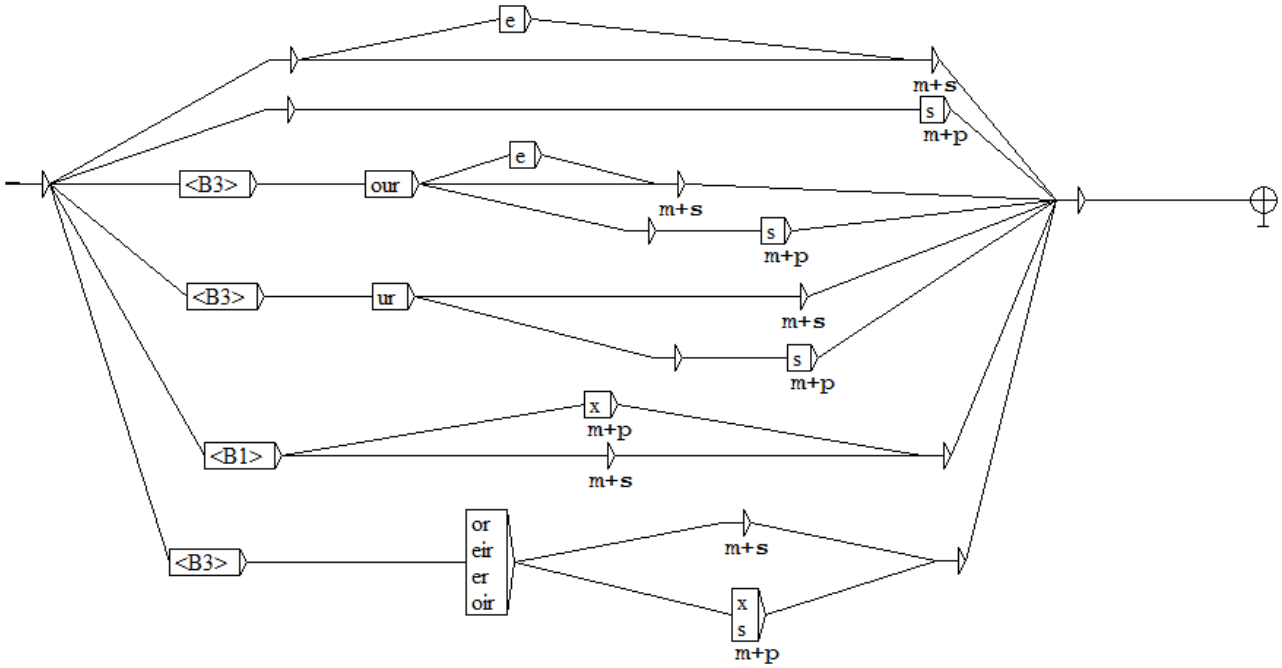


Figure 67. La flexion des noms communs qui se terminent par « eur »

3.3.3 Compilation du dictionnaire électronique en moyen français

La phase de compilation de dictionnaire consiste à associer les paradigmes flexionnels et dérivationnels déterminés aux différentes entrées lexicales du dictionnaire afin de générer la

liste des formes fléchies potentielles. Prenons ainsi un exemple simple de l'entrée lexicale « *agriculteur*» :
agriculteur,NC+FLX=DOCTEUR+Profession

En utilisant le graphe de flexion de la figure 67 qui correspond au paradigme « DOCTEUR », le nom commun « *agriculteur* » qui n'est associé à aucune dérivation permet de générer 22 formes fléchies qui se présente comme suit :
agriculteurs,agriculteur,NC+m+p+Profession
agricultoure,agriculteur,NC+m+s+Profession

Les entrées générées par notre dictionnaire électronique contiennent plus d'informations que le dictionnaire standard à savoir :

- La forme fléchie générée par le compilateur
- La virgule « **,** » comme séparateur de champ ;
- Le lemme qui correspond à l'entrée lexicale ;
- La virgule « **,** » comme deuxième séparateur du champ ;
- La partie du discours : dans notre exemple « NC » correspond au nom commun ;
- Le séparateur « **+** » ;
- Le genre de nom « m » qui correspond au « masculin » ;
- Le nombre « p » pour le premier exemple qui correspond au pluriel ;

Certaines de ces formes ne sont pas attestées dans notre corpus. Cependant, elles peuvent exister dans d'autres textes en moyen français. En effet, les règles de flexion appliquées aux entrées lexicales ont été déterminées à partir des variantes qui sont de même classe grammaticale et qui ont la même flexion. Par exemple, à l'entrée lexicale « agriculteur », les règles de flexion issues des entrées lexicales « docteur » et « laboureur » sont appliquées. Par conséquent, notre dictionnaire électronique une fois compilé a une couverture beaucoup plus large que le dictionnaire standard. En effet, la liste des formes fléchies générée automatiquement contient environ 450.200 formes fléchies alors que le dictionnaire standard contient environ 160.000 formes. En effectuant des tests sur des textes de genres différents de notre corpus « MEDITEXT » à savoir des textes de théâtre, notre dictionnaire des formes fléchies a pu couvrir environ 96% du vocabulaire.

4. Conclusion

Dans ce chapitre, nous avons fait le tour d'horizon des principales caractéristiques des dictionnaires électroniques les plus répondus à savoir le réseau lexical WordNet, le standard *Lexical Markup Framework* LMF, les dictionnaires DELA et les dictionnaires NooJ. Avec sa méthode de compilation et les multiples opérateurs proposées, les dictionnaires NooJ présentent un intérêt pour une description des éléments du vocabulaire ayant une orthographe non-normalisée qui s'inscrivent dans une approche de formalisation des langues pour répondre à des besoins applicatifs, constituant plus précisément des outils d'annotation automatique du corpus.

Plusieurs modules de notre processus de construction du dictionnaire électronique du moyen français ont été automatisés afin de minimiser l'intervention humaine qui est limitée à la validation manuelle de quelques milliers d'entrées de notre dictionnaire et à l'annotation semi-automatique de quelques textes de notre corpus. Ce processus a été conçu de manière à permettre un enrichissement incrémental du dictionnaire. En effet, à partir d'une liste des formes les plus fréquentes de MEDITEXT, nous avons interrogé des ressources linguistiques à savoir le « dictionnaire du moyen français » DMF et l'*Anglo-normand dictionnary* afin d'associer les différentes formes attestées à des entrées lexicales définies par ces ressources. Ce dictionnaire a été enrichi automatiquement par des nouvelles variantes qui figurent principalement dans le DMF et dans l'*Anglo-Normand dictionary*. Ensuite, nous avons appliqué une méthode itérative et incrémentale afin de constituer un dictionnaire standard en exploitant des textes du MEDITEXT annotés semi-automatiquement.

Nous avons mis en place un système de classification automatique qui, à partir des entrées lexicales associées à des règles de flexion, permet de regrouper un ensemble des entrées lexicales d'une même classe grammaticale ayant la même flexion. Finalement, grâce au compilateur des dictionnaires NooJ, nous avons procédé à une génération automatique des formes fléchies qui a permis de produire des formes non-attestées qui peuvent cependant potentiellement exister dans des textes en moyen français. Finalement, ce dictionnaire compilé est utilisé par un analyseur NooJ afin d'annoter chaque forme du texte avec des informations linguistiques essentiellement un lemme et une partie du discours. Cette annotation hors contexte, aussi appelée « étiquetage a priori », est une des étapes indispensables pour toute application d'annotation automatique telle que l'étiquetage morphosyntaxique et la reconnaissance des entités nommées.

Chapitre 5

Analyse lexicale des textes en moyen français

1. Introduction

L'analyse lexicale est le premier niveau d'analyse des données textuelles et elle est souvent considérée comme indispensable pour faciliter, voire rendre possible des traitements à mettre en œuvre ultérieurement. En effet, les traitements applicatifs ne se contentent pas d'établir une correspondance entre une forme dans un texte et une entrée lexicale dans le dictionnaire électronique standard de la langue mais ils nécessitent des analyses plus fines des ALUs afin de proposer tous les découpages potentiels ainsi que toutes les descriptions morphosyntaxiques correspondantes. Le but d'un analyseur lexical est donc d'identifier les unités du vocabulaire du texte auxquelles on associe un ensemble de descriptions morphosyntaxiques correspondantes (Silberztein, 2015). Concrètement, il s'agit de mettre en place un ensemble de traitements automatiques qui cherchent à proposer toutes les segmentations possibles d'une séquence de caractères et à y associer les informations linguistiques correspondantes. Ces traitements dépendent principalement de la langue, de la nature des textes à analyser et des traitements applicatifs à réaliser.

La mise en place d'un analyseur lexical repose donc sur le développement de ressources linguistiques spécifiques (Pierrel, 2000). En plus des ressources linguistiques que sont les dictionnaires électroniques, les grammaires morphologiques lexicales et productives et les grammaires locales, notre méthode d'analyse lexicale fait appel aux analyseurs NooJ qui, selon le phénomène linguistique à observer, offrent la possibilité d'effectuer des traitements mono-niveau d'une façon séquentielle ou des traitements multi-niveaux en cascade. Les analyseurs NooJ permettent de gérer l'ordre de priorité quant à l'application des ressources linguistiques et au choix du processus d'analyse effectuée selon la taille des séquences reconnues par la grammaire. En effet, ils disposent de méthodes de représentation et de stockage des données afin de rendre possible l'accès aux analyses effectuées et la récupération des informations de la TAS. Comme le montre le schéma de la figure 68, notre analyseur lexical des textes en moyen français se décompose en trois phases à savoir (i) des analyses typographiques qui permettent la résolution des problèmes liés à l'utilisation des caractères comme les signes de ponctuation et l'apostrophe, (ii) la reconnaissance des mots simples qui

propose une ou plusieurs segmentations du texte en suite de formes et (iii) la reconnaissance des mots composés qui permet le regroupement d'une séquence d'ALU. À l'issue de ces opérations, toutes les hypothèses lexicales et les ambigüités potentielles sont représentées dans la TAS.

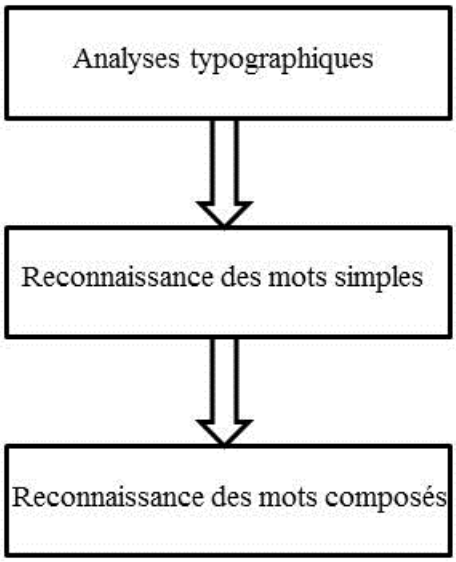
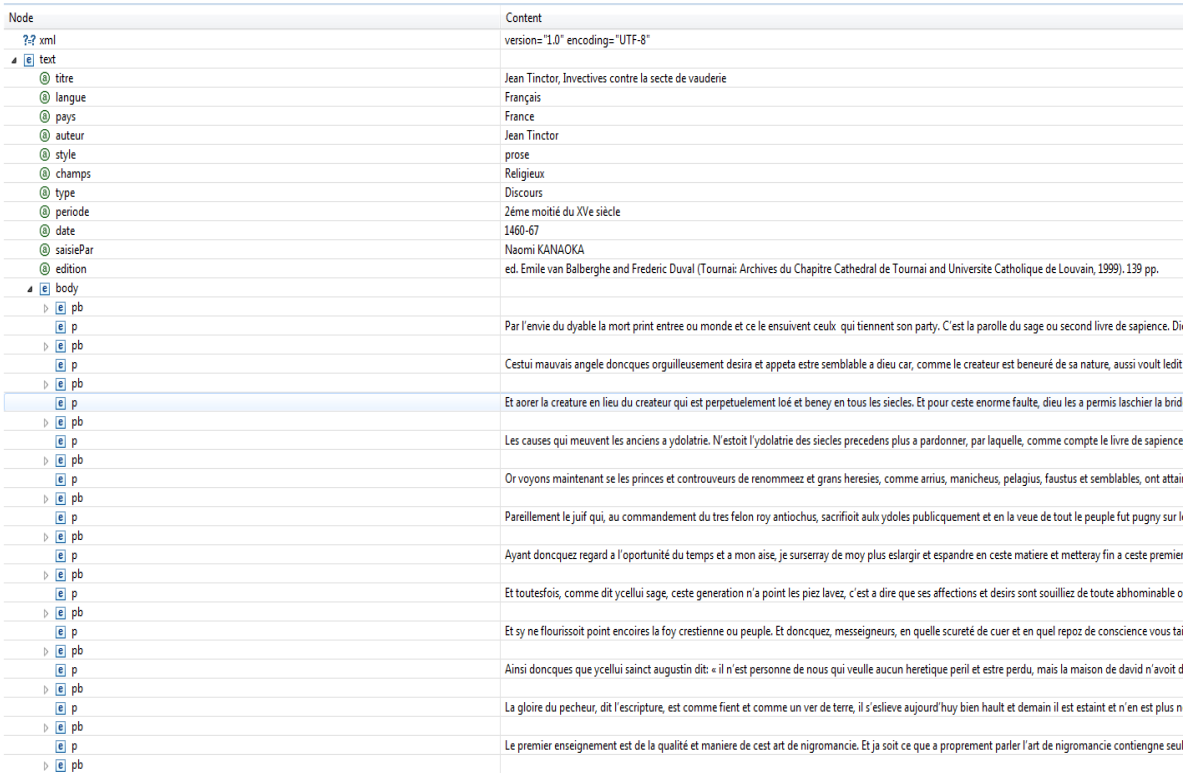


Figure 68. Les phases de l'analyse lexicale

2. Analyse typographique

Revenant au MEDITEXT, les caractères de ce dernier sont représentés grâce à l'encodage universel des caractères, UNICODE et ils sont stockés dans des fichiers au format XML où chaque fichier correspond à un texte. Chaque texte est organisé à l'aide d'une structure arborescente mise en place en utilisant les standards et les recommandations du XML-TEI. Comme le montre l'exemple de la figure 69, cette structuration des données permet une séparation entre les métadonnées et le contenu du texte. En effet, les métadonnées présentent plusieurs informations qui décrivent le texte telles la langue, la période, la date, l'origine, l'auteur, l'édition et le genre textuel, tandis que le contenu du texte regroupe, lui, toutes les informations qui permettent de restituer ce dernier, à savoir des informations de formatage qui conservent la mise en forme, des informations sur la structure qui préservent la pagination et les différents niveaux des titres, ainsi que d'autres informations qui apparaissent dans les textes plus spécifiquement les commentaires, les notes et les références bibliographiques.



Node	Content
?-? xml	version="1.0" encoding="UTF-8"
text	
titre	Jean Tinctor, Invectives contre la secte de vauderie
langue	Français
pays	France
auteur	Jean Tinctor
style	prose
champs	Religieux
type	Discours
periode	2éme moitié du XVe siècle
date	1460-67
saisiePar	Naomi KANAOKA
edition	ed. Emile van Balberghe and Frederic Duval (Tournai: Archives du Chapitre Cathedral de Tournai and Universite Catholique de Louvain, 1999). 139 pp.
body	
pb	
p	Par l'envie du dyable la mort print entree ou monde et ce le ensuivent ceulx qui tiennent son party. C'est la parolle du sage ou second livre de sapience. Di
pb	
p	Cestui mauvais angele doncques orguilleusement desira et appeta estre semblable a dieu car, comme le createur est beneuré de sa nature, aussi voult ledit
pb	
p	Et aorer la creature en lieu du createur qui est perpetuelement loé et beney en tous les siecles. Et pour ceste enorme faulte, dieu les a permis laschier la brid
pb	
p	Les causes qui meuvent les anciens a ydolatrie. N'estoit l'ydolatrie des siecles precedens plus a pardonner, par laquelle, comme compte le livre de sapience
pb	
p	Or voyons maintenant se les princes et controuveurs de renommeez et grans heresies, comme arrius, manicheus, pelagius, faustus et semblables, ont attai
pb	
p	Pareillement le juif qui, au commandement du tres felon roy antiochus, sacrifioit aulx ydoles publicquement et en la veue de tout le peuple fut pugny sur l
pb	
p	Ayant doncquez regard a l'oportunité du temps et a mon aise, je surserray de moy plus eslargir et espandre en ceste matiere et metteray fin a ceste premie
pb	
p	Et toutesfois, comme dit ycellui sage, ceste generation n'a point les piez lavez, c'est a dire que ses affections et desirs sont souilliez de toute abhominable o
pb	
p	Et sy ne flourissoit point encoires la foy crestienne ou peuple. Et doncquez, messeigneurs, en quelle scureté de cuer et en quel repoz de conscience vous tai
pb	
p	Ainsi doncques que ycellui sainct augustin dit: « il n'est personne de nous qui veulle aucun heretique peril et estre perdu, mais la maison de david n'avoit d
pb	
p	La gloire du pecheur, dit l'escripture, est comme fient et comme un ver de terre, il s'eslieve aujourd'huy bien hault et demain il est estaint et n'en est plus n
pb	
p	Le premier enseignement est de la qualité et maniere de cest art de nigromancie. Et ja soit ce que a proprement parler l'art de nigromancie contiengne seul
pb	

Figure 69. Structuration des textes en XML-TEI

A ce stade, nous nous intéressons à mettre en place une structure qui dépasse la contrainte imposée par la TEI qui consiste à associer une annotation unique à chaque séquence de caractère. Ainsi, afin de répondre aux spécificités du moyen français qui sont principalement la non-standardisation de l'orthographe et l'aspect évolutif de la morphologie comme de la syntaxe, nous adoptons pour structurer un texte la structure d'annotation de texte (TAS) qui supporte diverses analyses potentielles effectuées de façon massive sur le contenu textuel. Par conséquent, nous avons analysé et interprété automatiquement le métalangage du XML-TEI afin de reconstituer le contenu textuel à analyser. En effet, nous avons exploité la structuration du texte en conservant toutes les informations qui peuvent nous être utiles pour des analyses automatiques ultérieures. De ce fait, nous avons transformé en format textuel des éléments du métalangage XML-TEI principalement les marques de fins de pages, de paragraphes et de phrases ainsi que les différents niveaux de titres. Comme l'illustre la figure 70, les traitements effectués ont permis de transformer la structure XML-TEI du texte en une séquence stricte de caractères alphabétiques et non alphabétiques avec une première segmentation en phrases marquée par des retours à la ligne. Ces textes générés automatiquement ont été employés pour l'élaboration du corpus MEDITEXT en format (.noc) adopté par la structure d'annotation du texte (TAS) dans laquelle les analyses des différents niveaux linguistiques ont été effectuées.

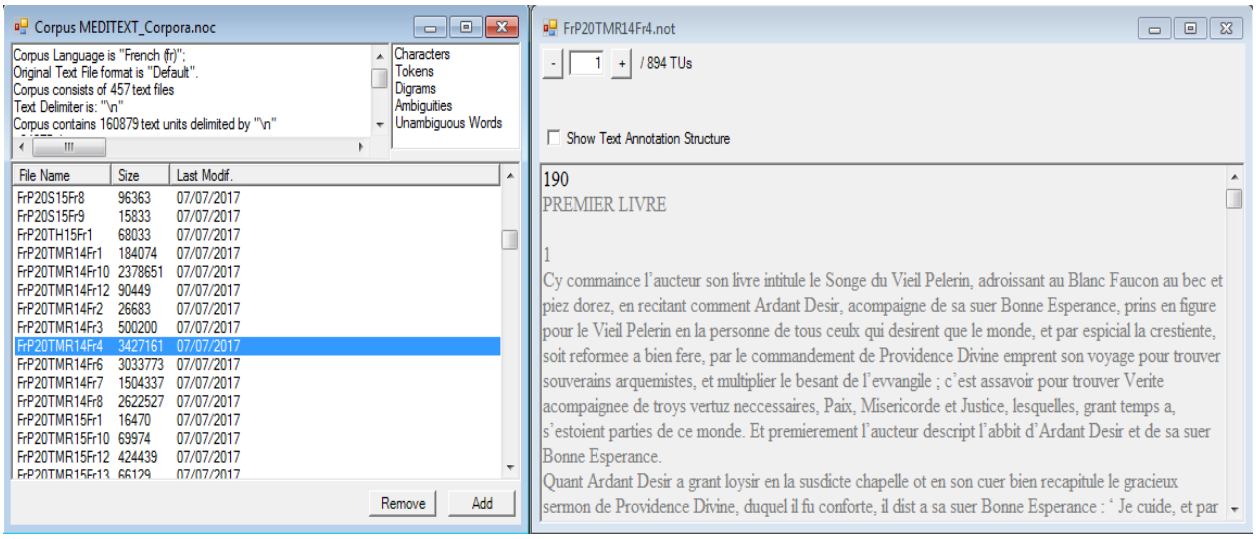


Figure 70. Importation du contenu textuel et constitution du corpus .noc

Nous signalons que la mise en forme n'était pas systématiquement établie lors de la numérisation des textes, du fait du manque ou de l'absence d'informations sur la structure dans le texte d'origine ou du fait de l'insuffisance des efforts accomplis pour la restitution de la structure originale. Par conséquent, une analyse typographique nous a semblé nécessaire à ce niveau afin de traiter les signes de ponctuation et des signes non alphabétiques comme l'apostrophe pour faciliter la reconnaissance des ALU. Ces différentes analyses, illustrées par la figure 71, permettent donc une première phase de segmentation en marquant automatiquement les fins de phrases, les nombres et l'apostrophe. Finalement, l'analyse typographique établie est capable de traiter des situations qui n'ont pas été prises en compte lors du formatage des textes en TEI.

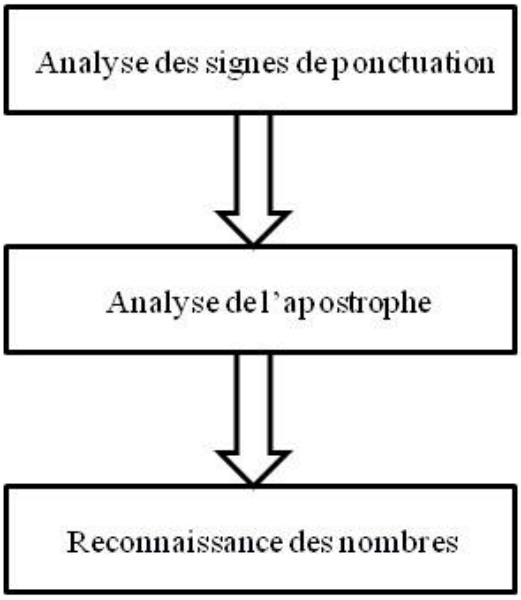


Figure 71. Les phases d'analyse typographique

2.1. L'analyse des signes de la ponctuation et la segmentation en phrases

Les signes de ponctuation sont un ensemble de caractères non-alphanumériques qui permettent d'ordonner le texte en le découpant en unités élémentaires et en marquant les pauses. Chacune de ces unités élémentaires peut correspondre à une « phrase ». Mais en linguistique, la phrase est une notion floue et mal définie malgré les nombreuses tentatives qui essaient de lui attribuer une définition autonome (C. Marchello-Nizia, 1977). Elle reste cependant souvent considérée comme l'unité supérieure de l'analyse linguistique. Dans une perspective d'analyse automatique, nous considérons qu'une phrase est un segment textuel qui résulte du découpage d'un texte en se basant sur les signes de ponctuation. Ce découpage détermine donc la suite du traitement automatique du texte (Friburger et al., 1993).

On distingue deux types de signes de ponctuations, terminaux et non terminaux :

- Les signes non-terminaux servent à indiquer les pauses courtes et à séparer les parties d'une phrase telle que la virgule, les deux points, les guillemets, les parenthèses et les crochets.

- Les signes terminaux servent à marquer les fins de phrases comme le point, le point d'interrogation, le point d'exclamation, le point-virgule, les trois points de suspension et le saut de paragraphe.

Nous rappelons que, lors de la numérisation du MEDITEXT, un découpage du texte en chapitres, pages, paragraphes et phrases a été effectué. Ce découpage a été automatiquement sauvegardé et il a été restitué au format textuel afin d'éviter les éventuelles erreurs d'une segmentation automatique causées généralement par l'ambiguïté de certains signes de ponctuation. Nous constatons qu'environ 70 % des marques de fin de phrase n'ont pas été prises en considération. En effet, le découpage des paragraphes en phrases n'a pas été toujours respecté dans la phase de constitution du corpus, d'où le besoin d'une segmentation automatique qui permette de meilleures analyses ultérieures.

La ponctuation dépend du type de texte, car la ponctuation du vers est bien différente de celle de la prose. Cependant, notre système est utilisé pour segmenter en phrases des textes des deux types. En effet, les vers sont marqués par des retours à la ligne systématique généralement précédés par un signe de ponctuation. Mais ces retours à la ligne permettent une segmentation des phrases prosodique et rythmique plutôt que syntaxique et sémantique. Dans un vers, le point joue principalement le rôle de séparateur terminal que ce soit pour des raisons rythmiques ou syntaxiques. Pour la prose comme pour le vers, nous constatons que le point d'exclamation et le point d'interrogation, souvent ajoutés par le copiste ou l'éditeur,

sont des marques de fins de phrases. En outre, le point-virgule et les trois points de suspension lorsqu'ils sont suivis d'une majuscule marquent la fin d'un segment textuel.

La majorité des textes de notre corpus environ 96 % contiennent trois ou quatre signes de ponctuation. Le point, la virgule et la majuscule sont les plus présents, avec une utilisation du point-virgule pour les textes juridiques et l'utilisation des guillemets pour les discours. La fréquence d'usage des signes de ponctuation varie d'un texte à l'autre mais nous considérons qu'elle est faible par rapport à la fréquence d'usage des textes en français moderne. La grammaire illustrée à la figure 72 permet donc d'ajouter une balise <SENTENCE> pour marquer les fins de phrases en analysant les signes de ponctuation fréquents et non-ambigüs.

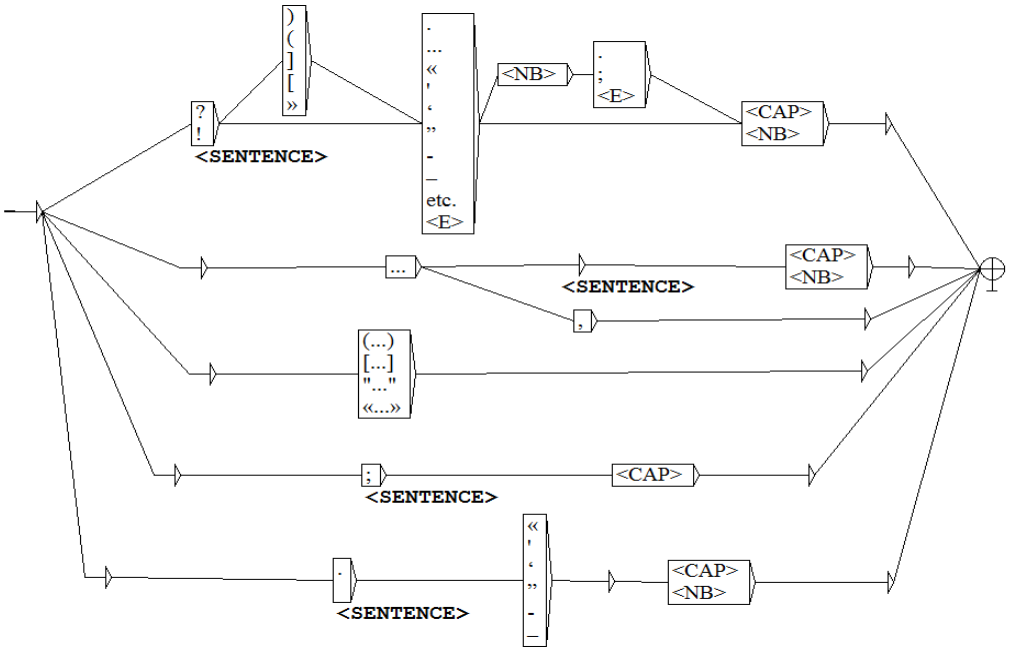


Figure 72. Segmentation en phrase selon les signes de la ponctuation

Cette grammaire ne permet pas de traiter l'ambiguïté posée par certains signes de ponctuation. Citons l'exemple du point, qui est le signe de ponctuation le plus utilisé, mais pour lequel on distingue quatre cas qui peuvent aboutir à des ambiguïtés dans MEDITEXT :

- Expressions anthroponymiques qui contiennent des abréviations dans la civilité, titre, profession, prénom ou nom d'une personne par exemple *M. Nicolas-de-Plancy* ou *M. de Bourbon* ou *J. Menon*.

- Les nombres en chiffres romains peuvent contenir des points à titre d'exemple le nombre *III.C.XX.IIII* utilisé pour exprimer l'âge d'un seigneur ou *C. livres* qui permet de designer une quantité monétaire.

- Les nombres arabes peuvent contenir des points lorsqu'ils sont imbriqués dans une structure comme montre l'exemple suivant : un nombre qui fait partie d'une expression de date *1428. 29 juin*.

- Les abréviations sont fréquemment utilisées dans les textes juridiques et parlementaires. À titre d'exemple, nous mentionnons *art. 1* pour désigner le premier article.

Au contraire du français moderne, le point peut marquer la fin d'une phrase sans être suivi d'une majuscule. En effet, on a répertorié de nombreux cas où le point joue le rôle d'un délimiteur de phrase bien qu'il soit suivi d'une lettre minuscule. À l'inverse, dans certains textes, on remarque l'absence de point malgré la présence d'une majuscule qui marque le début d'une phrase. En guise de conclusion, certains cas ambigus ne peuvent être traités qu'après une analyse lexicale, morphosyntaxique et sémantique plus profonde. Cependant, notre grammaire, qui se base sur les signes de ponctuation et un ensemble de caractères spéciaux, nous a permis d'atteindre une segmentation du texte en phrases avec un taux de réussite d'environ 87 %.

2.2. Analyse de l'apostrophe

L'apostrophe n'est pas considérée comme un signe de ponctuation (Silberztein, 2015). Dans notre corpus, elle sert essentiellement de signe typographique formant des mots composés ou marquant l'élision de la lettre finale.

La grammaire de la figure 73 montre l'analyse des différents cas d'utilisation de l'apostrophe. En effet, les trois sous-graphes, qui composent cette grammaire, correspondent aux trois utilisations différentes de l'apostrophe que nous avons recensées dans MEDITEXT à savoir, l'identification d'élision, la formation des mots composés et l'usage du simple guillemet.

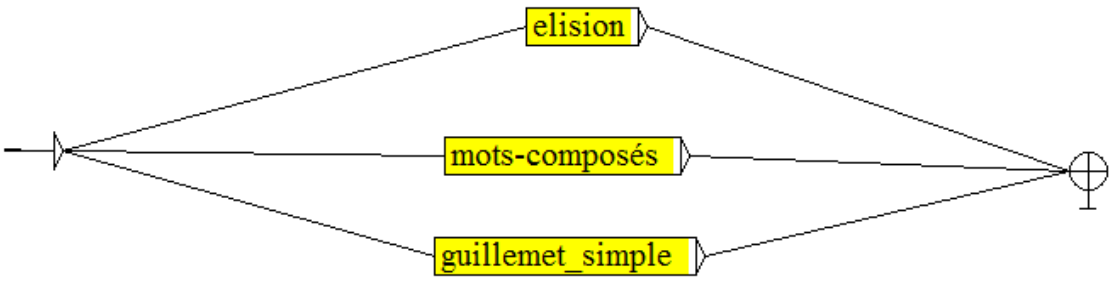


Figure 73. Grammaire d'analyse de l'apostrophe

- L'identification de l'élision consiste à reconnaître un phénomène qui se produit entre deux formes en contact lorsque la lettre finale de la première forme est remplacée par une

apostrophe et n'est pas suivie d'un espace (Silberztein, 2015). Ainsi l'élaboration d'un dictionnaire qui recense les différentes formes qui peuvent être élidés, permet d'analyser l'élision afin de segmenter les deux formes. En français moderne, l'apostrophe ne concerne que 19 formes (Silberztein, 20015). Tandis qu'en moyen français, nous avons recensé un dictionnaire de 36 formes élidées dont nous exposons quelques entrées à la figure 74.

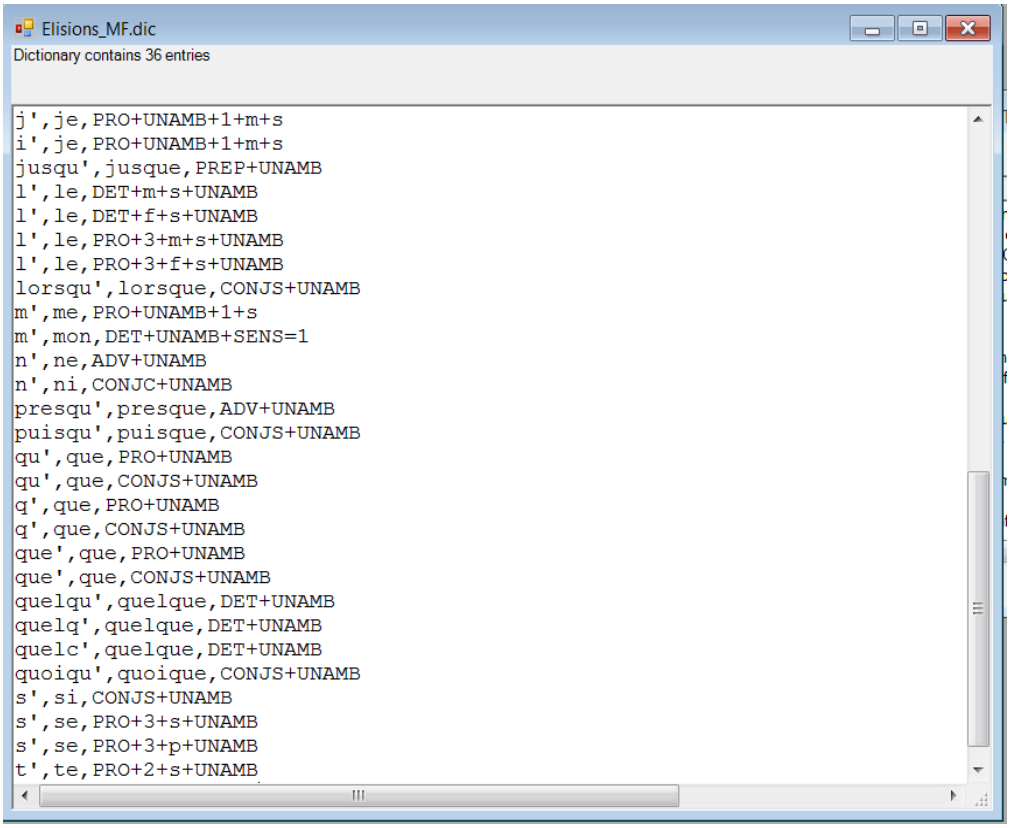


Figure 74. Dictionnaire des formes élidées en moyen français

Dans le cas d'une élision, les formes élidées contenues dans le dictionnaire sont donc suivies d'une forme ayant une initiale vocalique. A titre d'exemple : *l'umain*, *d'aultres*, *qu'ilz*, etc. Certaines formes comme « *puisqu* », « *presqu* » et « *lorsqu* » ne correspondent pas à des ALUs. Par conséquent, elles ne sont pas ambiguës lorsqu'elles sont suivies par une apostrophe et leur annotation est possible grâce au dictionnaire électronique des formes élidées. Cependant, d'autres formes comme « *l*» et « qu' » peuvent être ambiguës. La reconnaissance et la désambiguïsation de ces formes nécessitent le développement d'une grammaire qui utilise le dictionnaire des formes élidées et les entrées du dictionnaire du moyen français. La grammaire de la figure 75 permet la désambiguïsation de la forme élidée « l' ». En effet, cette grammaire permet de lever l'ambiguïté de la forme « l' » en lui attribuant l'étiquette d'un déterminant <DET> lorsqu'elle est suivie par un nom ou un adjectif ou un

pronom ou en lui annotant comme pronom <PRO> lorsqu'elle est suivie d'un verbe ou d'une forme « en » ou « y » suivie d'un verbe.

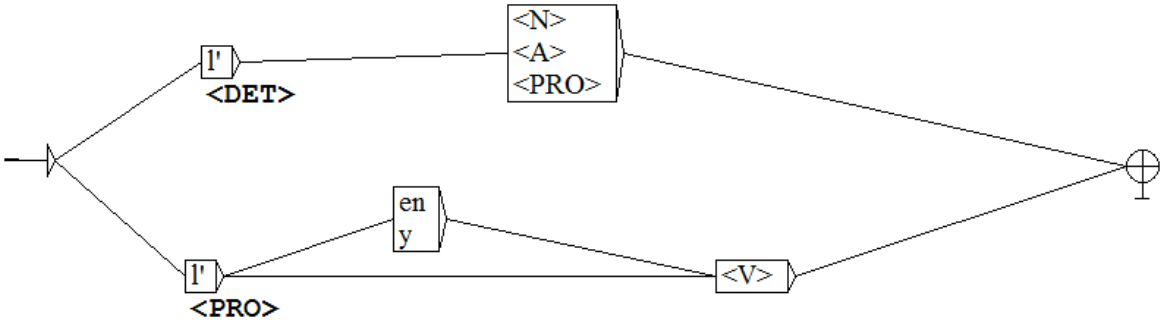


Figure 75. Reconnaissance et désambiguïsation de la forme « l' »

- Nous constatons que l'apostrophe peut aussi jouer un rôle orthographique comme celui du trait d'union pour former des mots composés, en notant une liaison interne ou une fausse coupe des formes, comme dans le cas des formes « *aujourd'hui* », « *prud'homme* », « *s'entr'aimer* » et « *entr'ainer* ». Ces formes correspondent donc à des ALUs. Certaines d'entre elles sont traitées en les recensant dans le dictionnaire comme pour l'ALU « *aujourd'hui* » et ses différentes variantes. Lorsqu'il s'agit d'un mot composé, ils peuvent être, dans certains cas, reconnus à l'aide d'une grammaire locale comme le montre l'exemple ci-dessous de la figure 76 qui fait apparaître les différentes variantes de « *prud'homme* ».

Figure 76. Grammaire de reconnaissance de la forme composée « prud'homme »

- L'apostrophe peut être utilisée comme guillemet pour signaler un exemple « *le 'chevalier'* ». Le sous-graphe « guillemet_simple » de notre grammaire traite cette utilisation en identifiant les cas d'utilisation des apostrophes comme des guillemets. De fait certains de ces cas sont traités lors de l'analyse par notre graphe de la figure 72 qui permet la segmentation du texte en phrase. A ce stade, il s'agit d'identifier les cas où les apostrophes sont généralement utilisées à l'intérieur d'une phrase. Dans ce cas le guillemet ouvrant apparaît obligatoirement au début d'un mot ; le guillemet fermant apparaît obligatoirement, lui, à la fin d'un mot. Cette reconnaissance peut poser des problèmes d'ambiguïté puisque certaines formes comme « q » peut être ou non suivie d'une apostrophe. En effet, en moyen français, « q » peut signifier la forme « que » ou « qu' » lorsqu'elle est suivie d'une apostrophe. Par conséquent, comme montre la grammaire de la figure 77, afin d'éviter de générer des ambiguïtés supplémentaires lors de l'analyse, nous avons pris le parti d'annoter

les apostrophes comme des guillemets lorsque l'apostrophe ouvrante et l'apostrophe fermante ne sont pas suivie d'une forme élidée.

Figure 77. Grammaire identifie l'utilisation des apostrophes comme guillemets

2.3. Reconnaissance des nombres

Les nombres sont formés d'un petit nombre de chiffres. Cependant, ils peuvent constituer un ensemble varié et infini d'éléments. Par conséquent, dresser une liste exhaustive de tous les nombres, qui peuvent apparaître dans un texte, est une tâche qui s'est révélée irréalisable. En effet, le recours aux développements des grammaires morphologiques et locales est une solution plus adaptée qu'un simple accès aux dictionnaires électroniques pour la reconnaissance des nombres. Dans les textes en moyen français, les nombres apparaissent dans diverses expressions pour décrire et dénombrer divers objets à titre d'exemple les quantités monétaires, les expressions de temps, la numérotation des titres, la numérotation des articles dans les textes juridiques, etc.

Nous constatons l'utilisation des nombres cardinaux et ordinaux qui peuvent être décrits tant en chiffres arabes qu'en chiffres romains.

- Les chiffres arabes sont l'ensemble des chiffres décimaux {0,1,2,3,4,5,6,7,8,9} qui servent à décrire les nombres cardinaux et ordinaux. La reconnaissance de ces derniers est possible par la grammaire de la figure 78. En effet, NooJ dispose d'un symbole spécial <NB> qui permet la reconnaissance d'une séquence des chiffres arabes pour identifier les nombres cardinaux. En outre, la reconnaissance de tous les nombres ordinaux supérieurs à 9 est rendue possible grâce à l'utilisation du caractère spécial prédéfini <D> qui représente n'importe quel chiffre de 0 à 9. Finalement, les nombres ordinaux de 1 à 9 tels que 1er, 2ème et 3ème ont été traité en spécifiant les chiffres arabes. Tous les nombres ordinaux ont été annoté comme adjectif en utilisant le filtre <A+NA>.

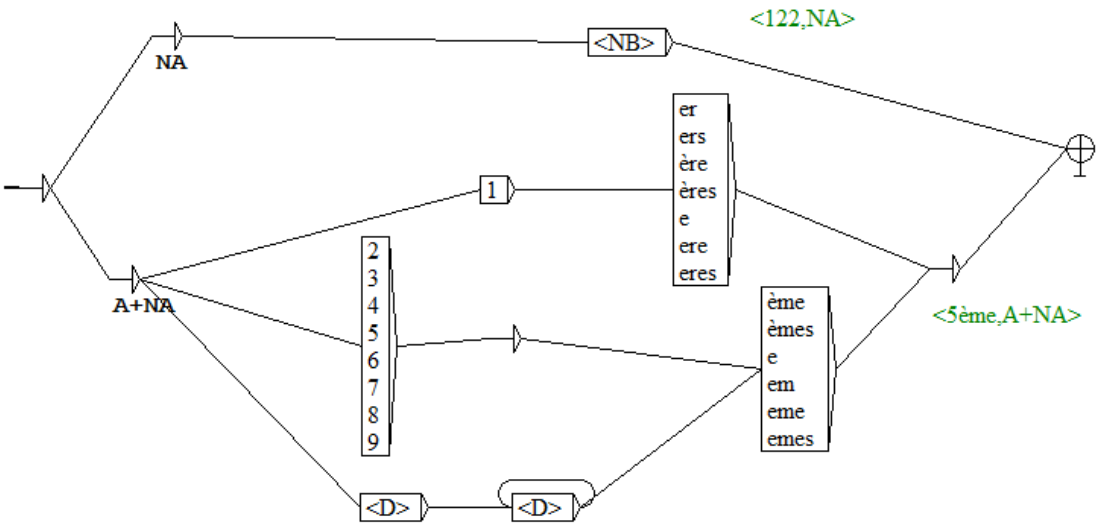


Figure 78. Reconnaissance des nombres cardinaux et ordinaux en chiffres arabes

- Le système de numérotation romain est dit additif du fait de qu'il combine des symboles alphabétiques { I, V, X, L, C, D et M} qui représentent respectivement les nombres 1,5,10,50 ,100,500 et 1000. Les nombres cardinaux apparaissent soit en majuscules soit en minuscules. Tandis que les nombres ordinaux sont généralement décrits par une cohabitation des majuscules et des minuscules. En effet, lorsque les chiffres romains en majuscules représentent des adjectifs ordinaux, ils sont suivis par un suffixe comme « *ème* », « *eme* », « *em* » ou « *e* ». Les nombres ordinaux en minuscules tels que « *xivem* » sont bien présents dans notre corpus mais peu nombreux. Notre grammaire de la figure 79 reconnait les nombres en chiffres romains cardinaux comme *XIV* et ordinaux comme *XIVème* ou *xivem*.

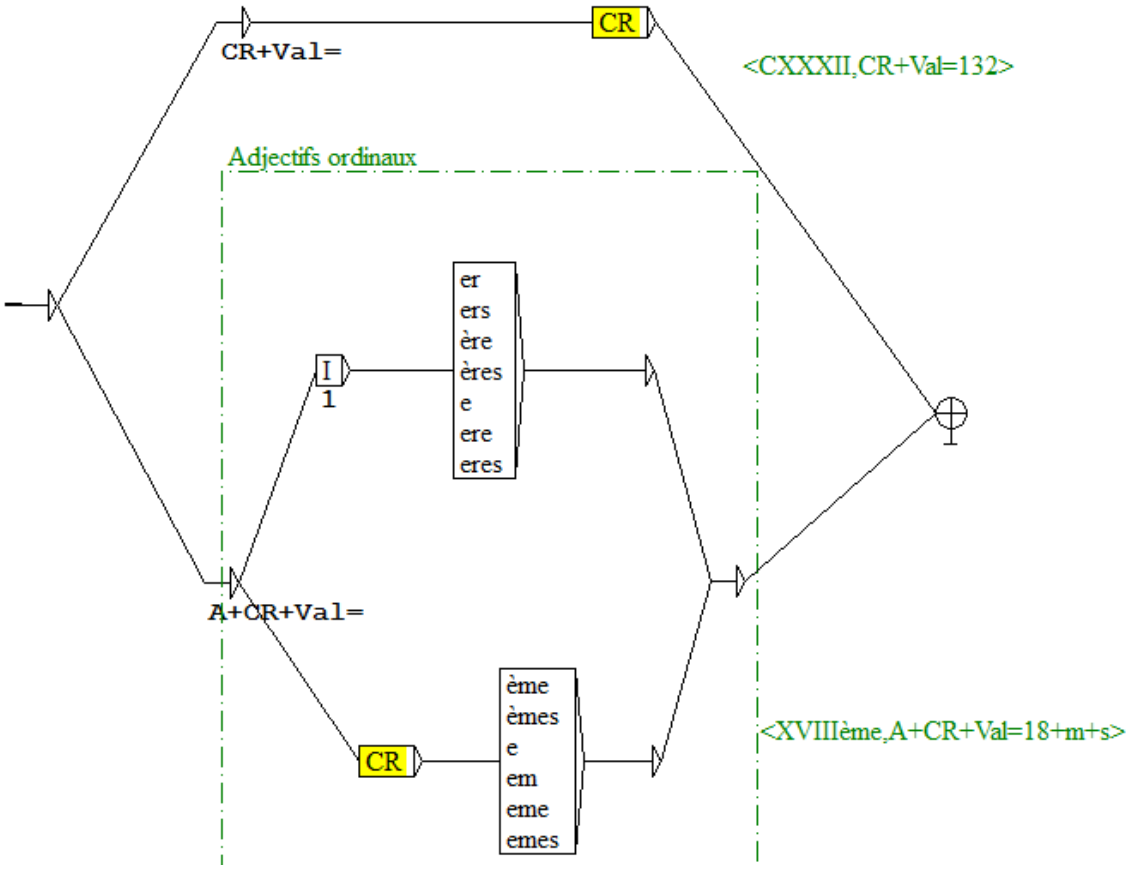


Figure 79. Reconnaissance des chiffres romains cardinaux et ordinaux

Le système de numérotation romaine est décrit par le sous-graphe *CR* qui combinent les unités, les dizaines, les centaines et les milliers.

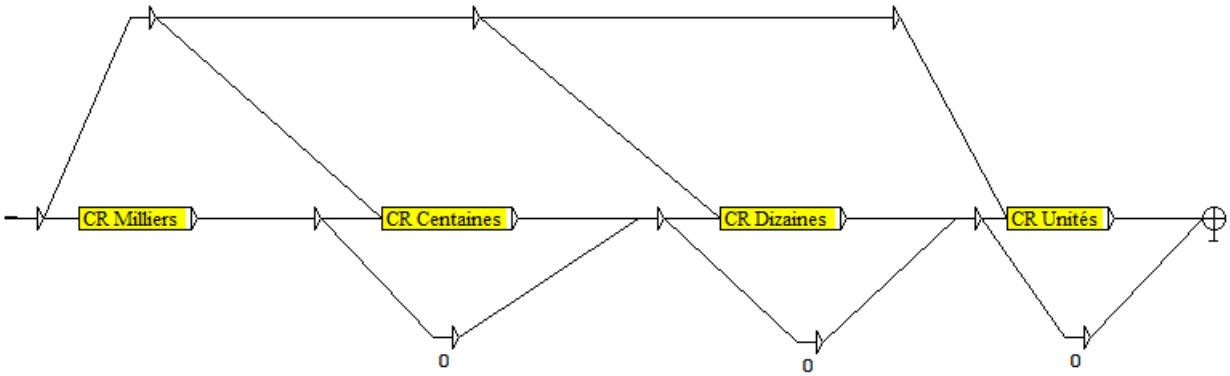


Figure 80. Reconnaissance des chiffres romains

Par exemple, la description des centaines représentée par le sous graphe « *CR Centaine* » de la figure suivante utilise trois symboles « *c* », « *d* » et « *m* » qui représentent respectivement les nombres 100, 500 et 1000. La grammaire fonctionne par une combinaison de symboles qui permet l'addition, à titre d'exemple, de C pour 100, de CC pour 200 et de DC pour 600 et la soustraction de CD pour 400 et de CM pour 900.

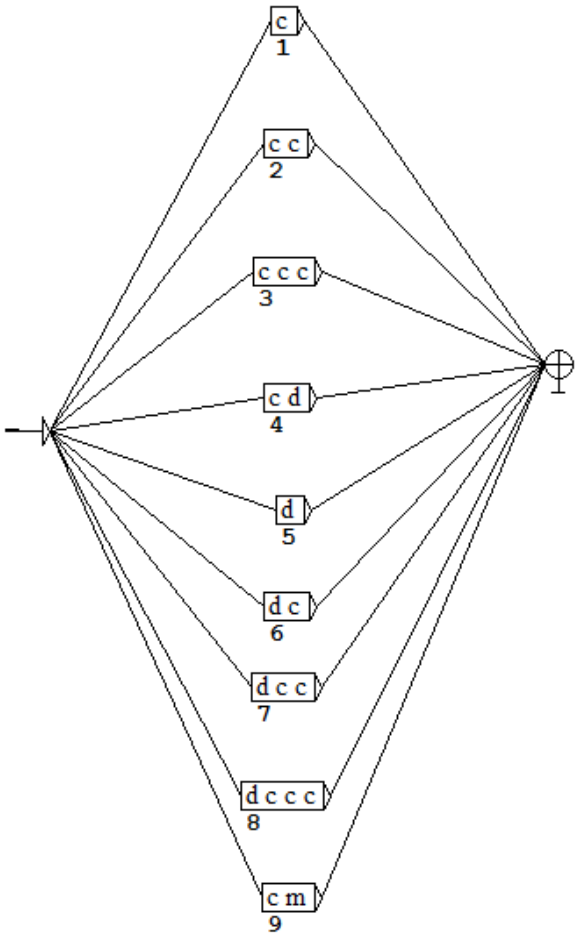


Figure 81. Reconnaissance des centaines

3. Reconnaissance des mots simples

La notion de mot soulève d'importants problèmes d'identification (Pierrel, 2000). Nous considérons qu'un mot simple est une ALU qui s'orthographie sous la forme d'une séquence de lettres et pouvant être prononcée en isolation avec un contenu sémantique ou pragmatique (Silberztein, 2015). En effet, le système d'écriture des textes en moyen français utilise une règle simple d'alternance formes/séparateurs pour segmenter les textes. De ce fait, l'identification d'une grande partie des mots simples est possible grâce à des expressions rationnelles qui extraient une séquence de lettres délimitée par deux séparateurs. Cependant, malgré les efforts de rationalisation du moyen français qui marquent une stabilisation débutante de l'orthographe, il existe des phénomènes linguistiques fréquents qui créent des exceptions faisant de la reconnaissance des mots simples une tâche non-triviale qui nécessite d'effectuer des analyses des phénomènes linguistiques comme la contraction, l'agglutination et la désagglutination.

3.1. Analyse des contractions

La contraction consiste à réunir au moins deux formes en une seule unité. En effet, cette combinaison des formes entraine une transformation de l'orthographe des formes utilisées en contraction. Le traitement de ce phénomène linguistique consiste à identifier les formes contractées par exemple « *au* » est la combinaison des formes contractées « *à* » et « *le* ».

Ce traitement améliore considérablement les performances des applications d'annotation automatique et des applications d'interrogation des textes. En effet, les contractions sont fréquentes dans les textes et ne pas en tenir compte produit soit un taux de silence considérable soit un important taux d'erreur qui diminue considérablement la précision du système. A titre d'exemple l'analyse de la contraction « *au* » est nécessaire à la recherche et à l'annotation de toutes les occurrences des groupes nominaux constitués d'un déterminant « *le* » suivi d'un nom commun « *NC* » pour analyser par exemple des séquences comme « *au roi* » et « *aux seigneurs* ».

Le moyen français se caractérise par une fréquente utilisation des contractions. Cependant, contrairement au français moderne, cet usage n'est pas stabilisé, rendant plus complexe l'automatisation de traitement.

3.1.1. Contraction avec « *à* »

La préposition « *à* » est souvent placée avant un complément pour décrire plusieurs situations telles qu'exprimer une localisation géographique, indiquer une distance, décrire une composition, des expressions temporelles comme la date et l'heure, mentionner un prix et exprimer des relations de contenance et d'appartenance. Nous avons recensé 4 variantes graphiques possibles de la préposition « *à* » à savoir « *à* », « *a* », « *as* » et « *ad* ».

Nous illustrons dans notre graphe de la figure 82 l'analyse de trois cas d'utilisation de la préposition « à » en contraction :

- lorsqu'elle est suivie d'un article défini masculin singulier « *le* » ou un article défini au pluriel « *les* » ;
- lorsqu'elle est suivie du pronom « lequel » au masculin singulier ou du pronom au pluriel « lesquels » ou « lesquelles » ;
- lorsqu'elle est suivie d'un article défini singulier masculin « le » ou d'un article défini au pluriel « les », eux-mêmes suivis de participe passé « dit ».

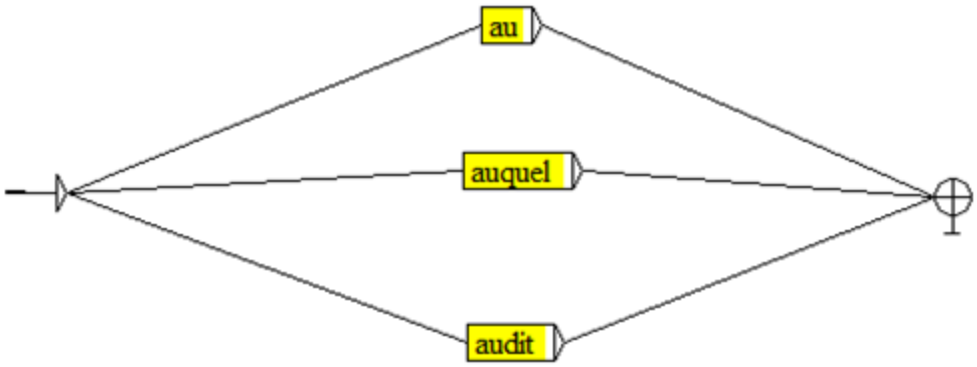


Figure 82. Contraction avec « *à* »

En français moderne, il existe deux occurrences « *au* » et « *aux* » qui résultent de la contraction de la préposition « *à* » avec les articles définis « *le* » et « *les* ». En moyen français, nous avons recensé trois variantes de l'occurrence « *au* » à savoir « *au* » et « *al* » et quatre variantes de l'occurrence « *aux* » à savoir « *aux* », « *aulx* », « *aus* » et « *as* ».

Notre grammaire de la figure 83 permet d'analyser la contraction de la préposition « *à* » avec les articles définis « *le* » et « *les* » en décomposant les variantes des occurrences « *au* » et « *aux* » en préposition « *à* » suivi du déterminant « *le* ». De plus, elle permet de lever d'éventuelles ambiguïtés posées par certaines variantes ou par la syntaxe.

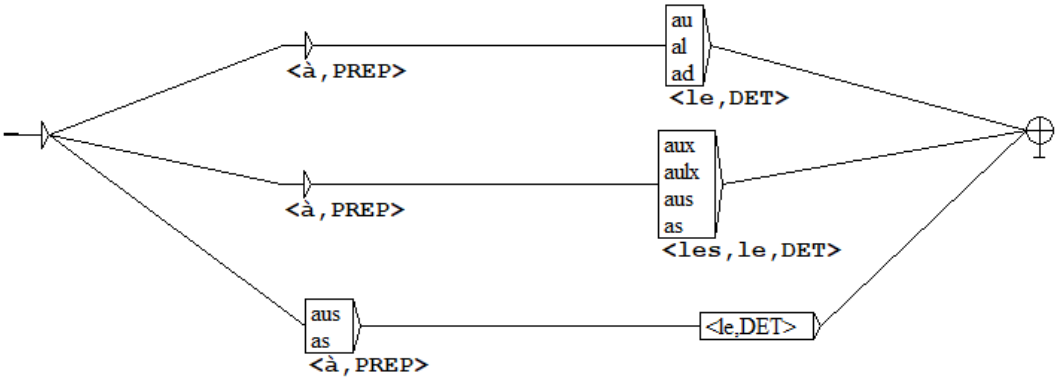


Figure 83. Analyse des contractions des variantes des occurrences « *au* » et « *aux* »

En effet, certaines variantes des occurrences « *au* » et « *aux* » comportent des ambiguïtés. A l'instar de la forme « *as* » dont les expressions suivantes illustrent deux différents contextes d'apparition. En effet, l'expression *(i)* montre l'apparition de « *as* » comme une variante de la préposition « *à* », tandis que l'expression *(ii)* illustre l'utilisation de « *as* » comme une variante de la forme « *aux* ». Notre grammaire permet de résoudre cette ambiguïtés en annotant la forme « *as* » par <*à, PREP*> lorsqu'elle est suivie par l'article définit « *le* ».

- establies as l'indignacion de nostre dit seigneur (i)
- Les heritages soient departiz as enfanz (ii)

Au contraire du français moderne, l'occurrence « *aux* » peut être suivie par le déterminant « *les* » comme montre l'exemple *(iii)*. Ces cas rares mais bien présents dans notre corpus sont traités par notre grammaire en annotant « *as* » par la préposition « *à* » et « *les* » comme déterminant « *le* ».

- mes l'en doit eschiver as les euvres humaines (iii)

Quand la préposition « *à* » introduit le pronom relatif « *lequel* » au masculin singulier ou au pluriel, elle génère l'une des trois occurrences suivantes : « *auquel* », « *auxquels* » et « *auxquelles* ». Nous avons recensées une liste non-exhaustive des variantes des occurrences « *auquel* », « *auxquels* » et « *auxquelles* », à savoir respectivement, 2 variantes, 11 variantes et 7 variantes.

La grammaire de la figure 84 permet donc d'analyser les formes contractées en les décomposant en deux unités à savoir la préposition « *à* » suivi du pronom relatif « *lequel* ».

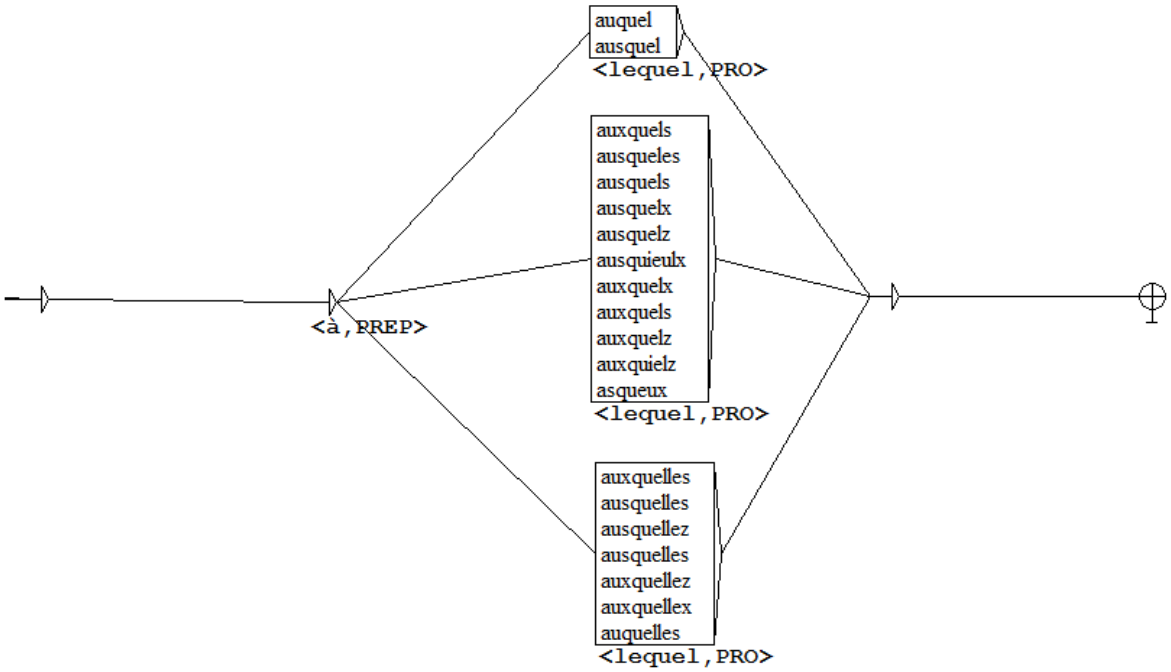


Figure 84. Analyse des contractions des variantes des occurrences « *auquel* », « *auxquels* » et « *auxquelles* »

De même pour les occurrences « *audit* », « *auxdits* » et « *auxdites* » qui résultent de la contraction de la préposition « à » avec le déterminant « le » suivi de participe passé « dit », la grammaire de la figure 85 liste les variantes de ces occurrences « *audit* », « *auxdits* » et « *auxdites* » et génèrent trois étiquettes qui correspondent à la succession de la préposition « *à* » <*à,PREP*>, du déterminant « *le* » <*le,DET*> et de participe passé « *dit* » <*dit,V*> et qui seront, chacun, sauvegardé dans la TAS.

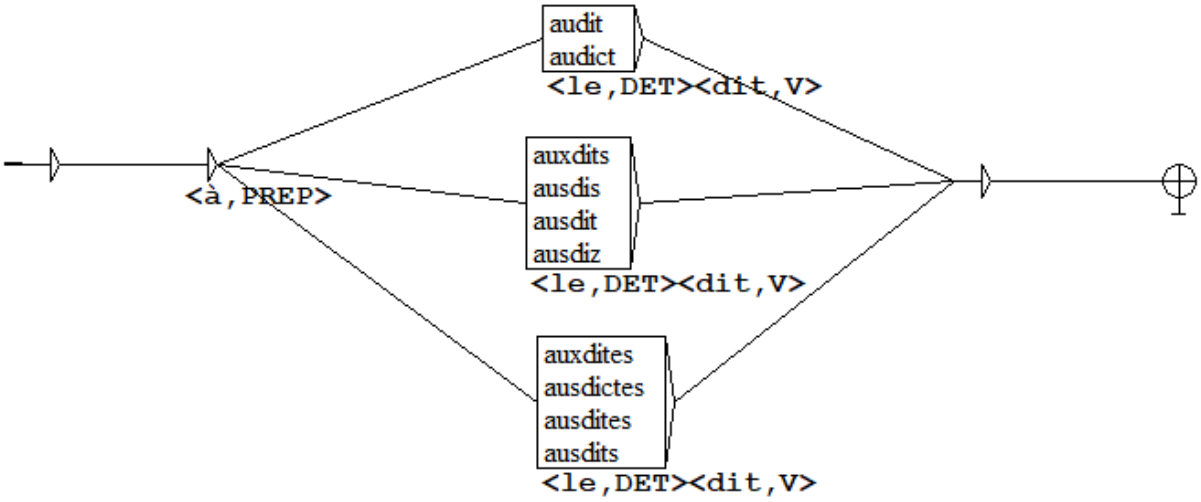


Figure 85. Analyse des contractions des variantes des occurrences « *audit* », « *auxdits* » et « *auxdites* »

3.1.2. Contraction en *« de »* et en « *du* »

A l'inverse du français moderne, l'utilisation des formes simples « *de* » et « *du* » n'est pas stable. En effet, la forme « *de* » peut être une variante de la forme « *du* » comme montre l'exemple *(i)* et vice versa la forme « *du* » peut être une variante de « *de* » comme illustre l'exemple *(ii)*.

- *Il embusoigne en ascun tiel Prince de Monde. (i)*
- *ainsné filz du Roy du France. (ii)*

De ce fait, nous considérons que les formes « *du* » et « *de* » peuvent avoir trois étiquettes possibles à savoir le déterminant partitif *du* <*du,DET*>, préposition *de* <*de,PREP*> et l'article indéfini *un* <*un,DET*>.

Les formes *« de »* et *« du »* peuvent être des occurrences qui résultent des formes contractées de la préposition « *de* » et du déterminant « *le* » ou des formes contractées de la préposition « *de* » et du déterminant « *du* ».

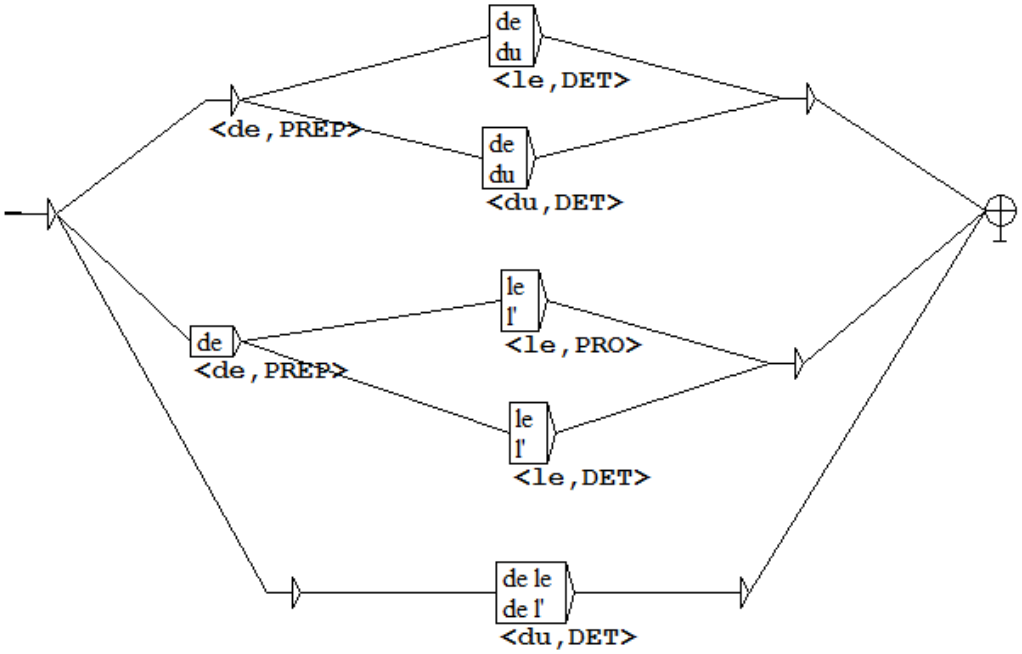


Figure 86. Analyse des contractions des occurrences « de » et « du »

Cependant, quand les formes *« de »* et *« du »* sont suivies par le pronom *« le »*, elles seront annotées en proposition avec l'étiquette *<de,PREP>*. On constate que la séquence « *de la* » est bien présente dans notre corpus. Mais, la forme « *la* », lorsqu'elle est annotée en pronom ou en déterminant, n'est jamais précédé par la préposition « *du* ». Il existe trois analyses possibles de la séquence ambiguë « *de la* ». Cette dernière peut correspondre à une préposition « *de* » suivie soit du pronom « *la* » soit du déterminant « *la* » comme elle peut être la forme féminine du déterminant partitif « *du* ».

Si le déterminant partitif « *du* », ou sa forme au féminin « *de la* », sont suivis d'un mot à initiale vocalique, il s'élide en « *de l'* ». En effet, de même que la séquence « *de la* », notre grammaire de la figure 86 permet de générer trois analyses possibles de la séquence « *de l'* ». Il convient de signaler que la séquence « *du l'* » est non attestée dans MEDITEXT.

Nous avons développé un graphe, figure 87, qui permet d'analyser l'occurrence « *duquel* » en la décomposant en préposition « *de* » et en pronom « *lequel* » et l'occurrence « *dudit* » en la décomposant en préposition « *de* », en déterminant « *le* » et en participe passé « *dit* ».

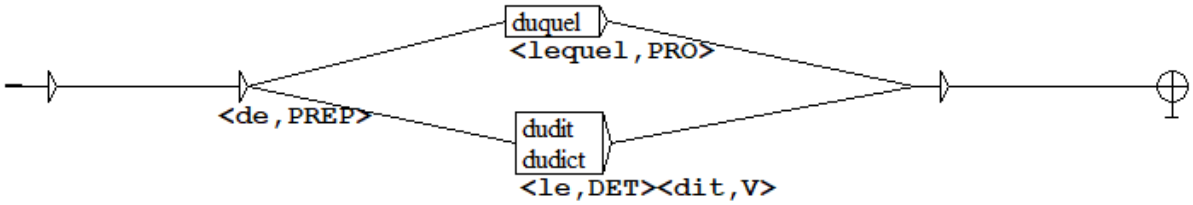


Figure 87. Analyse des contractions des occurrences « *duquel* » et « *dudit* »

3.1.3 Contraction en « *des* »

Comme en français moderne, la forme simple « *des* » peut avoir deux annotations possibles à savoir le déterminant indéfini « *un* » en lui attribuant l'étiquette <*un,DET*> ou la séquence d'annotation composée de la préposition « *de* » <*de,PREP*> suivi du déterminant « *le* » <*le,DET*> .

La forme « *des* » peut être utilisée en contraction avec d'autres formes afin de générer les variantes des occurrences « *desquels* », « *desquelles* », « *desdits* » et « *desdites* ». L'analyse de ces formes contractées est assurée par le graphe de la figure 88 qui permet de décomposer les occurrences « *desquels* » et « *desquelles* » en préposition « *de* » suivie du pronom « *lequel* » et les occurrences « *desdits* » et « *desdites* » en préposition « *de* » suivie du déterminant « *le* ». Il est important de souligner que nous avons recensé plusieurs variantes des occurrences « *desquels* », « *desquelles* », « *desdits* » et « *desdites* » et que la forme « *de* » issue de ces différentes contractions est toujours un pronom.

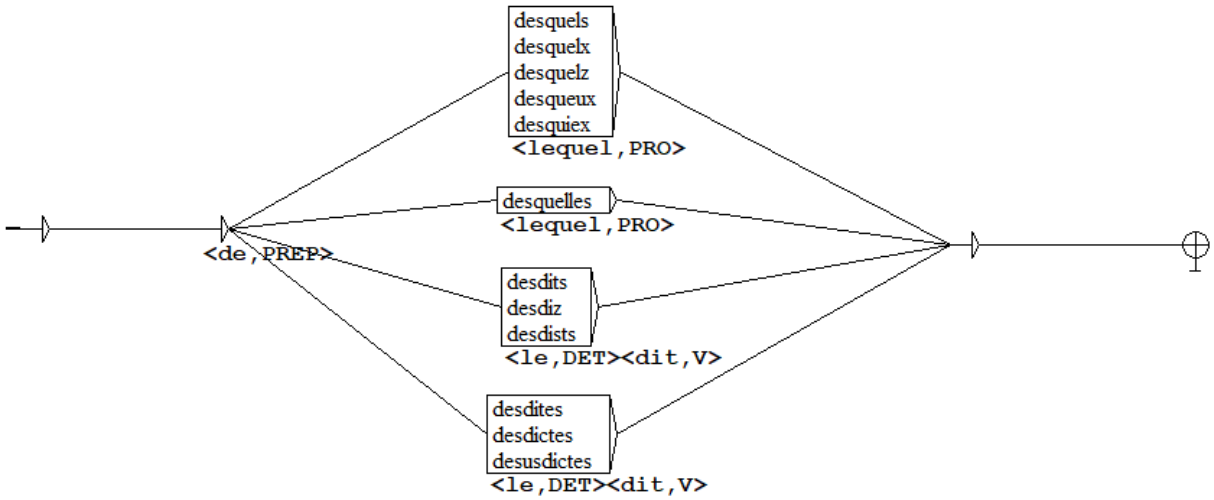


Figure 88. Analyse des contractions des occurrences « *desquels* », « *desquelles* », « *desdits* » et « *desdites* »

3.1.4. Contraction avec « *dit* »

Les analyses, vues précédemment, traitent les différentes variantes des occurrences issues des contractions des formes « *à* », « *de* », « *du* » et « *des* » avec le participe passé « *dit* » à savoir « *audit* », « *auxdits* », « *auxdites* » « *dudit* », « *desquels* », « *desquelles* », « *desdits* » et « *desdites* ». Nous analysons donc, par la suite, seulement un cas particulier de la contraction de participe passé « *dit* » à savoir avec le déterminant « *le* ». En effet, cette contraction donne lieu à quatre occurrences « *ledit* », « *ladite* », « *lesdits* » et « *lesdites* ». De ce fait, nous avons recensé une liste des variantes pour chacune de ces formes. Ces listes, qui ont été utilisées par

notre grammaire présente à la figure 89, ont été décomposées en déterminant « *le* » suivi d'un participe passé « *dit* ».

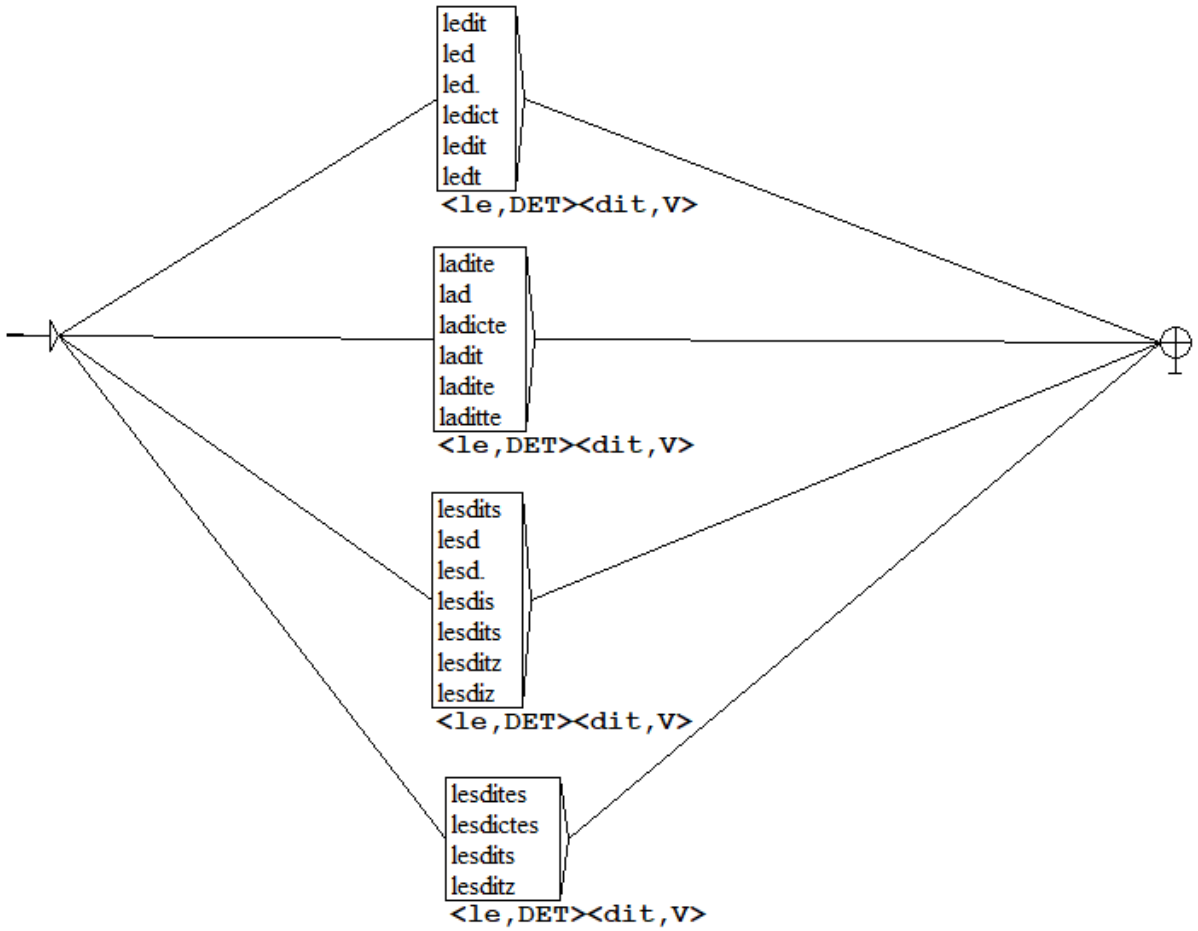


Figure 89. Analyse des contractions des occurrences « *ledit* », « *ladite* », « *lesdits* » et « *lesdites* »

3.2. Analyse des agglutinations

L'agglutination est un procédé qui consiste à réunir en une seule forme orthographique une séquence d'ALU (Silberztein, 2015). Elle se distingue de la composition, qui permet de former une ALU à partir d'une séquence des formes, par l'absence des séparateurs tels que l'espace, le trait d'union et l'apostrophe faisant ainsi une soudure complète des ALU. Conceptuellement, nous considérons que les contractions, vues précédemment au 3.1, sont des agglutinations. La méthode d'analyse à appliquer sur des formes contractées constitue la seule véritable différence perceptible entre la contraction et les autres phénomènes d'agglutination. En effet, la contraction concerne un ensemble fini de formes qui peuvent être recensées et analysées en mettant en place des traitements spécifiques à chaque forme contractée. Tandis que l'agglutination est productive et concerne un nombre non-restreint de formes potentielles et une liste exhaustive de leurs formes est impossible à établir. L'analyse de l'agglutination

suppose une capacité de retrouver les ALU au sein de ces formes agglutinées et de les représenter sous forme d'une séquence d'annotations.

La formation des ALU agglutinés en moyen français se réduite aux phénomènes de préfixation et de suffixation. Cependant, la résolution de l'agglutination fait appel à plusieurs traitements possibles selon le phénomène à analyser. On distingue des phénomènes d'agglutination qui s'appliquent sur un grand nombre d'ALU d'une façon régulière. La *morphologie productive* est une méthode efficace pour traiter ces phénomènes. Elle consiste à mettre en place des mécanismes de préfixation et de suffixation permettant la production de nouvelles entrées lexicales à partir d'une entrée lexicale donnée. Or, l'analyse de ce type d'agglutination doit pouvoir établir une relation entre l'entrée lexicale principale et les entrées lexicales produites en présentant les transformations morphologiques effectuées, les différentes propriétés héritées et partagées et les propriétés spécifiques à chaque entrée lexicale. L'ajout de toutes les entrées lexicales produites et les informations indiquant les relations existantes entre les différentes entrées peut provoquer une croissance exponentielle des entrées du dictionnaire et par conséquent une masse de données difficilement gérable que ce soit pour parser les données textuelles, ou pour mettre en œuvre manuellement un dictionnaire standard de la langue et des grammaires analysant différents phénomènes. A titre d'exemple, en moyen français comme pour le français moderne, l'ajout du préfixe « *re* » au verbe peut dans certains cas exprimer une action répétée comme pour les verbes « *recommander* », « *recommencer* », « *recoucher* » et « *recourir* ». Cette agglutination dite donc « productive » peut être décrite à l'aide des grammaires morphologiques en modélisant les règles de dérivation comme la grammaire exposée à la figure 90 qui permet d'ajouter le préfixe « *re* » aux différentes formes du verbe.

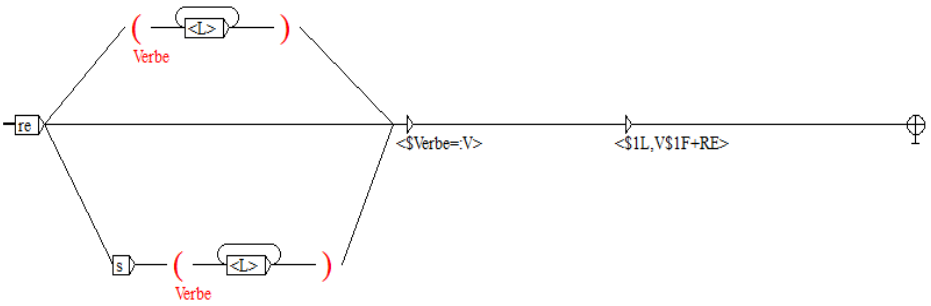


Figure 90. Génération des verbes en utilisant la préfixation « *re* »

Cette grammaire dérivationnelle sera associée aux verbes à l'aide du modificateur *DRV* :

commander,V+FLX=Aimer+DRV=Re

D'autre part, il existe des ALU qui sont des entrées lexicales à part entière du dictionnaire et s'agglutinent avec d'autres ALU. En d'autres termes, ces ALU seront considérées comme des préfixes qui s'ajoutent à un ensemble d'ALU. On distingue deux analyses possibles de ces agglutinations : certaines seront considérées comme des éléments du vocabulaire et seront donc ajoutées au dictionnaire et d'autres seront décomposées en une séquence d'ALU reconnue par une grammaire locale. En effet, certaines agglutinations telles que « *lendemayn*», « *chaussetrappes* », « *passetempz*» et « *chauvesouris*» constituent des formes associées respectivement à des entrées lexicales indépendantes « *lendemain* » « *chausse-trappe*», « *passe-temps*» et « *chauve-souris*». Ces entrées lexicales préfixées ont évoluées et ont pris un sens et des propriétés autonomes au point qu'il est indispensable qu'elles soient décrites comme des entrées lexicales à part entière dans le dictionnaire. Cependant, d'autres préfixes qui sont associés à des classes dites « productives » permettent la constitution des ALU comme « *interlocuteur*», « *monopole*», « *avant-garde*» et « *postposer* ». Ainsi, un préfixe comme « *post* » peut se combiner avec n'importe quel verbe d'action, tandis que le préfixe « *avant* » peut précéder un ensemble important de noms afin d'exprimer généralement une antériorité. D'autres préfixes s'appliquent sur un ensemble plus restreint de noms comme « *mono* » qui sert à indiquer l'unicité du nom préfixé. Pour l'analyse de ces ensembles d'agglutinations, nous procédons à la mise en place de grammaires morphologiques dérivationnelles qui seront associées aux entrées lexicales du dictionnaire.

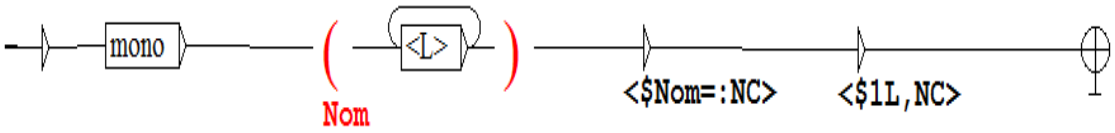


Figure 91. Génération des noms communs en utilisant la préfixation « *mono* »

Au contraire du français moderne, le moyen français se caractérise par la possibilité d'utiliser des agglutinations irrégulières qui sont causées principalement par la non-fixation de l'orthographe, par l'absence de normes de transcription du son phonétique et par le non-respect d'utilisation des séparateurs. Comme illustre le tableau suivant, nous constatons que ces agglutinations ne concernent que quelques préfixes à savoir la préposition « *de* », le déterminant « *le* » et la forme « *que* » lorsqu'ils apparaissent dans certaines structures syntaxiques.

Préfixe	**Structures syntaxiques des agglutinations**	**Occurrences**	**Formes agglutinées**
Préposition « *de* »	De + Nom Propre De + Nom Commun De + Déterminant	*Dangleterre* *donneur* *dune*	*De + Angleterre* *De + Honneur* *De + Un*
Déterminants « *le* »	Déterminant + Nom Commun	*Lesuniversités* *loccasion*	*Le + Université* *Le + Occasion*
Conjonction de subordination « *que* »	Conjonction de Subordination + Pronom	*qil* *quilz*	*Que + Il* *Que + Il*

Tableau 11. Exemples des agglutinations irrégulières

La préposition « *de* » peut être agglutinée avec trois classes grammaticales possibles à savoir un « *Nom Propre* », un « *Nom Commun* » ou un « *Déterminant* ». La grammaire de la figure 92 permet de reconnaitre les agglutinations irrégulières de la préposition « *de* » et de les décomposer en un préfixe suivi d'une forme préfixée. En effet, cette grammaire d'agglutination commence par reconnaitre le préfixe en le sauvegardant dans la variable « *pre* » pour ensuite vérifier si cette dernière correspond à une préposition. Si c'est le cas, le reste de la forme est considérée comme une forme préfixée. Son contenu sera stocké dans la variable « *formePre* » sur laquelle des tests seront effectués pour vérifier qu'il existe dans le dictionnaire une forme ayant une de ces 3 étiquettes <*DET*>, <*NC*> ou <*NP*> correspondant respectivement aux classes grammaticales suivantes : un déterminant, un nom commun et un nom propre. Donc toute forme sur laquelle cette suite déterminée d'opérations séquentielles

est exécutée sera reconnue comme une agglutination de la préposition « *de* ». Cette grammaire d'agglutination produit comme résultat l'annotation *<$1L, PREP>* dans laquelle *$1L* représente le préfixe qui correspond est la préposition « *de* » et l'annotation *<$2L, $2C$2S$2F>* dans laquelle *$2L* correspond à la forme préfixée, *$2C* correspond à la classe grammaticale de cette dernière qui peut être donc un déterminant ou un nom commun ou un nom propre. Finalement *$2S* et *$2F* représentent respectivement les propriétés syntaxiques et morphologiques.

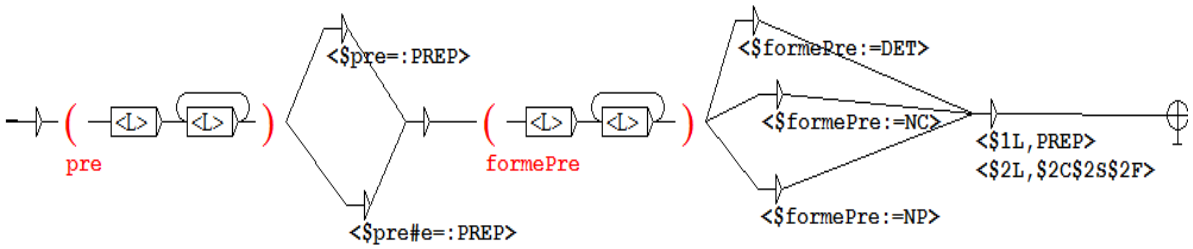


Figure 92. Agglutinations irrégulières de la préposition « *de* »

De même, la figure 93 illustre la grammaire d'agglutination irrégulière où le préfixe correspond au déterminant « *le* » et la forme préfixée correspond à un nom commun. Elle permet donc de déglutiner des formes comme « *loccasion* » et « *lesuniversités* » en produisant deux annotations à savoir l'annotation *<$1L, PREP>* avec *$1L* représente le préfixe plus précisément le déterminant « *le* » et l'annotation *<$2L, $2C$2S$2F>* avec *$2L* correspond à la forme préfixée ayant l'étiquette grammaticale identifiée grâce à l'opérateur $2C qui sera donc un nom commun accompagné de ses propriétés syntaxiques et morphologiques extraites grâce aux opérateurs *$2S* et *$2F*.

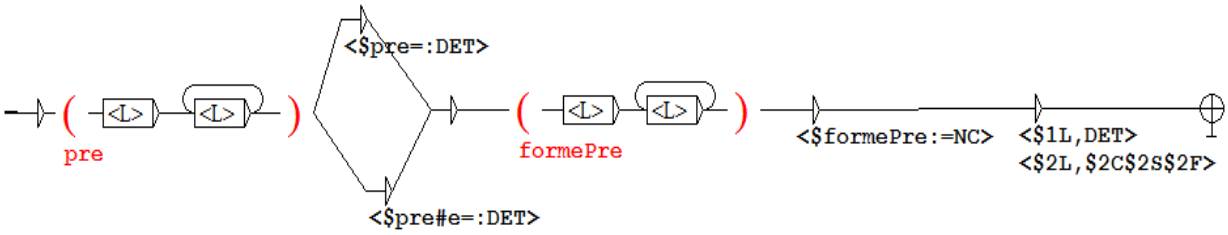


Figure 93. Agglutinations irrégulières du déterminant « *le* »

Les agglutinations irrégulières de la forme « *que* » permet de reconnaitre et de décomposer des formes comme « *qil* » et « *quilz* ». En effet, le préfixe correspondra à une des variantes de la conjonction « *que* » à savoir « *q* », « *qu* » et «*que* » quelle que soit sa classe grammaticale une conjonction de subordination, un pronom ou un adverbe. Tandis que la forme préfixée sera obligatoirement un pronom puisque les variantes des pronoms de

troisième personne du masculin que ce soit singulier ou pluriel sont les seules formes que nous avons recensées pour cette agglutination. Cette grammaire permet la production de l'annotation *<$1L, $1C>* dans laquelle *$1L* représente le préfixe ayant comme classe grammaticale extraite grâce à l'opérateur $2C qui correspond à l'une des trois classes grammaticales précisées lors des tests par les conditions *<$pre :=CONJS>*, *<$pre :=PRO>* et *<$pre :=ADV>* à savoir une conjonction de subordination, un pronom et un adverbe et l'annotation *<$2L, PRO$2S$2F>* dans laquelle *$2L* correspond à la forme préfixée qui sera par un étiquetage forcé associé à l'étiquette du pronom « *PRO* » suivi de ses propriétés syntaxiques et morphologiques qui sont indiquées et capturées par les opérateurs *$2S* et *$2F*.

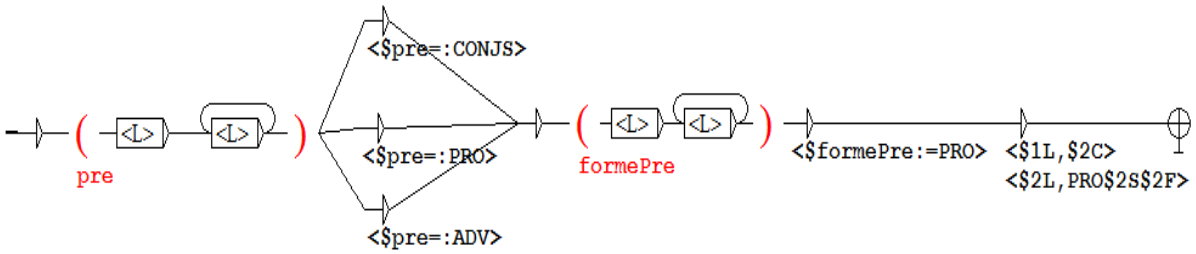


Figure 94. Agglutinations irrégulières de la conjonction « *que* »

3.3. Analyse des désagglutinations

Nous définissons la désagglutination comme étant un découpage de la chaine parlée qui ne respecte pas le dictionnaire et les normes fixés à l'écrit. Elle peut être d'origine phonétique et/ou écrite et elle est considérée donc comme une segmentation erronée d'une forme en une séquence de morphèmes. La résolution de la désagglutination consiste à analyser la séquence des morphèmes issue de la désagglutination soit en considérant que chacun de ces morphèmes correspond à une entrée lexicale à part entière, soit en comptant toute la séquence comme une forme qui doit correspondre à une seule ALU.

Les textes en moyen français se singularisent par la forte utilisation de la désagglutination. Elle est fréquemment présente lorsqu'il s'agit de transcrire une forme qui peut être découpée en deux morphèmes dont chacun d'entre eux peut correspondre à une entrée lexicale du dictionnaire à titre d'exemple « *pour tant* », « *le quel* » et « *le dit* ». Nous avons procédé par l'analyse des désagglutinations fréquemment et régulièrement présentes dans MEDITEXT en dressant une liste des 10 désagglutinations. Comme illustre le tableau 12, pour chacune de ces désagglutinations, nous avons analysé la structure syntaxique des formes déglutinées, l'entrée lexicale associée et sa classe grammaticale correspondante.

Formes déglutinées	**Structure syntaxique**	**Entrée lexicale**	**Classe grammaticale**
le quel	Déterminant+Adjectif	lequel	Déterminant
le dit	Déterminant+participe passé	ledit	Déterminant
toutes voies	Déterminant+Nom Commun	toutevoie	Adverbe
par fois	Préposition+Nom Commun	parfois	Adverbe
lors que	Adverbe+ Conjonction de subordination	lorsque	Conjonction de subordination
par tout	Adverbe +Adverbe	partout	Adverbe
si comme	Adverbe + Adverbe	sicomme	Adverbe
de par	Préposition+Préposition	Depar	Préposition (par)
pour tant	Préposition + Adverbe	pourtant	Adverbe
pour quoi	Préposition +Pronom	pourquoi	Conjonction de subordination

Tableau 12. Analyse des formes déglutinées

Le graphe principal illustré par la figure 95 est restreint à des appels aux sous graphes relatifs aux résolutions de l'ensemble des désagglutinations décrites dans le tableau précédent. Les formes agglutinées analysées ne se réduisent pas à la forme de l'entrée lexicale. En effet, les sous graphes mentionnent l'ensemble des variantes associées à l'entrée lexicale. Par exemple, pour les formes déglutinées « *le quel* », l'ensemble des variantes du déterminant « *le* » ainsi que toutes les variantes de l'adjectif « *quel* » sont reconnues. Par conséquent, le sous graphe « *le quel* » permet d'analyser des désagglutinations comme « *les quels* » et « *li quel* » et « *les quelz* ». Comme sortie, la grammaire attribue à l'ensemble des formes déglutinées l'entrée lexicale correspondante dans le dictionnaire ainsi que sa classe grammaticale.

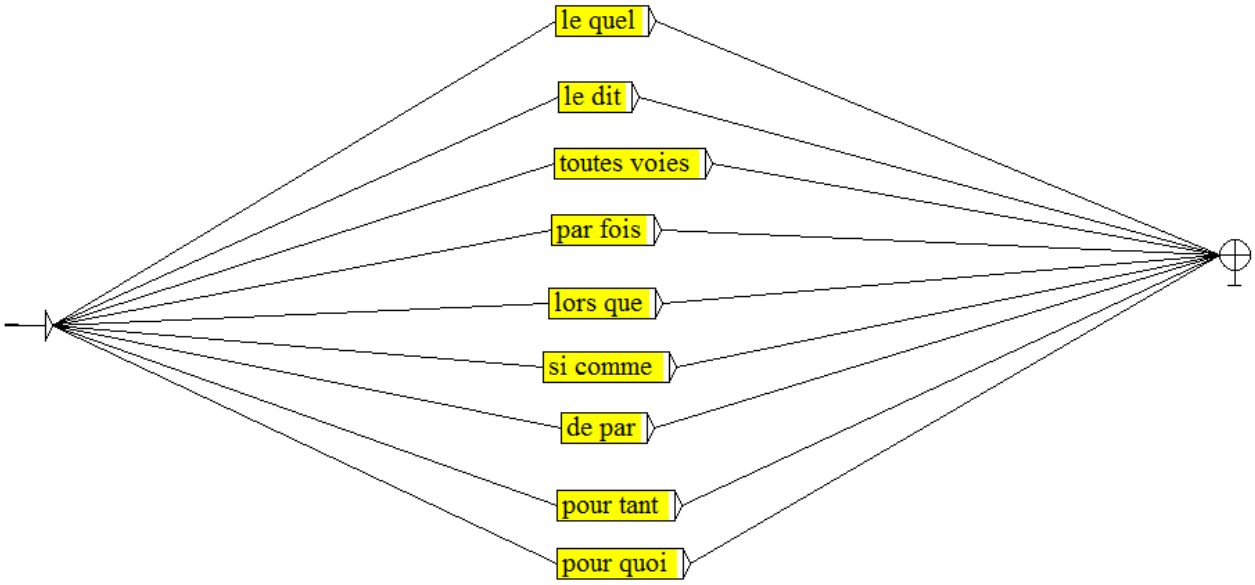


Figure 95. Analyse des désagglutinations

A titre d'exemple, nous détaillons l'analyse de l'ensemble des désagglutination reconnues par le sous graphe « *le quel* ». D'abord, il est important de souligner que même si l'utilisation des formes déglutinées du déterminant « *le quel* » est fréquente par rapport aux autres désagglutinations, la pratique la plus répandue surtout au XVème siècle est d'utiliser l'entrée lexicale du dictionnaire « *lequel* ». En effet, comme montre la figure 96 qui affiche les concordances des variantes de l'ALU « *lequel* » et les concordances des formes déglutinées « *le quel* », l'utilisation des formes déglutinées ne représente que 7.6%. La dominance de la pratique de transcrire le son avec l'ALU « *lequel* » constitue donc un argument qui justifie le recours à la normalisation des formes déglutinées pour qu'elles soient alignées avec la même annotation que les variantes de l'entrée lexicale « lequel ».

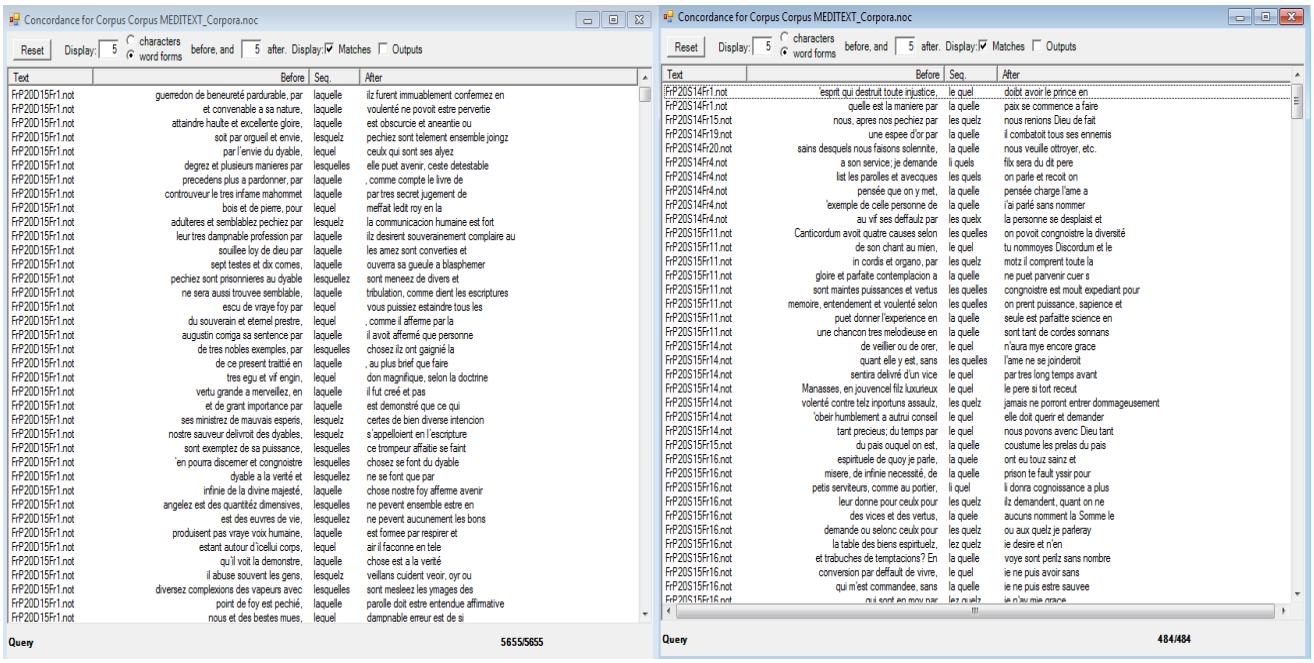


Figure 96. Fréquence d'utilisation de la désagglutination « le quel » et l'ALU « lequel »

La grammaire de la figure 97 reconnait les désagglutinations « *le quel* » qui se composent de séquences des formes constituées d'une variante du déterminant « *le* » <*le,DET*> suivie de l'adjectif « *quel* » <*quel,A*>. Elle annote ces séquences en déterminant « *lequel* » en produisant l'étiquette <*lequel,DET*>.

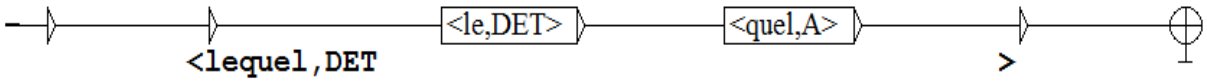


Figure 97. Analyse des formes déglutinées « *le* » et « *quel* »

4. Reconnaissance des mots composés

Un mot composé est une unité du vocabulaire (ALU) qui s'orthographie sous forme d'une séquence de plusieurs formes et d'un ou de plusieurs séparateurs (Silberztein, 2015). La reconnaissance des mots composés est indispensable pour toute analyse lexicale fiable. Elle permet d'identifier des hypothèses lexicales potentiellement présentes qui ne sont pas proposées par la reconnaissance des entrées lexicales simples. En effet, elle consiste à définir les séquences des ALU qui forment une seule unité syntaxique et sémantique et à les distinguer des séquences qui doivent être analysées par des ALU simples.

Trois séparateurs sont principalement utilisés pour orthographier les mots composés en moyen français à savoir l'espace (*arrière fief*), le trait d'union (*beau-fils*) et l'apostrophe (*aujourd'hui).* Or les mots composés n'étant pas standardisés, ils acceptent

plusieurs variantes comportant des séparateurs ou des variantes soudées sans séparateur : *arrière fief, arriere-fiefvez, aujourd'hui, aujourduy, beau-fils, beaufilz.*

4.1. Les mots composés de la langue standard

Comme tous les ALU, les mots composés sont décrits principalement à l'aide d'un dictionnaire électronique. Il ne s'agit pas à ce stade d'analyser les mots composés d'un domaine bien particulier que ce soit scientifique, comme la médecine et l'agronomie, ou technique lié à une activité, comme la cuisine et les services d'ordre juridique. En effet, nous nous intéressons aux mots composés qui font partie du vocabulaire de base tels que les adverbes « *aujourd'hui* » ou « *lonc-temps* » (*longtemps*), les adjectifs « *non-pareil*», les noms communs « *nord-est*» ou « *passe-temps* » et les noms propres « *pont de l'arche* ». Nous constatons que les noms composés sont très nombreux puisqu'ils permettent de spécifier des objets de la vie quotidienne comme *« livre de sapience » « flûte à bec »* et *« porte bois»* et de nommer des institutions, des métiers, des lieux et des personnes à titre d'exemples *«chambre des comptes », « porte bannière », « signy-l'abbaye »* et *« alexandre le grand».*

Rappelons que l'orthographe des mots composés n'est ni stabilisée ni standardisée. Les mots composés acceptent plusieurs variantes de séparateurs ou des variantes soudées sans séparateurs. Pour remédier à ce problème, nous utilisons des caractères spéciaux définis par les dictionnaires électroniques NooJ qui seront interprétés lors de la compilation :

- Le caractère spécial « = » pour représenter la variation entre espace et trait d'union

```
saint=rémi,NP
bois=grolland,NC
bourg=vieux,NP
```

- le caractère spécial « _ » pour représenter sous une forme unifiée les variantes soudées, avec trait d'union et avec espace d'une même ALU.

```
beau_fils,NC
belle_étoile,NC
```

Les mots composés acceptent des variations de leurs constituants. Par exemple, les mots composés « *garde robe », « garde robes », « garde-robe », « garde-robbe », « gardereube », « garderobbe », « garderobe*» sont toutes représentées par une entrée lexicale unique « *garde-robe*». En effet, certaines variations sont productives et peuvent être décrites à l'aide des paradigmes flexionnels et d'autres sont orthographiques et doivent être recensées dans le dictionnaire. Un dictionnaire électronique NooJ offre donc la possibilité d'associer une

description flexionnelle à un mot composé comme pour l'exemple de l'entrée lexicale « *garde_robe,NC+FLX=GardeRobe*» à laquelle on associe donc le paradigme flexionnel « *GardeRobe = <E>/m+s | s/m+p ;*» qui permet la production des six mots composés suivants « *garde-robe* », « *garderobe* », « *garde robe* », « *garde-robes* », « *garderobes* » et « *garde robes* ».

Les variantes orthographiques sont aussi reliées à l'entrée lexicale. En effet, l'acquisition de ces variantes est rendue possible grâce à une démarche exploratoire du corpus à l'aide des méthodes d'analyse textométrique comme « les concordances » en lançant des requêtes décrivant les variantes orthographiques potentielles.

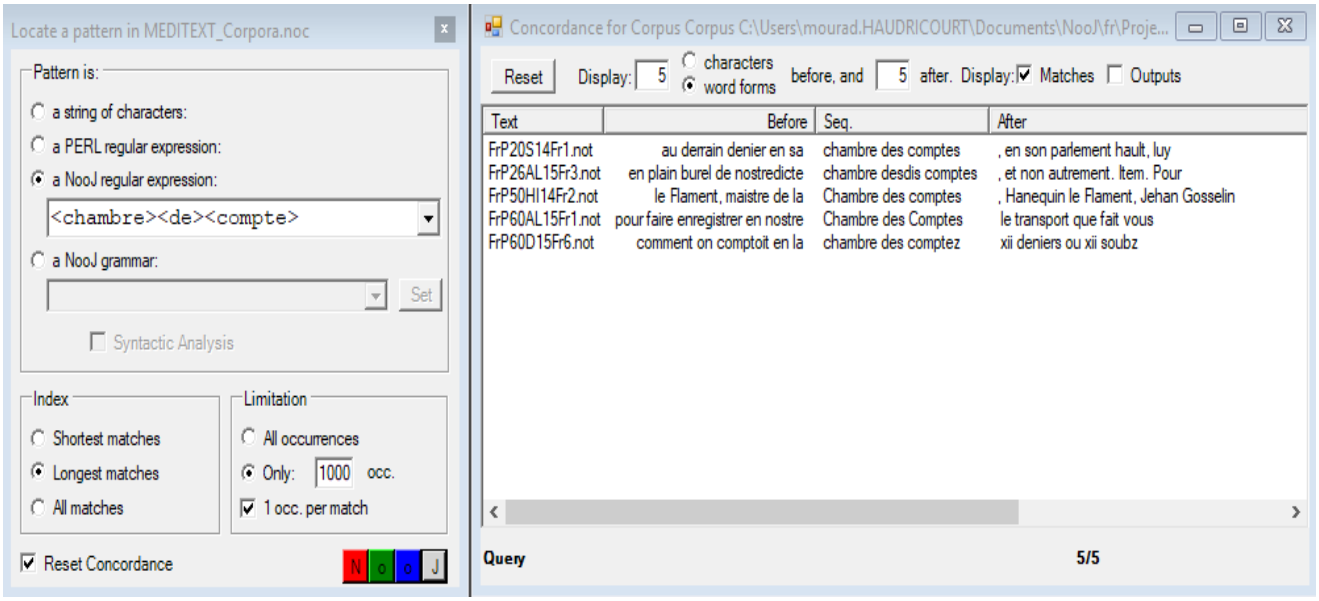


Figure 98. Acquisition de quelques variantes du mot composé « *chambre des comptes* »

Comme l'illustre la figure 98, l'utilisation du concordancier avec les expressions rationnelles NooJ et avec les expressions régulières mises en œuvre à l'aide des opérateurs du langage de programmation PERL a rendue possible l'acquisition de l'ensemble des variantes orthographiques des mots composés. A titre d'exemple, la requête NooJ *<chambre><de><compte>* a permis le recensement de cinq variantes possibles du mot composé « *chambre des comptes* ».

4.2. Analyse des juxtapositions

Le français aux XIVème et XVème siècles se caractérise par l'introduction massive de mots nouveaux. En effet, parmi les principales conséquences de l'usage du français comme la langue de l'administration royale et de l'appareil juridique, de lancement d'une opération de traduction d'œuvres latines et grecques en divers domaines comme l'histoire, la philosophie et

la politique, l'apparition des mots qui décrivent des phénomènes et des faits liés à l'époque. Et, « si le néologisme calqué sur le latin est peut-être le phénomène le plus frappant », d'autres phénomènes ont permis la création des nouveaux mots comme la « juxtaposition » (Marchello-Nizia, 2005).

La juxtaposition est définie par C. Marchello-Nizia (2005) comme un procédé qui consiste à accoler deux constituants qui peuvent être deux substantifs « *chef-lieu* », un substantif et un adjectif « *proces-verbal* », « *saige femme* », un substantif et un infinitif « *culbuter* », un impératif et un substantif « *portefais* », « *tapecul* » et deux substantifs reliés par une préposition « *maistre d'ostel* », « *dé à coudre* », « *sac à vin* ».

En général, les mots composés sont décrits comme une entrée lexicale d'un dictionnaire électronique à laquelle on associe des paradigmes flexionnels et dérivationnels ainsi que des variantes orthographiques. Cependant, dans le cas d'une juxtaposition, il est parfois intéressant de représenter les termes à l'aide d'une grammaire lorsqu'il s'agit de rassembler toutes les variantes terminologiques d'un concept donné.

A titre d'exemple, des historiens font des recherches autour du « *profit du roi* » (Fletcher, 2014), ce concept peut avoir plusieurs occurrences possibles dans les textes.

Grâce aux méthodes d'analyse textométrique à savoir les cooccurrences, les segments répétés et les concordances, nous avons recensé les variantes terminologiques de ce concept comme « *profit du reaume* », « *profit du pueple* », « *profit du roy* », « *profit al roialme* » qui peuvent être reconnues grâce à la grammaire de la figure 99.

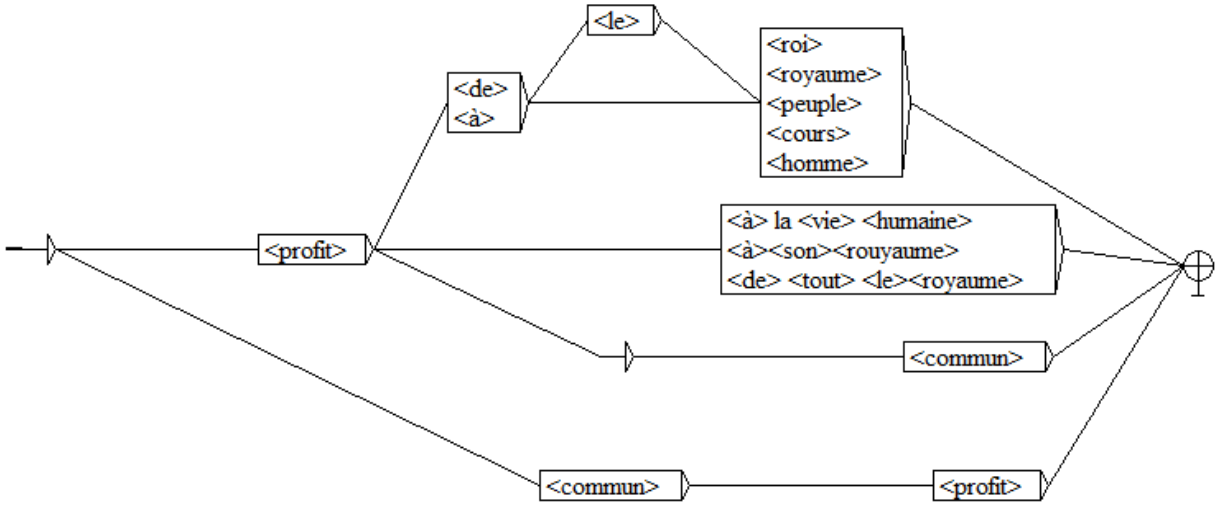


Figure 99. Famille de termes modélisant le concept « profit du roi »

De même, les lois au moyen âge font l'objet de plusieurs études (Genet, 2003), le graphe de la figure 100 permet de modéliser ce concept pour reconnaitre des termes comme « *les lois du païs* », « *les lois du royaume* », « *les lois escrites* », « *les lois profitables du païs* », «*les lois du prince et du pueple*».

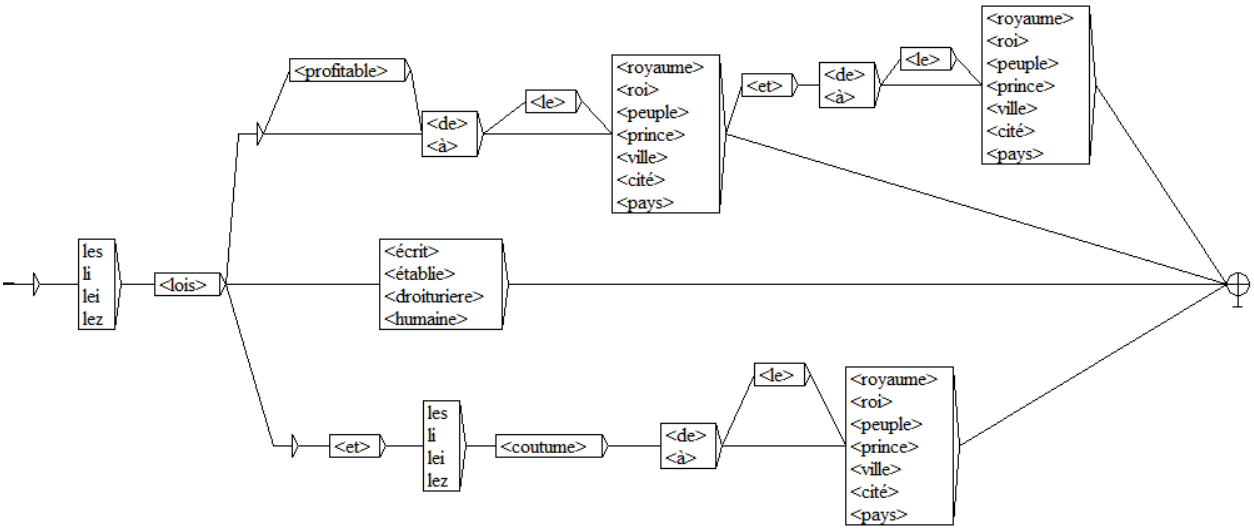


Figure 100. Famille de termes modélisant le concept « *les lois du pays* »

4.3. Analyse des déterminants numériques

L'inscription des nombres en lettres appelés « *noms des nombres* » est une ancienne pratique qui est fort présente dans les textes en moyen français. La plupart des déterminants numériques sont des mots composés comme le montre les exemples suivants : *vingt-quatre, vint sept, vint et deux, quatre vings mille, sexante neufviesme, dix-septisme ;*

Nous distinguons les déterminants numériques cardinaux qui expriment une quantité *« soixante-dix », « trois quart », « six cens », « deux mil»* des déterminants numériques ordinaux qui expriment un ordre « *deux premiers », « vingt sixiesme », « vingt deuxiesme*».

En effet, ces numéraux composés sont constitués d'une séquence de numéraux simples. Par conséquent, leur reconnaissance repose sur une analyse préalable des numéraux simples. Or au contraire du français moderne, ces derniers acceptent plusieurs variantes orthographiques qui devaient être recensées en explorant le corpus. Nous avons donc établi une liste des ALU simples qui correspondent à des nombres cardinaux à savoir *« un », « deux », « trois », « quatre », « cinq », « six », « sept », « huit », « neuf », « dix », « onze », « douze », « treize », « quatorze », « quinze », « seize », « vingt », « trente », « quarante », « cinquante », « soixante », « cent »* et *« mille »*. Ensuite, comme illustre l'exemple de l'ALU

« *vingt* », nous avons associé les variantes orthographiques aux entrées lexicales correspondantes.

vingt,DET+DnumCardinal
vings,vingt,DET+DnumCardinal
vingts,vingt,DET+DnumCardinal
vingtz,vingt,DET+DnumCardinal
vins,vingt,DET+DnumCardinal
vint,vingt,DET+DnumCardinal
vintz,vingt,DET+DnumCardinal

Les déterminants numériques cardinaux sont donc le produit d'une composition à partir des numéraux cardinaux simples. Cette composition peut être formalisée à l'aide de la grammaire locale de la 101. En effet, les numéraux simples forment les nœuds de la grammaire de reconnaissance des déterminants cardinaux avec la possibilité d'insérer un séparateur, plus précisément un trait d'union. Cette grammaire reconnait le numéral simple « *un* » qui sera annoté comme déterminant numérique singulier <*DET+DnumCardinal+s*> et fait appel aux trois sous graphes permettant d'analyser des dizaines de 2 à 99, des centaines de 100 à 999 et des milliers 1000 à 999999. Elle produit l'annotation <*DET+DnumCardinal+p*> qui correspond à un déterminant numérique pluriel.

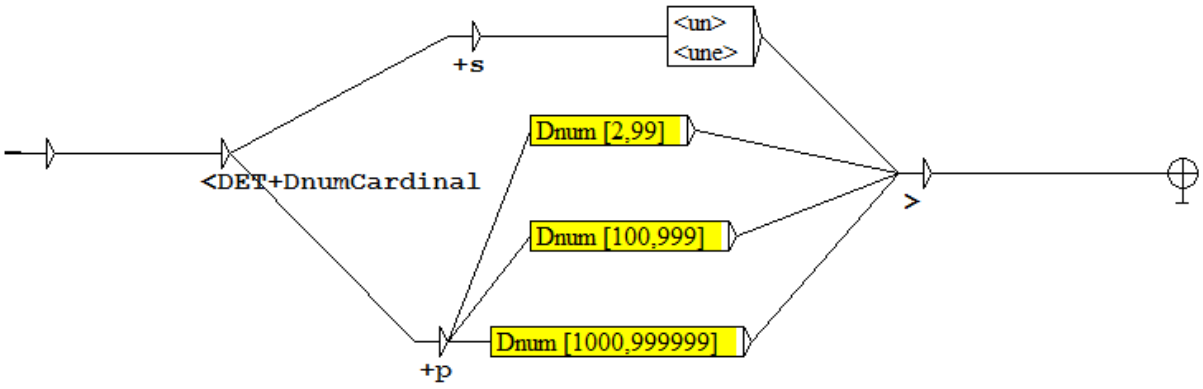


Figure 101. Reconnaissance des nombres cardinaux en lettres

Nous exposons à la figure 102 le premier sous graphe qui correspond à la grammaire de reconnaissance des numéraux cardinaux de 2 à 99. Cette dernière permet l'analyse d'une séquence des numéraux simples qui acceptent deux variantes de séparateurs à savoir l'espace et le trait d'union. Elle permet donc la reconnaissance :

- des déterminants numériques simples de 2 à 16 comme « *unze (11), dousze (12)*» ;
- les valeurs entières des dizaines à savoir de « *vingt* » à « *nonante* » comme « *querente (40), cessante (60) »* ;

- la valeur entière « *80* » qui peut être reconnue grâce à une composition des variantes des nombres « *quatre* » et « *vingt* » comme « *katere vingz (80)*» ;
- les valeurs entières « 60 » et « 80 » qui sont associées aux numéraux du 1 à 9 dans le but de reconnaitre les numéraux de 60 à 69 comme *« soixante chiench (65), cessante treys (63)»* et de 80 à 89 comme *« qater vint cynk (85)»*. Comme elles peuvent être associées aux numéraux de 10 à 19 afin de reconnaitre les numéraux du 70 à 79 comme *« soixante quinse (75)»* et du 90 à 99 comme «*catre vintz quatorze (94)» ;*
- le reste des valeurs entières des dizaines à savoir les variantes des numéraux *« vingt, trente, quarante, cinquante, soixante »* sont relié aux chiffres cardinaux de *un* à *neuf* à titre d'exemple *« vint-deux (22), trente troys (33), carente katere (44)»*.

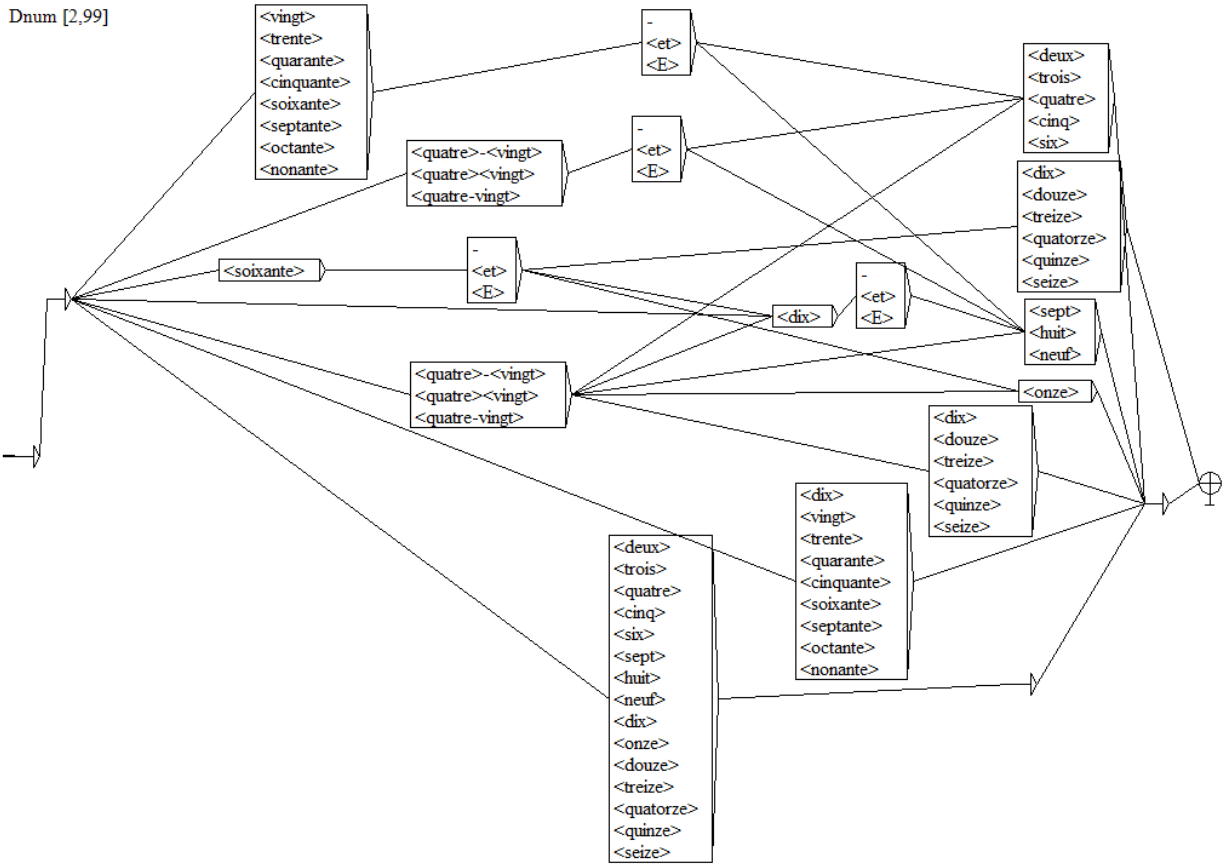


Figure 102. Reconnaissance des nombres cardinaux en lettres

En moyen français, les numéraux ordinaux peuvent avoir plusieurs formes comme « *premier »* ou « *prime* » pour 1er, « *second* » pour 2ème, « *tiers*» pour 3ème et « *quart* » pour 4ème. Ces formes indépendantes sont insérées au dictionnaire électronique comme des entrées lexicales auxquelles on associe les variantes orthographiques correspondantes.

tiers, DET+DnumOrdinal
tierce, tiers, DET+DnumOrdinal

treies, tiers, DET+DnumOrdinal
treis,tiers, DET+DnumOrdinal
tieres,tiers, DET+DnumOrdinal

Cependant, à l'instar du français moderne, la plupart des numéraux ordinaux supérieurs ou égaux à deux se forment à partir du cardinal en ajoutant le suffixe « *ième* ». Or, le suffixe « *ième* » a donné lieu à plusieurs variantes orthographiques à savoir *« ièm », « ème », « èmes », « em », « eme » et « emes »* qui s'attachent aux cardinaux par exemple « *troisiem », « onzieme », « cinquante-neufièm », etc*. La reconnaissance des numéraux ordinaux est rendue possible grâce à la grammaire dérivationnelle de la figure 103 qui permet à partir d'un cardinal de construire le numéral ordinal.

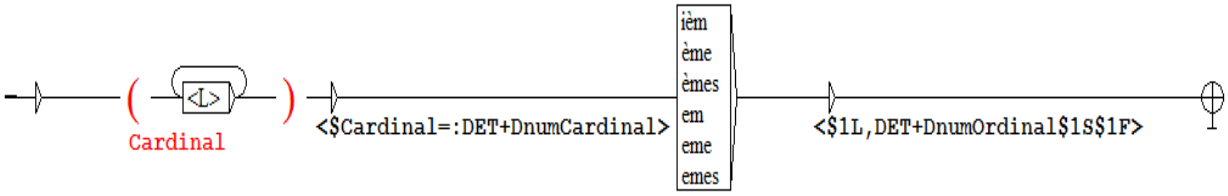


Figure 103. Génération automatique des nombres ordinaux

Cette grammaire morphologique productive s'applique donc aux cardinaux et à leurs variantes dans le but de produire toutes les variantes possibles du déterminant numérique ordinal. Par exemple, elle permet de produire 48 variantes de l'ordinal « *vingtième* » : *vintièm, vintème, vintzème, vintz, vingtzemes , vingtsème, vingtsem.*

5. Conclusion

L'analyse lexicale est la première étape de l'analyse automatique des données textuelles. Notre analyseur lexical des textes en moyen français permet d'identifier et de représenter toutes les unités automatiques d'un texte en utilisant essentiellement le dictionnaire électronique standard de la langue (décrit au chapitre 4). Or, un simple accès au dictionnaire électronique se révèle insuffisant pour identifier les ALU d'un texte en moyen français. En effet, la description de certains phénomènes linguistiques lexicaux est jugée indispensable pour l'identification de toutes les ALU. Afin de s'adapter aux spécificités du MEDITEXT, notre analyseur est décomposé en trois opérations essentielles à savoir les analyses typographiques, la reconnaissance des mots simples et la reconnaissance des mots composés.

Les analyses typographiques permettent de résoudre les problèmes liés à l'utilisation des caractères plus précisément les signes de ponctuations, l'apostrophe et les nombres en chiffres arabes et romains. Ensuite, un ensemble de traitements a été effectué afin d'analyser des phénomènes linguistiques comme la contraction, l'agglutination et la désagglutination pour

permettre une reconnaissance des mots simples. Finalement, nous avons fait appel à des techniques permettant le regroupement des séquences des ALU simples afin de décrire les mots composés de la langue standard. L'analyse lexicale a donc permis la production de toutes les hypothèses lexicales. En d'autres termes, plusieurs descriptions morphosyntaxiques ont été attribuées à une seule ALU sans recourir à une analyse du contexte. En guise de conclusion, les ALU peuvent être ambigus et « ce sont les analyseurs syntaxiques et sémantiques qui résoudront les ambigüités lexicales en tenant comptent du contexte local, syntaxique ou sémantique des ALU » (Silberztein, 2015).

Chapitre 6

L’étiquetage morphosyntaxique de textes en moyen français

1. Introduction

L’étiquetage morphosyntaxique est « une tâche de la syntaxe locale » (Silberztein, 2015) considérée par Paroubek & Rajman (2000) comme essentielle pour tout traitement ultérieur comme l'analyse syntaxique structurelle, sémantique ou pragmatique d'une langue. Elle permet une annotation grammaticale d’un texte en attribuant à chacune de ses formes la description morphosyntaxique correspondante. En effet, à partir d’un ensemble de couples (forme, étiquette morphosyntaxique), cette opération consiste selon Fleury (2009) à choisir pour chacune des formes du texte parmi les étiquettes morphosyntaxiques associées celle(s) qui correspond(ent) au contexte. La réalisation de cette tâche ne nécessite pas une analyse des phrases ou des énoncés mais plutôt la reconnaissance des constructions de petite taille ou des séquences plus conséquentes. La description de ces ensembles de séquences d’ALU est mise en place en faisant appel soit à une méthode par apprentissage, soit à une méthode symbolique, permettant ainsi la désambiguïsation des ALU en attribuant à chaque forme la description morphosyntaxique lui correspondant selon le contexte.

A notre connaissance, jusqu’à présent le peu de travaux concernant l’étiquetage morphosyntaxique et traitant de textes en moyen français est basé sur des méthodes par apprentissage. Citons à titre d’exemple l’étiqueteur probabiliste mis en œuvre à l’aide de *Treetagger* par Heiden & Prévost (2002) pour le Français Médiéval du XIème au XVème siècle. Nous présentons ensuite une méthode symbolique d’étiquetage morphosyntaxique qui prend en compte la spécificité du moyen français à savoir la non-standardisation de l’orthographe et l’évolution de la syntaxe remarquablement influencée par les aspects chronologique et géographique.

2. Processus d'étiquetage morphosyntaxique

Le processus d'étiquetage morphosyntaxique est composé, rappelons-le, essentiellement de trois phases (i) le prétraitement et la segmentation, (ii) l'étiquetage a priori et (iii) la désambiguïsation (Paroubek & Rajman, 2000).

Comme le montre la figure 104, notre méthode symbolique d'étiquetage morphosyntaxique repose sur une analyse lexicale fine (que nous avons décrite au chapitre 5) qui permet d'identifier les ALU d'un texte et d'y associer toutes les descriptions morphosyntaxiques correspondantes. Par conséquent, suite à l'analyse lexicale, toutes les segmentations possibles d'une séquence de caractères sont proposées et sont enrichies par les hypothèses morphosyntaxiques. Pour désambiguïser les ALU et faire, selon le contexte, un choix automatique d'une des étiquettes morphosyntaxiques associées à une forme, nous utilisons des grammaires locales appelées « grammaires de levée d'ambiguïté ». Ces grammaires en effet décrivent les contextes d'apparitions d'une ALU qui permettent de la désambiguïser dans de nombreuses occurrences. Rappelons que plusieurs études ont montré qu'une grosse part de l'ambiguïté est généralement détenue par un petit nombre des formes qui sont les mots grammaticaux les plus fréquents (chapitre 2, 3.3 page 60). Par conséquent, des grammaires qui décrivent les contextes immédiats de ces mots peuvent réduire considérablement la taille de la TAS. Il ne s'agit pas d'une description de toutes les phrases dans lesquelles la forme à désambiguïser apparait mais d'une description d'un voisinage d'une à quatre ALU qui représente les contextes caractéristiques de la forme en question. Ces contextes peuvent être déterminés par une analyse textométrique qui permet ainsi de prendre en compte les variantes orthographiques les plus fréquentes et les structures syntaxiques dominantes.

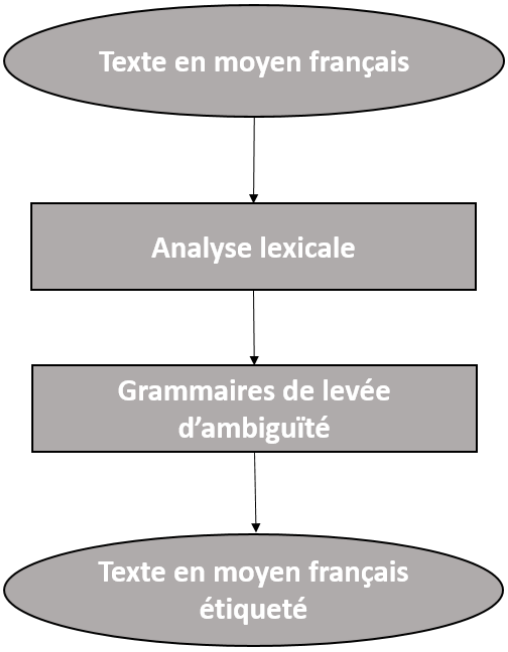
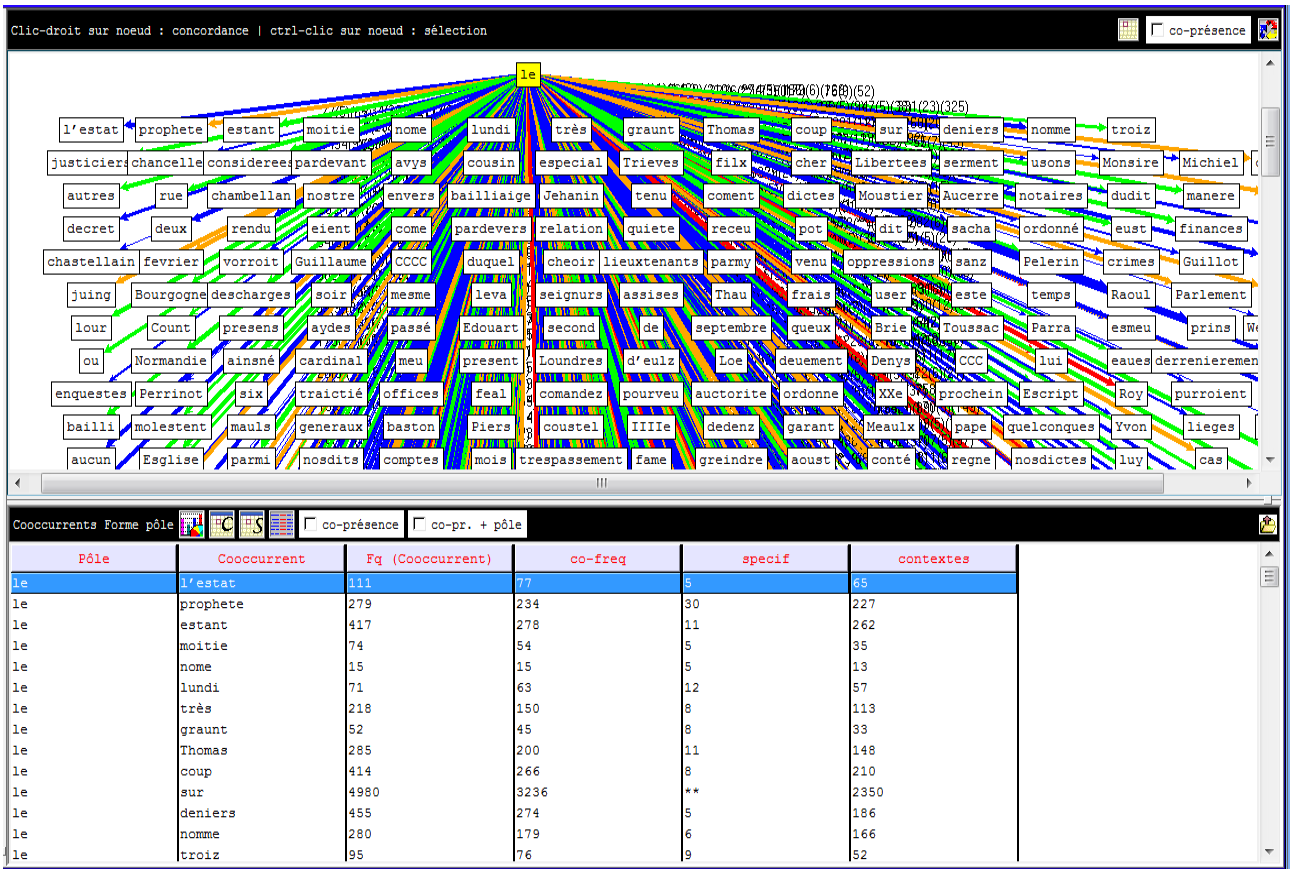


Figure 104. Processus d'étiquetage des textes en moyen français

La distribution des probabilités d'apparition des formes dans un corpus d'une langue non-standardisée, plus particulièrement le moyen français, est plus complexe que celle d'une langue comme le français moderne ou l'anglais qui contiennent un nombre fini de formes fréquentes et ambigües. En effet, en moyen français, il peut y avoir plusieurs variantes orthographiques pour une seule forme et la syntaxe, qui de fait n'est pas stable, évolue. Cependant, l'analyse textométrique rend possible de recenser les formes ambigües les plus fréquentes ainsi que leurs variantes et de détecter les structures syntaxiques minimales dans lesquelles elles apparaissent régulièrement. Nous présentons ces méthodes textométriques qui permettent une description des contextes immédiats des mots grammaticaux fréquents à savoir les cooccurrences, les segments répétés et les concordances.

La cooccurrence permet de déterminer les formes utilisées d'une façon régulière dans un contexte commun. Elle fait en effet, comme le soulignent Manning & Schütze (1999), référence au phénomène général par lequel des mots sont susceptibles d'être utilisés dans un même contexte. Elle indique donc si deux formes sont associées l'une à l'autre dans un contexte commun. Salem (1986) rappelle que des couples de formes présentes dans les mêmes phrases à des distances variables et dans des contextes immédiats différents sont donc identifiés. A titre d'exemple, la figure 105 montre les cooccurrents de la forme « *le* » dont nous constatons la présence avec un nombre important des noms comme « *lundi* », « *cousin* » ou les variantes de « *seigneur* », des verbes comme les variantes de « *ordonner* », « *faire* » et

« *avoir* », quelques adjectifs comme « *grant* », « *beau* » ou « *bon* » et des adverbes comme « *très* », « *autrement* » et « *coment* ».

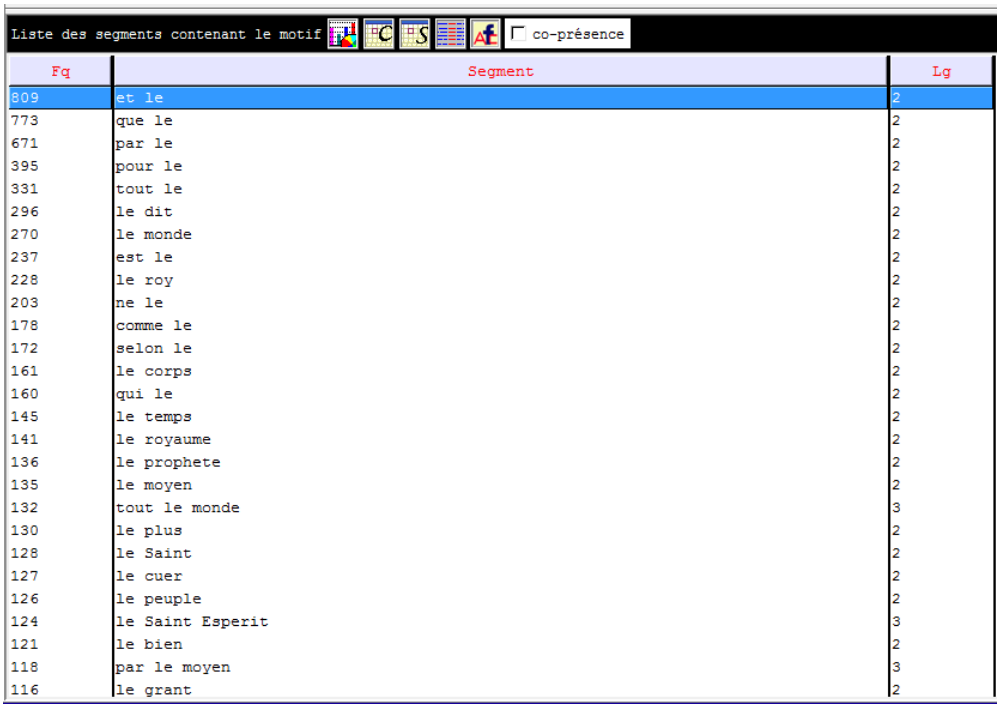


Pôle	Cooccurrent	Fq (Cooccurrent)	co-freq	specif	contextes
le	l'estat	111	77	5	65
le	prophete	279	234	30	227
le	estant	417	278	11	262
le	moitie	74	54	5	35
le	nome	15	15	5	13
le	lundi	71	63	12	57
le	très	218	150	8	113
le	graunt	52	45	8	33
le	Thomas	285	200	11	148
le	coup	414	266	8	210
le	sur	4980	3236	**	2350
le	deniers	455	274	5	186
le	nomme	280	179	6	166
le	troiz	95	76	9	52

Figure 105. Les cooccurrences de la forme « *le* »

L'opération des segments répétés permet la reconnaissance des séquences d'ALU qui apparaissent dans le même ordre et de façon fortement récurrente dans le corpus.

Comme le montre la figure 106, les segments répétés permettent de recenser des séquences d'ALU contenant la forme « *le* » et ayant une distribution régulière dans le corpus. Ces segments répétés permettent donc d'identifier les contextes immédiats les plus fréquents de la forme « *le* ». L'analyse des voisinages récurrents permet d'appliquer aux segments répétés les méthodes statistiques utilisées pour analyser la distribution des formes simples (Salem, 1986). A partir des segments résultants, nous pouvons donc lancer une analyse des concordances.



Liste des segments contenant le motif ☐ co-présence

Fq	Segment	Lg
809	et le	2
773	que le	2
671	par le	2
395	pour le	2
331	tout le	2
296	le dit	2
270	le monde	2
237	est le	2
228	le roy	2
203	ne le	2
178	comme le	2
172	selon le	2
161	le corps	2
160	qui le	2
145	le temps	2
141	le royaume	2
136	le prophete	2
135	le moyen	2
132	tout le monde	3
130	le plus	2
128	le Saint	2
127	le cuer	2
126	le peuple	2
124	le Saint Esperit	3
121	le bien	2
118	par le moyen	3
116	le grant	2

Figure 106. Les segments répétés les plus fréquentes contenant la forme « *le* »

Les concordances permettent une description détaillée des contextes. Comme l'illustre la figure 107 des concordances du segment répété « *que le* ». Cette méthode permet de décrire les différents contextes immédiats de la séquence « *que le* » dans le but de désambiguïser la forme « *le* » en distinguant les cas selon qu'elle est déterminant ou pronom.

n°	Partie	ContexteGauche	Pôle	ContexteDroit
1	titre=FrP20D15Fr1.xml	poitier, car il semble a chascun d'eulx	que	le gaignage de son voisin amendrit le
2	titre=FrP20D15Fr1.xml	pechiez sont telement ensemble joingz et connexes	que	le mouvement d'orgueil est inseparablement tousjours acc
3	titre=FrP20D15Fr1.xml	seule illusion et oppression de fantasie et	que	le dyable par son malice abuse ainsi
4	titre=FrP20D15Fr1.xml	dont elles venoient. Si cuiderent les aulcuns	que	le feu gouvernast tout le monde et
5	titre=FrP20D15Fr1.xml	superflue qu'ilz avoient a leurs amys. Ainsi	que	le sage conte que ung pere,
6	titre=FrP20D15Fr1.xml	le pechié de vauderie est plus grief	que	le pechié d'ydolatrie. Mais ces traitres
7	titre=FrP20D15Fr1.xml	tres execrable profession. Il appert doncques assez	que	le crisme des vaudois est sans comparoison
8	titre=FrP20D15Fr1.xml	des vaudois est sans comparoison plus grief	que	le pechié de ydolatrie des payens.
9	titre=FrP20D15Fr1.xml	le pechié de vauderie est plus grief	que	le pechié de heresie. 6 Or
10	titre=FrP20D15Fr1.xml	le pechié de vauderie est plus grief	que	le pechié des machommetistes. Et aprés
11	titre=FrP20D15Fr1.xml	humain. Mais toutesfois je oze bien affermer	que	le contagieux et pestilencieux venin des vaudois
12	titre=FrP20D15Fr1.xml	et de tant excede cestui crisme l'autre	que	le sacrement de l'autel ouquel est vrayement
13	titre=FrP20D15Fr1.xml	sa proye ordinaire, et si a fiance	que	le fleuve de jourdan doive descendre et
14	titre=FrP20D15Fr1.xml	scismatique que aucuns pareillement jugeront plus doulz	que	le pain acoustumé de saincte obeissance.
15	titre=FrP20D15Fr1.xml	elle a commencié. Il est a croire	que	le tres perilleux temps vendra duquel daniel
16	titre=FrP20D15Fr1.xml	tirans et dirent tout au long ce	que	le sainct esperit leur enseignoit et ne
17	titre=FrP20D15Fr1.xml	prenes le tres fin et reluisant harnas	que	le sainct esperit a forgié pour entre
18	titre=FrP20D15Fr1.xml	malefices sans quelque doubte et crainte, et	que	le fol delaisse souvent sa folye et
19	titre=FrP20D15Fr1.xml	bien animer qu'ilz s'acquitent de la charge	que	le hault juge leur a baillié,
20	titre=FrP20D15Fr1.xml	la char pourrie et infecté, affin d'eschever	que	le surplus de la masse du corps
21	titre=FrP20D15Fr1.xml	trois enfans nommez ananie, azaie et misael	que	le tirant nabugodonosor fist jecter en la
22	titre=FrP20D15Fr1.xml	Au surplus, dont est ce peu venir	que	le tres victorieux charlemaine, roy des
23	titre=FrP20D15Fr1.xml	car en tous ces enseignemens elle suppose	que	le dyable puist estre forcié et constraint
24	titre=FrP20D15Fr1.xml	comme je croy, personne qui face doubte	que	le bon ou mauvais angele au commandement
25	titre=FrP20D15Fr1.xml	angelez? La fondacion des consideracions de ce	que	le dyable puet faire a la verité
26	titre=FrP20D15Fr1.xml	de la nature haulte, et pour ce	que	le mouvement de lieu en lieu est
27	titre=FrP20D15Fr1.xml	pourtant les aultres y soient. Puis doncques	que	le mouvement local est le premier et
28	titre=FrP20D15Fr1.xml	est cellui qui ne sache tres bien	que	le createur a conferé aux corporelz elemens
29	titre=FrP20D15Fr1.xml	valoir et servir certaines applications d'aulcunez semences	que	le dyable congnoist tres bien. Par
30	titre=FrP20D15Fr1.xml	preserve. Le tiers enseignement parle des chosez	que	le dyable fait illusioirement et par seule
31	titre=FrP20D15Fr1.xml	sans quelque verité. Ja soit ce doncques	que	le dyable puist faire moult de chosez
32	titre=FrP20D15Fr1.xml	face de certaine et determinee matiere et	que	le semblable soit engendré de son semblable
33	titre=FrP20D15Fr1.xml	fantosme sans verite, ou que la chose	que	le voit soit ung corps forme de

Figure 107. Concordances de la forme « *le* »

3. Grammaires de levée d'ambiguïté

Les grammaires de levée d'ambigüité appelées aussi « grammaires locales de désambiguïsation » reposent sur une production précise des contraintes lexicales selon le contexte décrit. Par exemple, pour le contexte exprimé par l'expression rationnelle « *il le <V>* », la contrainte lexicale « la forme *le* est un pronom » est produite. Les grammaires locales sont utilisées dans la plupart des cas pour ajouter des annotations, dans un contexte de désambiguïsation automatique, elles produisent les contraintes lexicales sous forme de filtres permettant ainsi de supprimer de la TAS toutes les annotations incompatibles. Par exemple, le filtre *<PRO>*, produit par la forme « *le* » dans notre exemple, sera utilisé pour supprimer toutes les annotations de la TAS qui ne sont pas compatible avec l'étiquette « *PRO*». Par la suite, nous présentons les grammaires de désambiguïsation développées pour des mots grammaticaux en moyen français les plus fréquents dans MEDITEXT permettant également de lever l'ambigüité des formes présentes dans leurs contextes immédiats.

3.1. Désambigüisation des déterminants/pronoms « le », « la », « li », « l' » et « les »

Les formes « *le* », « *la* », « *l'* », « *li* » et « *les* » peuvent être, dans la plupart des cas, soit des articles définis soit des pronoms. Au contraire du français moderne, nous constatons, dans bon nombre de cas, l'absence du déterminant devant le substantif, cependant, l'utilisation de l'article défini reste fréquente. La grammaire de la figure 108 décrit en détails les contextes de chacune de ces formes dans le but de les désambiguïser. En effet, elle fait appel à des sous-graphes qui permettent la désambiguïsation non pas seulement de la forme en question mais aussi de plusieurs formes appartenant à ses contextes immédiats. Nous notons qu'en cas d'agglutination avec un nom ou en contraction avec les prépositions « *de* », « *a* » et « *en* », les formes « *le* », « *la* », « *l'* », « *li* » et « *les* » peuvent être désambigüisées par notre grammaire puisque ces formes ont été déglutinées suite à l'analyse lexicale.

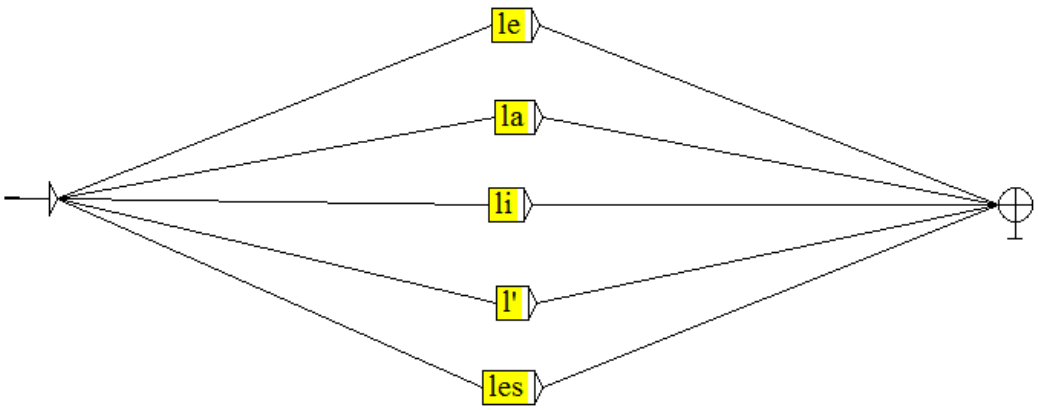


Figure 108. Grammaire de désambiguïsation des formes « le », « la », « l' » et « les »

La forme « *le* » peut donc être soit un déterminant, soit un pronom, soit une variante du nom commun « *lé* » qui désigne « *largeur* ». Comme l'illustre le sous-graphe de la figure 109, notre grammaire de désambiguïsation présente une description de deux ensembles de contextes : le premier produit le filtre « *DET* » permettant ainsi d'annoter « *le* » en déterminant et de supprimer de la TAS les annotations « *PRO* » et « *NC* » et le deuxième ensemble spécifie les situations où « *le* » est utilisé comme pronom en ne gardant dans la TAS que l'annotation « *PRO* ».

La forme « *le* » apparait comme déterminant lorsqu'elle précède un nombre cardinal ou ordinal en chiffres romains, en chiffres arabes ou en lettres comme le montrent les exemples suivants « *le quinziesme* », « *le onziesme* », « *le XXVIII* », « *le XXXVI* » et « *le 8e* ».

En moyen français, « *le* » est fréquemment utilisée comme déterminant d'un animé ou d'un non animé décrit à l'aide d'un des deux adverbes « *plus* » ou « *très* » suivi d'un adjectif. A titre d'exemple, nous citons « *le plus fort* », « *le plus petit* » et « *le très sage* ». La forme « *le* » peut également être déterminant d'un substantif. C'est le contexte le plus fréquent de l'emploi global de la forme « *le* » avec environ 74% des cas. Il est décrit par la forme « le » suivie d'un nom commun ou d'un nom propre comme « « *le monde* », « *le texte* » et « *le feu* ». Le nom peut être précédé d'un adjectif afin de reconnaitre les séquences d'ALU comme « *le bon message* » ou d'un adjectif précédé d'un adverbe comme « *le plus parfait mouvement*» et *« le tres vaillant roy*». En effet, dans notre grammaire, ces contextes sont reconnus par la simple expression rationnelle « *le (<E>|<ADV>)(<A>)(<NC>|<NP>)| le (<NC>|<NP>)*». En examinant, les séquences reconnues par cette expression, nous constatons qu'elles sont souvent précédées par un ensemble d'ALU qui peuvent être désambigüisées à savoir:

- les conjonctions de subordination « *que* », « *comme* » et « *quand* » permettant la reconnaissance des séquences d'ALU comme « *que le sage conte* », « *que le bon ange* », « *comme le plus digne seigneur* », « *comme le bon filx* » et « *quand le nouveau duc* » ;
- l'adjectif « *tout* » afin de reconnaître des séquences comme « *tout le gros navire* », « *tout le monde* » et « *tout le tamps* » ;
- les prépositions « *par*», « *pour* » et ses variantes « *por* » et « *pur* », « *de* », « *du* », « *sur* » et « *selon* » pour former des séquences d'ALU comme « *par le abbé* », « *de le roialme* », « *selon le proverbe* », « *sur le bon governement* » , « *pour le jeune roy* » et « *pour le trés grant bien* » ;

- les conjonctions de coordination « *et* » et « *ne* » variante de « *ni* » afin de reconnaître des séquences d'ALU comme « *et le plus parfait mouvement* », « *et le plus grant maistre* », « *ne le Fils* » et « *ne le Saint* » ;
- l'adverbe « *comment* » pour former des séquences d'ALU comme « *comment le souverain Empereur* », « *comment le saige roy* » et « *comment le saint prophete* » ;
- le verbe être au troisième personne du singulier « *est* » précéder ou non par le pronom « *c'* » permettant la reconnaissance des séquences d'ALU comme « *est le plus grant maistre* », « *c'est le tres crestien royaulme* » et « *c'est le Saint Esperit*».

Notre grammaire modélise aussi les contextes d'apparition de « *le* » pronom. Dans ce cas, il est forcément suivi d'un verbe. Ce verbe peut à son tour être précédé d'un pronom « *lui* » ou « *leur* » mais cette utilisation, au contraire du français moderne, reste très rare. Les séquences reconnues par l'expression rationnelle composée de la forme « *le* » suivi ou non d'un des pronoms « *lui* » et « *leur* » suivi du verbe « *le (lui|leur)<V>* » peut être précédé par ces ALU :

- les pronoms personnels « *je* », « *tu* », « *il* », « *elle* », « *on* », « *nous* », « *vous* », « *ils* » et « *elles* » afin d'identifier des séquences d'ALU comme « *nous le trouverons* », « *je le demande* », « *il le croit* », « *nous le verrons* » et « *vous le mettez*» ;
- l'adverbe « *comment* » permettant la reconnaissance des séquences d'ALU assez rares comme « *comment le tournez* » et « *comment le sera* » ;
- les conjonctions de subordination « *comme* » et « *qui* » pour former des séquences d'ALU comme « *comme le pria* », « *comme le declaire* », « *comme le demonstre* », « *qui le devaient* » et « *qui le bat* » ;
- l'adverbe de négation « *ne* » pour former des séquences d'ALU « *ne le rappelle* », « *ne le trouveroit* » et « *ne le doit* » ;
- les prépositions « *de* », « *à* » et sa variante « *a* » et « *pour* » ainsi que ses variantes « *por* » et « *pur* » permettant la reconnaissance des séquences d'ALU « *pour le distribuer* », « *pour le payer* », « *de le trahir* », « *de le delivrer* » et « *de le mettre*».

-

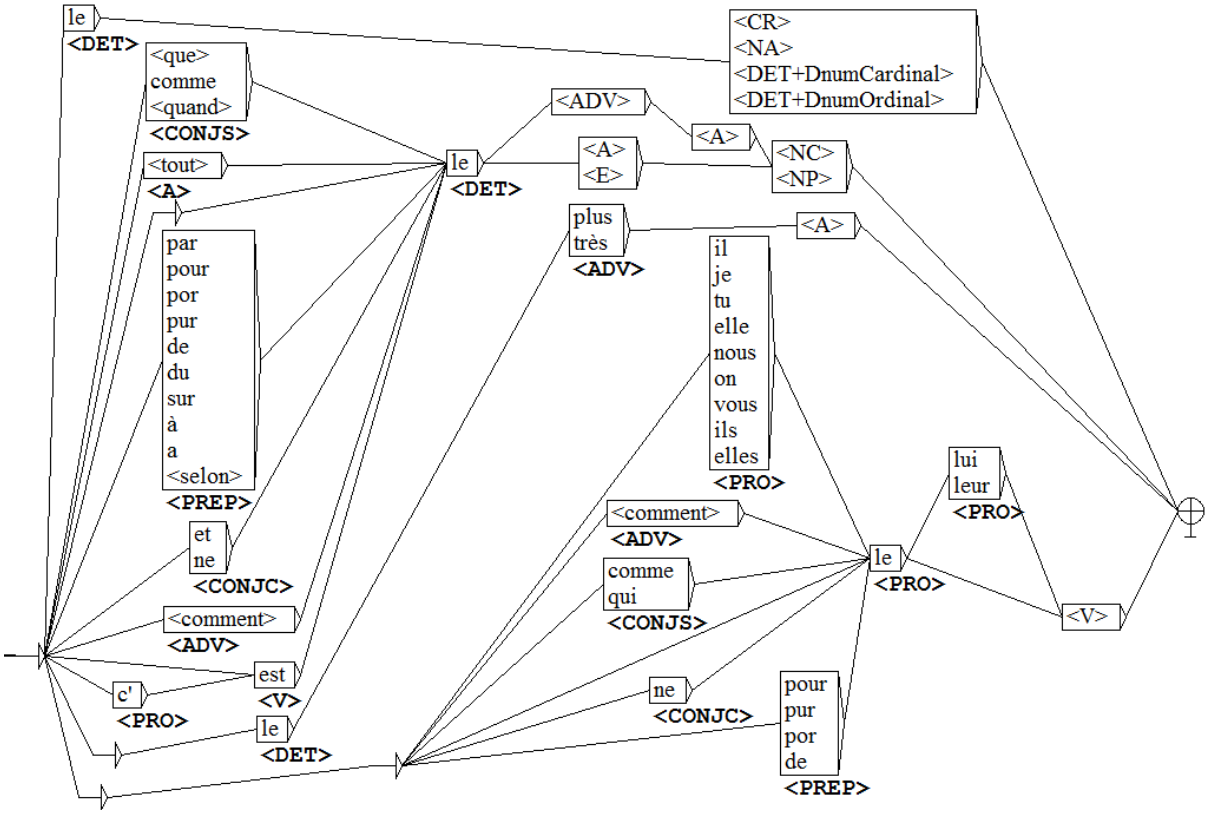


Figure 109. Grammaire de désambiguïsation de la forme « *le* »

En moyen français, la forme « *la* » est fréquemment utilisée. Tout comme la forme « *le* », elle peut être un déterminant précédant un nombre cardinal ou ordinal en chiffres arabes, en chiffres romains ou en lettres. On la rencontre souvent comme déterminant devant un substantif, par exemple, « *la couronne* », « *la lumiere* », « *la dignité* », « *la vie* » et « *la sagesse* ». Dans bon nombre de cas, le nom peut être accompagné d'un adjectif précédé ou non d'un adverbe. La description de ce nombre restreint de contexte via une simple règle rationnelle comme « *la (<E>|<ADV>)(<A>)(<NC>|<NP>)| la (<NC>|<NP>)*» permet de recenser environ 77% d'utilisation de la forme « *la* ». A titre d'exemples des séquences d'ALU reconnues par cette expression, nous citons « *la couronne* », « *la congnoissance* », « *la divine lumiere* », « *la noble nature* » et « *la tres saincte foy* ». Comme l'illustre notre grammaire de la figure 110, ces contextes ont fait l'objet d'une analyse textométrique permettant d'identifier les ALU qui apparaissent fréquemment avec les séquences reconnues par notre expression rationnelle. Ces ALU ont donc permis une description plus détailler et plus fiable de certains contextes immédiats du déterminant « *la* » et par conséquent la désambiguïsation de plusieurs ALU. Plus précisément, le déterminant « *la* » est souvent précédé par :

- les prépositions « *de* », « en », « *par* », « *selon* », « *sur* » , « *à* » et ses variantes et « *pour* » et ses variantes « *pur* » et « *por* » permettant la reconnaissances des séquences d'ALU comme « ***en la cité*** *de paradis* », « ***en la bataille*** *espirituelle* », « *de la mort* », « *de la passion* », « *à la punition* », « *par la gorge* », « par la rue », « ***pour la paix*** » et « sur la terre et la fait a son » ;

- la conjonction de coordination « *et* » pour former des séquences d'ALU comme « *et la plus dure servitude* » et « ***et la gloire*** *n'en aiés.*» ;

- la conjonction de subordination « *que* » afin d'identifier des séquences d'ALU comme « ***que la multiplicacion*** *et infinité des maulx d'iceulx* » et « *que la nature* » ;

- l'adjectif « *toute* » permettant la reconnaissance des séquences comme « ***toute la nuit****, c'est bien assavoir* », « *toute la matiere de nostre euuangile* » et « ***toute la journee****, et en a plus* ».

- l'adverbe « *comment* » pour reconnaitre des séquences comme « *comment la pape*», « *comment la grant vision* » et « *comment la riche marchandie*» ;

- le verbe être au troisième personne de singulier « *est* » précéder ou non par le pronom « *c'* » afin d'identifier des séquences d'ALU comme « *est la parolle* », « *est la Vierge Marie* », « *est la plus petite lettre* », « *c'est la vraye estoile* » et « *c'est la glorieuse Vierge* ».

La forme « *la* » peut également être pronom lorsqu'elle apparait devant un verbe précédé ou non d'un des pronoms « *lui* » ou « *leur* » par exemple « *la copa* », « *la sauva* », « *la lui baillast* », « la lui allast » et « *la leur vouloit* ». Ces séquences reconnues par l'expression rationnelle « *la (lui|leur)<V>* » sont, dans bon nombre de cas, précédé par :

- les pronoms personnels « *je* », « *tu* », « *il* », « *elle* », « *on* », « *nous* », « *vous* », « *ils* » et « *elles* » afin d'identifier des séquences d'ALU comme « *il la voult* », « *nous la verrons* », « *on la recommence* », « *nous la congnoissons* » et « *on la lui envoiroit* » ;

- Les prépositions « *de* » et « *pour* » permettant, par exemple, de former « *bien matiere* ***de la trouver*** », « ***pour la garder*** *et deffendre plus convenablement*», « ***pour la faire*** *monter* » et « ***pour la fonder*** *et ordonner* » ;

- Conjonction de coordination « *et* » et l'adverbe de négation « *ne* » afin de reconnaitre des séquences comme « *il* ***ne la laisse*** *point* », « *on* ***ne la peult*** *scavoir* », « *et* ***ne la retreuve*** *mie* », « *et la merveilleuse habondance* » et « *et la prudence* » ;

- Conjonction du subordination « qui » pour reconnaitre des séquences à l'instar de « *qui la doivent* », « *qui la recoit* » et « *qui la prennent* ».

La forme « *la* » peut être aussi une variante de l'adverbe « *là* ». Notre grammaire contient la modélisation des contextes permettant sa désambiguïsation en produisant le filtre <*ADV*>. En effet, la forme « *la* », suivie de la conjonction de subordination « *où* » ou sa variante « *ou* », d'un groupe nominal, qui peut être composé essentiellement d'un pronom ou d'un substantif, et d'un groupe verbale, qui commence par un verbe, est annotée comme étant l'adverbe « *là* ». Elle permet donc de reconnaitre des longs séquences comme « ***la ou il doit*** *estre employe.* », « ***la ou ilz auront*** *joie eternelle;* », « ***la ou elle est tres dure*** *et quant* », « *la ou la royne Verite tenoit son secret* » et « ***la ou les hommes sont*** *d'une singuliere condicion* ».

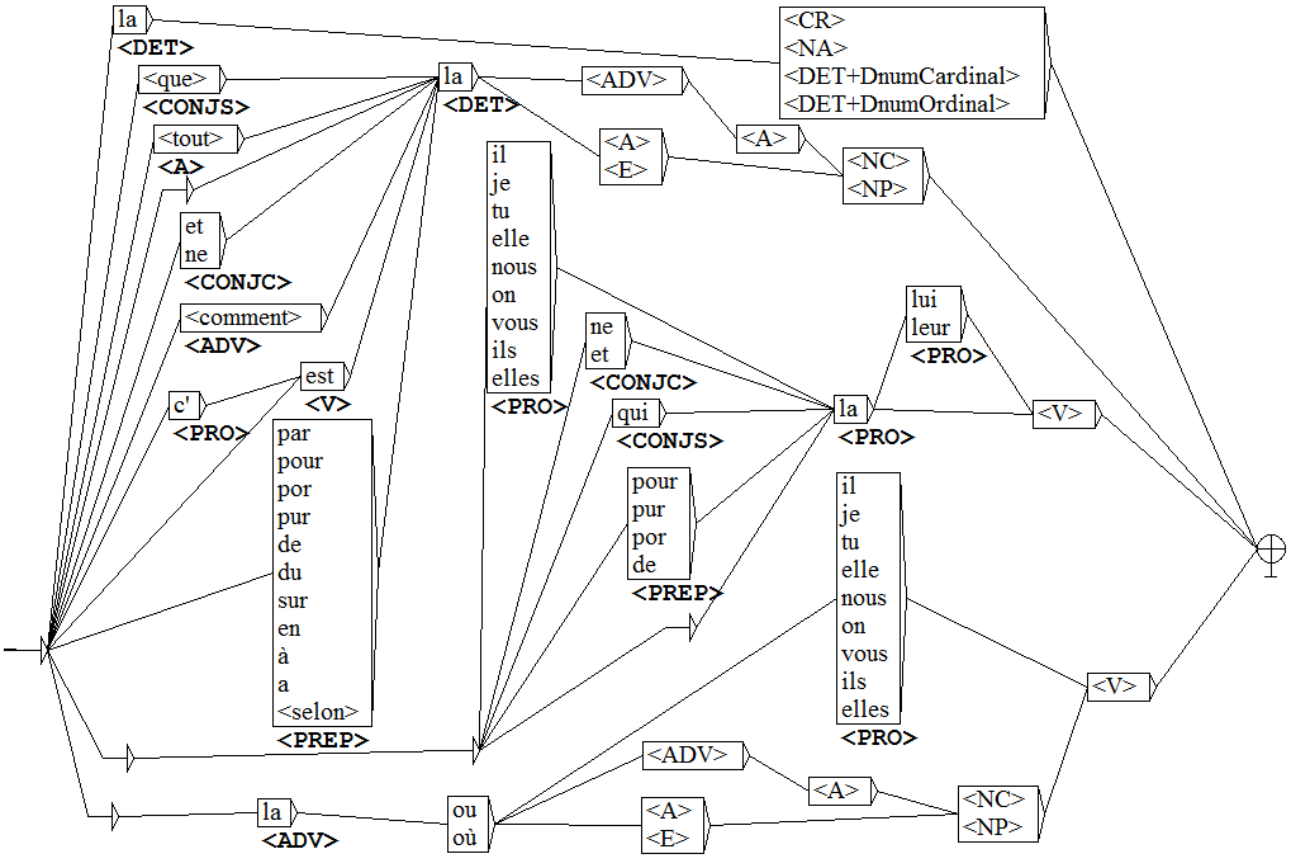


Figure 110. Grammaire de désambiguïsation de la forme « *la* »

« *li* » peut être un déterminant, un pronom ou une variante du verbe « *lire* ». Nous constatons que cette forme est peu employée environ 3486 fois dans l'ensemble du corpus. Elle est surtout fréquente dans les textes du début du XIVème siècle. Cependant, elle reste employée jusqu'au XVème siècle par quelques auteurs, on la rencontre, par exemple, dans quelques textes du Jean Gerson.

« *li* » est déterminant lorsqu'elle est suivie d'un nombre ou d'un nom commun « *uns* ». Tout comme « *le* » et « *la* », « *li* » est fréquemment placé devant un substantif précédé ou non d'adjectif. La reconnaissance des séquences d'ALU, qui constituent des groupes nominaux, permet donc la désambiguïsation de « *li* » en ne gardant dans la TAS que l'étiquette « *DET* ». Cependant, une meilleure description des contextes est possible par l'ajout à l'expression rationnelle « *li (<E>|<ADV>)(<A>)(<NC>|<NP>)| li (<NC>|<NP>)*» des ALU comme :

- les conjonctions de subordination « *que* », « *quand* », « *comme* » et « *dont* » afin d'avoir une description des contextes comme *« que li lieux », « que li petis enffans », « Quant li maison », « quant li rois », « dont li peuples », « dont li philosophe »* et *« dont li rois ».*
- les variantes de l'adjectif « *tout* » permettant une désambiguïsation des éléments des contextes comme « *tous li sains* », « *tous li membre* » et « *tout li royaumes* ».
- les prépositions « *de* », « en », « *par* », « *selon* », « *sur* », « *à* » et « *pour* » et ses variantes « *pur* » et « *por* » afin de reconnaitre des séquences comme *« a li amour », « a li trop grant amor » et « a li femmes ».* Nous constatons que « *li* » est rarement précédé par la préposition « *en* », en effet, elle n'apparait que deux fois « *en li grant abundance* » et « *en li choses* » et c'est dans un même texte à savoir « *Gilles de Rome* ». Nous signalons également que la préposition « *de* » ne précède en aucun cas la structure composé d'un déterminant « *li* » suivi d'un substantif.
- les conjonctions de coordination « *ne* », qui est une variante de « *ni* », permettant la reconnaissance des séquences comme *« ne li père », « ne li princes »* et *« ne li seignur ».*

De même que les formes « *le* » et « *la* », « *li* » est pronom lorsqu'elle est suivie d'un verbe. Des séquences comme « *li tournoit* », « *li sueffrent* » et « *li demonstre* » permettent donc de la désambiguïser. Une modélisation plus détaillée de ces contextes permet une désambiguïsation des ALU :

- les pronoms personnels « *je* », « *tu* », « *il* », « *elle* », « *on* », « *nous* », « *vous* », « *ils* » et « *elles* » afin de reconnaitre des séquences d'ALU comme « *il li aura* », « *je li voys* » et « *il li laissera* »

- les prépositions « *de* » « *pour* » et « à » afin de recenser des séquences comme *« a li sentir », « a li servir », « a li destruire », « de li verser », « de li venir »* et *« De li tourner »*
- l'adverbe de négation « *ne* » afin d'identifier des expressions comme *« car riens **ne li est** que servir Dieu », « ne li sueffrent »* et *« ne li sauerons »*.
- conjonction de subordination « *comme* », « *dont* » et « *que* » pour reconnaitre des séquences à l'instar de *« A fin **que li puist** ordener », « Et ensi apert **que li voloirs** des amis »* et *« li fais **que li doi** portent »*. Nous constatons l'inexistence dans « MEDITEXT », de « *li* » pronom précédé par la relative « *Quand* » et un emploi très rare avec la relative « *dont* » à savoir seulement une seule fois « *dont li vint* » au XVème siècle dans le texte « *Le Chapel des trois fleurs de lis* » de Philippe de Vitri.

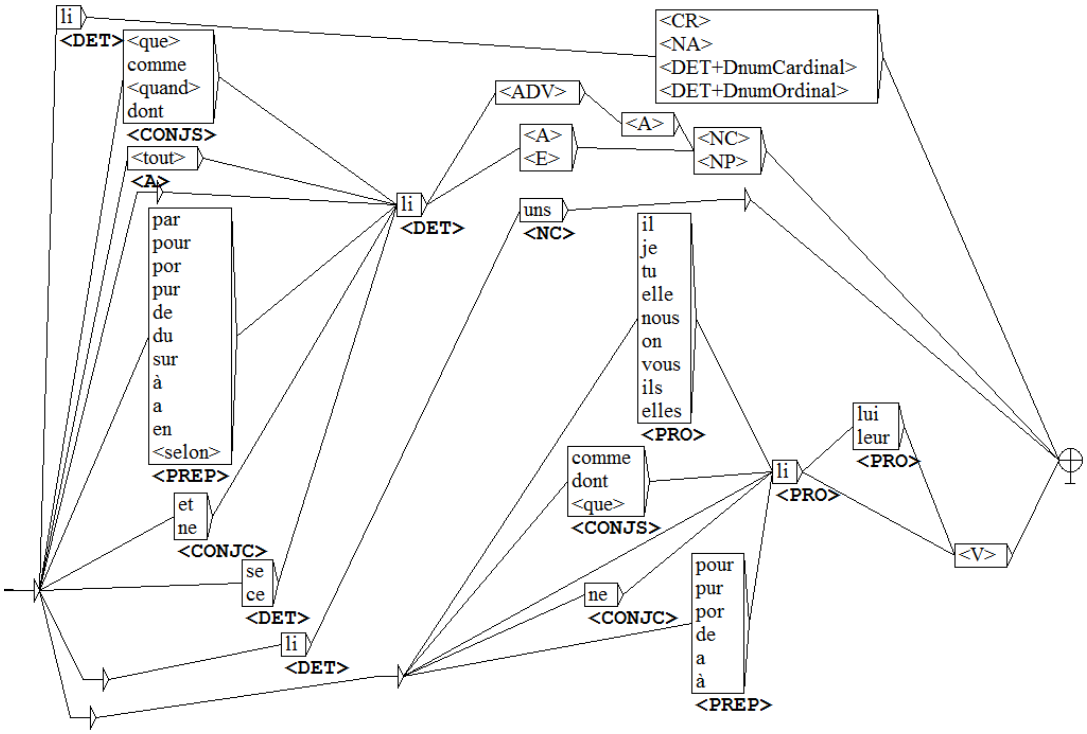


Figure 111. Grammaire de désambiguïsation de la forme « *li* »

Même si on rencontre les formes « *le* » et « *la* » devant voyelle et devant *h* muet ou aspiré comme par exemple « *le article* », « *la esglise* » et « *le emperour* », dans la plupart des cas, elles s'élident pour former des séquences comme « *l'ame* », «*l'ennemi* » et « *l'onneur* ».

Fq	Segment	Lg
2251	de l	2
1175	a l	2
1162	que l	2
1068	en l	2
835	et l	2
445	par l	2
308	pour l	2
302	à l	2
279	ne l	2
243	de l'eglise	2
240	comme l	2
220	l'an de	2
211	l'an de grace	3
192	qui l	2
165	Comment l	2
158	l'an de grace mil	4
154	l'umble supplication	2
152	droit et l	3
137	quant l	2
136	est l	2
135	l'umble supplication de	3
134	receu l	2
131	quoi l	2
129	avoir receu l	3
127	receue l'umble	2
127	avoir receue l'umble	3
126	nous avoir receue l'umble	4
123	comment l	2

Figure 112. Segments répétées les plus fréquents contenant la forme « l' »

De même que « *le* » et « *la* », « *l'* » est un déterminant lorsqu'elle est placée devant un substantif. Comme le montre la grammaire de désambiguïsation de la figure 113, elle peut donc être précédée par des prépositions comme « *de* », « *en* », « *par* » et « *pour* » afin de décrire des contextes comme « *de l'ennemy* », « *de l'eglise* », « *a l'eure* », « *a l'exemple* », « *en l'orde* », « *par l'envie* » et « *pour l'affection* », par des conjonctions de subordination « *qui* », « *que* », « *quand* » et « *comme* » pour reconnaitre des séquences à l'instar de « *qui l'euvre* » « *quant l'ame* » et « *quant l'aveugle* », par l'adverbe de négation « *ne* » pour modéliser des expressions comme « *ne l'omme* » et « *ne l'ombre* » ou l'adverbe « *tout* » et ses variantes pour identifier et désambiguïser des séquences comme « *tout l'empire* », « *tout l'effort* » et « *tout l'ouvrage* ». Cependant, nous constatons que la forme élidée « *l'* » est rarement utilisée devant un nombre sauf pour les numéraux ordinaux commençant par voyelle comme « *l'onziesme* ».

« *l'* » est pronom lorsqu'il est suivi d'un verbe, cependant, il n'est jamais suivi des pronoms « *lui* » et « *leur* ». La simple expression « *l'*<*V*> » pourrait donc désambiguïser les cas où « *l'* » est pronom. Afin de permettre une description plus fine et une désambiguïsation des ALU dans les contextes immédiats de « *l'* », les séquences reconnues par cette expression rationnelle pourraient donc être précédé par les prépositions « *de* », « *à* » et « *pour* » afin de

reconnaitre des expressions comme *« de l'empescher », « de l'enseigner », « a l'escouter », « a l'entendre »* et *« pour l'aider »*, par les conjonctions de coordination « et » et « ne » pour former des séquences à l'instar de *« et l'accorda », « et l'envahir », « et l'ordonna », « ne l'appella »* et *« ne l'explique »* ou par des conjonction de subordination comme « qui » qui est fréquemment employé *« qui l'envoye » ,« qui l'ensuivent »* et *« qui l'espousera », « comme »* qui est peu fréquente *« comme l'exposent », « comme l'afferma »* et ***« comme l'avons** ordonné », « quant » qui est rarement utilisée « **quant l'ai** laissie » et* qu'on la rencontre uniquement au XIVème siècle dans un texte religieux de Guillaume de Digulleville et la conjonction de subordination « que » qui est souvent employé suivi du verbe « *être* » et d'un participe à titre d'exemple nous citons *« q**ue l'avoit** fait », « **que l'ay** nourry », « **que l'avons** juré », **« que l'aurez** acceptée » et « **que l'eus** demandée »*. Nous notons également que l'emploi de l'adverbe interrogatife « *comment* » avec le pronom « *l'* » est très rare, nous avons recensé un exemple unique en vers « ***Comment l'as** tu osé songier? »* dans un poème au XVème siècle d'un auteur anonyme.

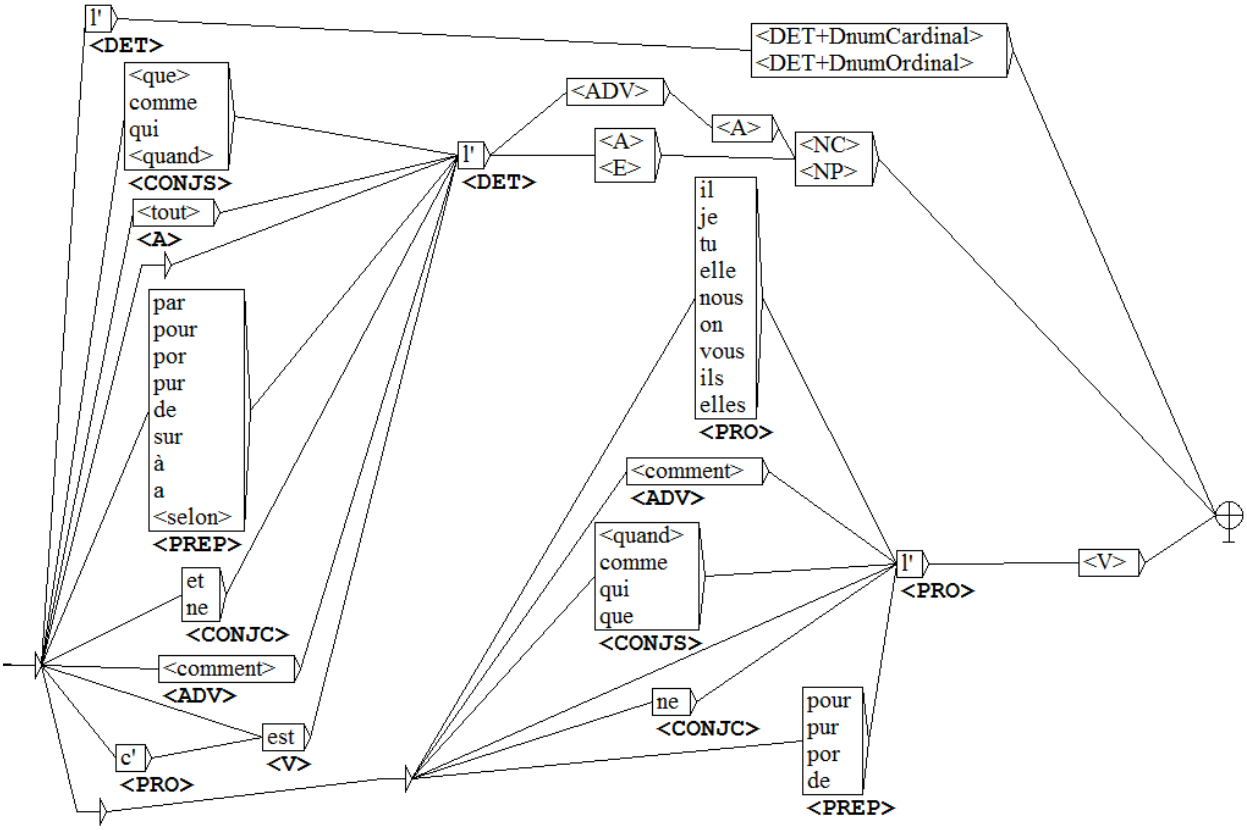


Figure 113. Grammaire de désambiguïsation de la forme « *l'* »

« *les* » peut-être un article définit, pronom, une variante du verbe « *laisser* », du nom « *lé* » ou de la préposition « *lez* ». La grammaire de désambiguïsation de la figure 114 décrit les contextes d'apparition de la forme « *les* » comme déterminant ou pronom. En effet, « *les* »

fonctionne comme déterminant lorsqu'elle précède un substantif. Elle peut également être précédé par de nombreuses prépositions qui sont fréquemment utilisées comme « *par* », « *pour* », « *contre* », « *entre* », « *sur* » et « *avec* » afin de décrire des contextes comme par exemple *« par les detestables suppoz », « par les anciennes histoires », « pour les juges »*, « *contre les orguilleux clers* », « *contre les mauvais ennemis* », *« entre les vieilles sorcieres »*, « *avec les philozophes* » et « *en les jennes genz* » et d'autres moins fréquentes comme « *de* » permettant la reconnaissance de *« de les coustumes », « de les comptes »* et *« De les chartres »*, par des conjonctions de subordination « *comme* », « *qui* » et « *que* » afin d'identifier des séquences telles *« que les tres nobles princes », « qui les ames », « qui les lettres », « comme les anciens Romains »* et *« comme les bons anges »* et par des conjonctions de coordination comme *« et »* et *« ne »* afin de désambiguïser les séquences *« et les tenebres », « et les detestables erreurs », « et les très abhominables ordures », « ne les chiens »* et *« ne les sciences »*. Nous signalons également l'utilisation fréquente de *« les »* déterminant suivi du nom commun *« autres »* et de *« les »* précédée par l'adjectif *« toutes »* ou *« tous »*.

« *les* » peut être désambiguïsée en préservant l'annotation du pronom *« PRO »* dans la TAS dans le cas où elle est suivie par un verbe. Une modélisation plus détaillé des contextes permettent de désambiguïser des ALU comme les conjonctions de coordination *« et »* et l'adverbe de négation *« ne »* permettant la reconnaissance des expressions telles « *et les tient* », « *et les deffendent* », « *on* ***ne les soufferra*** *paisiblement* » « *et* ***ne les veoit*** *on es lieux de leurs* » et « *se pecheur* ***ne les demandoit*** *pas et les ressoingnoit.* », les prépositions « *de* », « *en* » et « *pour* » donnant la possibilité d'identifier les expressions « *garder* ***de les acomplir*** », « *pour les gaignier* » et « *et l'****en les vouloit*** *gaiger pour cause et occasion* » et finalement les conjonctions de subordination, notons que « *qui* » est la plus fréquente permettant de former des séquences comme « ***qui les verra*** *mettre* », « ***qui les obligent*** *a entendre* » et « ***qui les prie*** *en alegeant sa povreté* », cependant « *comme* » et « *que* » sont moins fréquentes et on les retrouve plus au XVème siècle comme le montre ces exemples extraits d'un acte de Louis XI « « *affin* ***que les montrez*** *aux gens* », « ***que les vueillez*** *croire* » et « *vous mandons* ***que les croiez*** *de tout ce qu'ilz* ».

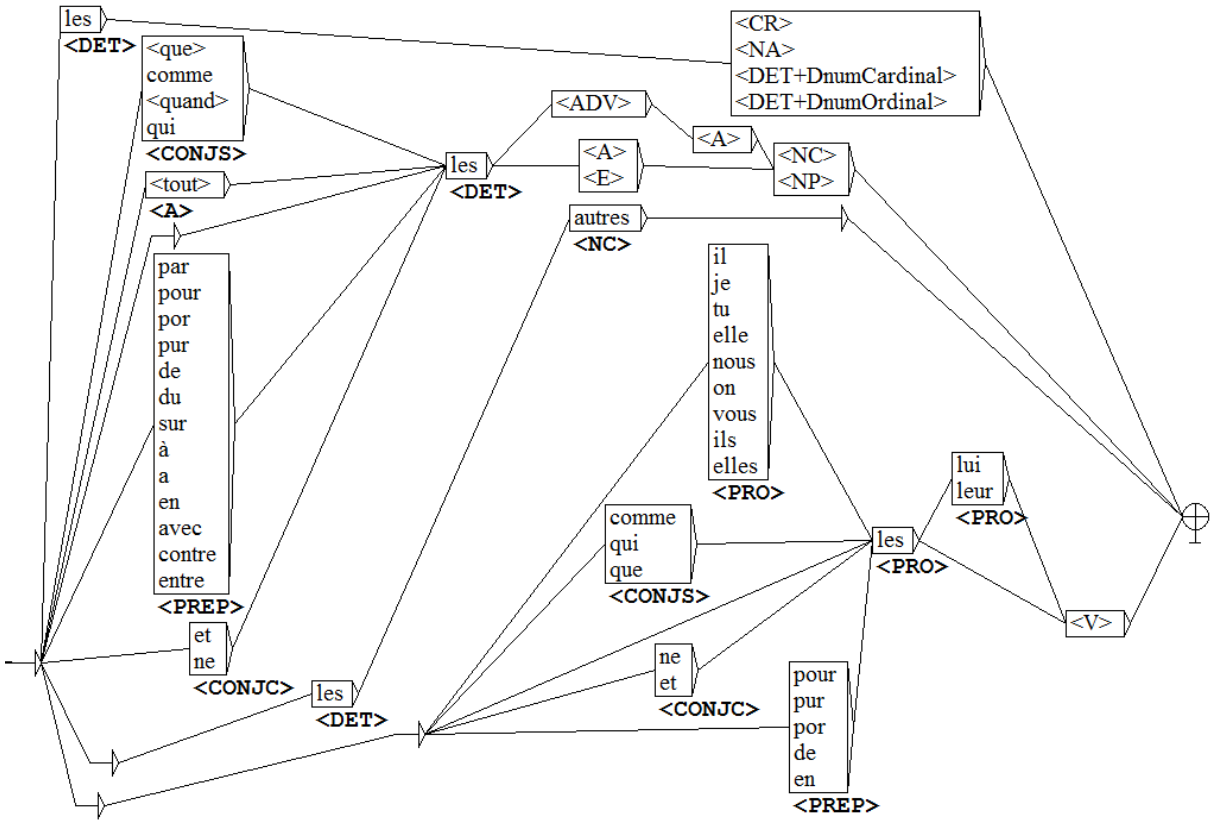


Figure 114. Grammaire de désambiguïsation de la forme « *les* »

3.2. Désambiguïsation des articles indéfinis « un », « une », « de » et « des » et des articles partitifs « de » et « des »

Bien qu'elles soient rarement utilisées, les formes déclinées de « *un* » comme « *unes* », « *uns* », « *unz* » ou « *ungs* » sont bien attestées dans certains textes de MEDITEXT de périodes et d'origines différentes. Ces articles indéfinis pluriels, qui ont été surtout employés au XIVème siècle, servent à déterminer des substantifs pluriels plus précisément des noms collectifs comme « *uns roys* », « *uns philosophes* » et « *ungs ciseaulx* ». D'autres articles indéfinis pluriels sont employés plus régulièrement que les formes déclinées de « *un* » sur l'ensemble du corpus et surtout au XVème siècle à savoir « *des* » et « *de* ». Mais ces derniers restent peu fréquents à cause de la pratique répondue à l'époque de ne pas employer un déterminant devant un substantif pluriel.

Comme montre la grammaire de la figure 115, les articles indéfinis peuvent être utilisés devant un substantif ou un groupe adjectival, qui peut être composé d'un ou deux adjectifs précédés ou non d'un adverbe, suivi d'un substantif pour former des séquences comme « *ung roy* », « *ung simple pescheur* », « *une noble et riche dame* », « *une forte et sainte pensée* », « *une tres haulte montaigne* » et « *ung tres mauvais tresor* ». Ils peuvent également être employés devant les différentes variantes des indéfinis « *autre* », « *tel* », « *chacun* » et

« *même* » afin de reconnaitre des séquences comme « *une aultre beste* », « *ung mesme lieu* » et « *un tel prince* ».

L'emploi d'un indéfini devant un complément pour exprimer une comparaison permet la désambigüisation en produisant le filtre correspondant à un déterminant <*DET*>. Ces expressions de comparaison sont composées essentiellement du l'adverbe « *comme* » suivi de l'article indéfini suivi d'un nom commun. Nous citons, à titre d'exemple, les expressions « *comme une puissante et forte tour*», « *comme des mauvais* », « *comme ung homme* », « *comme ung bon chevalier* » et « *comme une tres forte tour* ». La description du contexte à l'aide de l'expression rationnelle « *il y <avoir>* » permet de désambigüiser les indéfinis qui suivent cette expression comme par exemple « *il y a de belles dames* », « *il y a des poissons* », « *il y ait ung gibet*», « *il y aura un contreroleur* » et « *il y avoit ung homme* ». Comme les articles définis non élidés, les articles indéfinis au singulier « *un* », « *une* » et leurs variantes orthographiques peuvent être employés devant un numéral cardinal ou ordinal. Mais cet emploi de l'article indéfini est rare et on le rencontre avec presque la même fréquence au XIVème et au XVème siècle, à titre d'exemple, nous citons « *un cent* », « *un troisième* » et « *une quinzisme* ». Un autre emploi bien particulier des articles indéfinis « *un* » et « *une* » est dans un contexte immédiat gauche commençant par « tout » à titre d'exemple nous citons *« tout ung pays »* et *« tout ung grant royaulme ».*

Notre grammaire annote en préposition les articles partitifs « *de* » et « *des* ». Comme le montre notre grammaire de la figure suivante, ils sont fréquemment utilisés dans une construction syntaxique de type « *NOM PREP NOM* » comme « *createur du monde* » et « *lumiere des tenebres* ». Ils sont également employés après les adverbes de quantité comme « *assez* », « *autant* », « *tant* », « *beaucoup* », « *moins* », « *mout* », « *petit* », « *peu* », « *plus* », « *trop* », « *pas* », « *point* » et « *mie* » afin de reconnaitre des séquences comme « *plus de gens* », « *moult de choses* » et « *peu de negligence* ». Cependant, les cas où les partitifs précèdent un défini tels que « *le* », « *la* », « *li* », « *l'* » et « *les* », ils sont traités par les grammaires concernant ces derniers en 3.1.

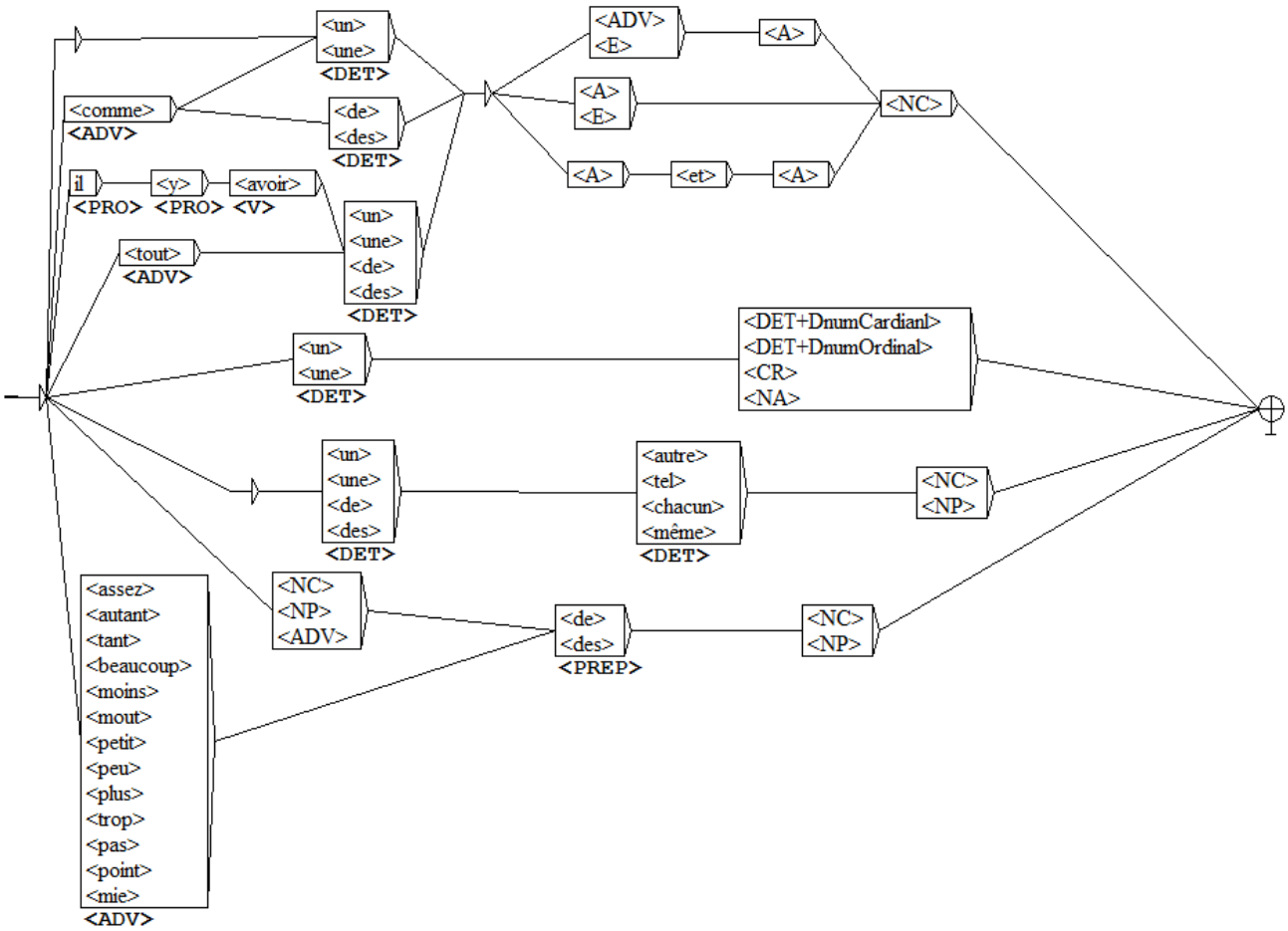


Figure 115. Grammaire de désambiguïsation des formes « *un* », « *une* » « *de* » et « *des* »

3.3. Désambiguïsation des pronoms et déterminants démonstratifs

Nous avons recensé plusieurs variantes des déterminants démonstratifs singuliers et pluriels à savoir « *ce* », « *cet* », « *cel* », « *cest* », « *cist* », « *cil* », « *ceste* », « *cele* », « *cette* » et « *ces* ». L'analyse textométrique montre que dans la plupart des cas, les déterminants démonstratifs sont suivis soit par un nom, soit par un adjectif ou par un adverbe. Nous faisons référence aux travaux sur l'évolution des démonstratifs effectués par (Dees, 1971) dans lesquels il montre que les déterminants démonstratifs masculins singuliers « *cist* » et « *cil* » sont utilisés dans des cas sujet à la première moitié du XIVème siècle, les déterminants démonstratifs masculins singuliers « *cel* » et « *cest* » continuent à être employés dans des cas régimes devant voyelle jusqu'au XVème siècle. Ils disparaissent laissant la place aux « *cet* » et « *ce* », tandis que le déterminant démonstratif féminin singulier « *cette* » remplace les formes « *ceste* » et « *cele* ».

Le déterminant démonstratif « *ce* » est le plus fréquemment utilisé, tandis que « *cet* » est la forme la plus rare avec seulement 19 occurrences dans MEDITEXT. Nous constatons également que « *ces* » est la forme la plus courante devant un substantif pluriel.

Comme montre la figure 116, plusieurs segments répétés fréquents contenant la forme « *ce* » font partie d'un pronom relatif composé comme « *ce que* », « *ce qui* », « *ce qu* », « *pour ce que* », « *de ce que* », « *ce que le* », « *et pour ce que* » et « *pour ce qu* ». Ces cas vont être reconnu et analysé grâce aux mots composés contenus dans notre dictionnaire électronique qui permettra donc d'annoter les mots composés « *ce que* », « *ce qui* » et « *ce qu* » ainsi que leurs variantes orthographiques en pronom.

Fq	Segment	Lg
4070	ce que	2
3197	pour ce	2
1842	de ce	2
1326	sur ce	2
1245	ce qu	2
1109	en ce	2
1103	pour ce que	3
1097	que ce	2
750	a ce	2
697	Et pour ce	3
666	ce qui	2
565	ce soit	2
547	tout ce	2
518	que ce soit	3
489	et pour ce	3
481	par ce	2
463	Pour ce	2
462	Et ce	2
422	ce monde	2
355	ce que dit	3
325	de ce que	3
323	pour ce qu	3
313	ce qu'il	2
311	et ce	2
307	ce fait	2
301	ce n	2
291	ce ne	2
278	ce que dit est	4

Figure 116. Segments répétés contenant la forme « *ce* »

La grammaire de la figure 117 permet de lever l'ambiguïté des déterminants démonstratifs singuliers et pluriels puisqu'ils sont souvent suivis soit d'un substantif, d'un adjectif ou d'un adverbe. L'analyse textométrique montre que, dans 45% des cas, ils sont précédés par une préposition. Les plus fréquentes de ces prépositions sont « *de* », « *sur* », « *en* », « *à* », « *par* », « *sans* » et « *pour* » permettant ainsi la reconnaissance des séquences comme « *à ce temps* », « *de ce monde* », « ***de ce dangereux** pas* », « ***de ces tres** horribles malefices* », « *en ce mespris* », « ***en ces derreniers** jours* », « *en cest estat* », « ***par ce roy** orgueilleux* » et « ***sur ce silence** perpetuel* ». D'autres séquences moins fréquentes dans lesquelles les déterminants démonstratifs apparaissent précédés d'un adverbe « *avant* » ou

« *après* » ou l'adjectif « *tout* » comme par exemple « *aprés cel jour* », « *apres cest present article* », « *avaunt ces houres* », « *totes ces choses* » et « *toutes ces manieres* ».

De même que les déterminants, les pronoms démonstratifs sont divers et ils ont subi une importante évolution. Les formes « *cestui celui* », « *ceus-ci* », « *ceus-la* », « *ceste* », « *cele* », « *cestes* » et « *celes* » sont les plus employées au XIVème siècle et les formes « *cist* » et « *cil* » sont très rares. Les pronoms « *cestui-ci* », « *ceste-la* », « *ceste-ci* », « *cele-la* », « *celes-la* », « *celui-là* » et « *cestui-là* » font leurs apparition au XVème siècle. Nous constatons que les pronoms possessifs ne sont pas fréquemment utilisés par les auteurs à l'époque du moyen français et que, dans 84% des cas, le pronom démonstratif est suivi par un verbe, soit par un nom, soit par un adjectif, soit par les par une préposition « *par* » ou « *de* » soit une relative « *que* » ou « *qui* ». Notre grammaire permet donc la reconnaissance des séquences comme « ***celui de*** *Lucifer* », « ***Celui par*** *qui sont honnourez* », « ***celui de*** *l'apostre* », « ***celui que*** *confesse* » et « ***celui qui*** *est de meilleur foy* ».

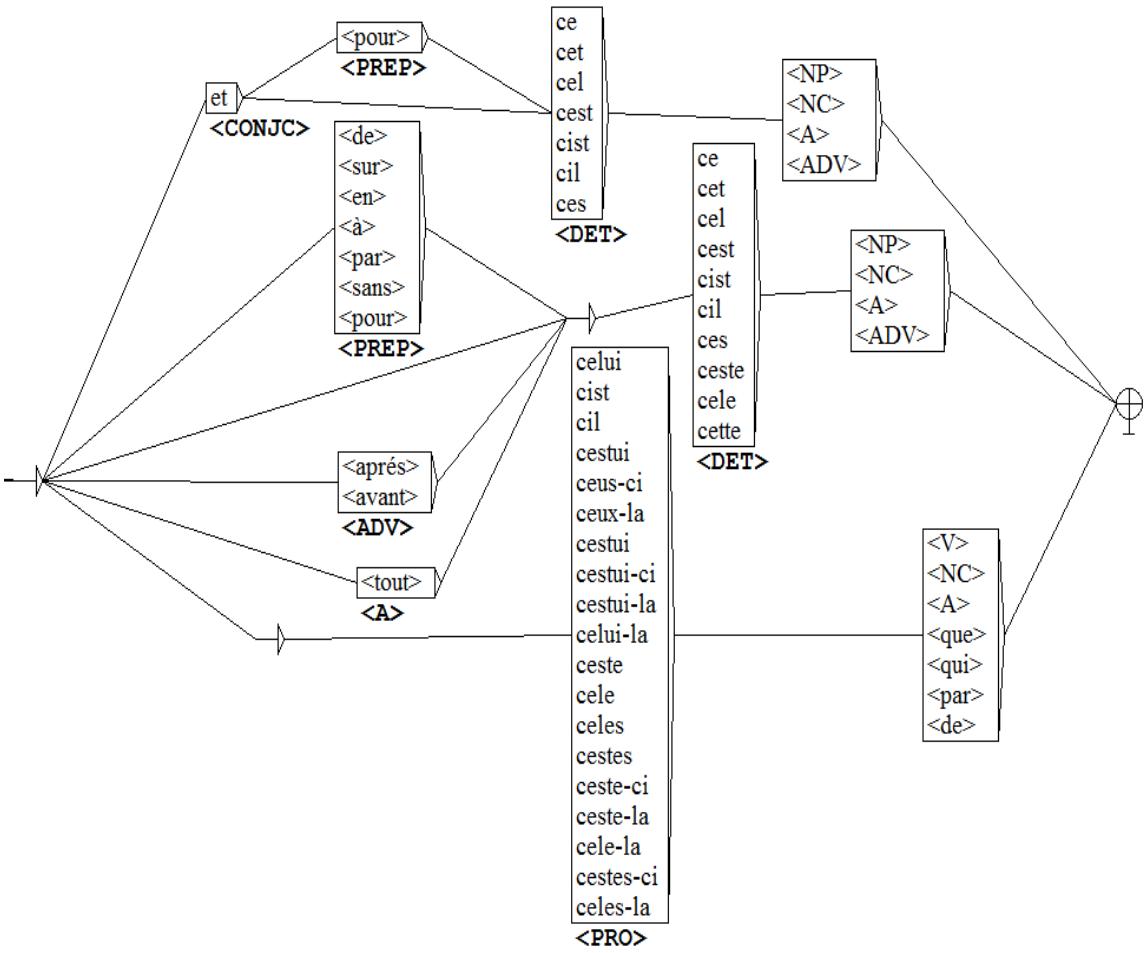


Figure 117. Grammaire de désambiguïsation des pronoms et des adjectifs démonstratifs

3.4. Désambiguïsation des adjectifs et pronoms possessifs

Les adjectifs possessifs varient selon le genre et le nombre du substantif qu'ils déterminent. A chaque adjectif possessif, nous avons associé différentes variantes orthographiques. Par exemple, les formes « *men* », « *mez* », « *moun* » et « *mun* » sont considérées comme des variantes orthographiques du l'adjectif possessif masculin singulier « *mon* ». De même, pour les formes « *no* », « *nostre* » et « *notres* » qui sont des possessifs rares qu'on rencontre dans des textes généralement influencés par des dialectes régionaux, ils n'ont pas été traité comme des possessifs autonomes, mais ont été considérés comme des variantes de l'adjectif possessif pluriel « *notre* ».

Tout comme pour le français moderne, l'adjectif possessif peut être suivi d'un nom c'est le cas le plus fréquent. Notre grammaire de désambiguïsation de la figure 118, illustre environ 80% des cas d'utilisation de l'adjectif possessif dans MEDITEXT à savoir lorsqu'il est suivi soit par un nom, par un adjectif ou par un adverbe afin de reconnaitre des séquences comme « *mon père* », « *sa robe* », « *son throsne* », « *leurs mains* », « *nostre Dame* », « *nostre Seigneur* », « *vostre pensee* », « *ta povre mere* », « *sa doulce mere* » et « *mon tres chier enfant* ».

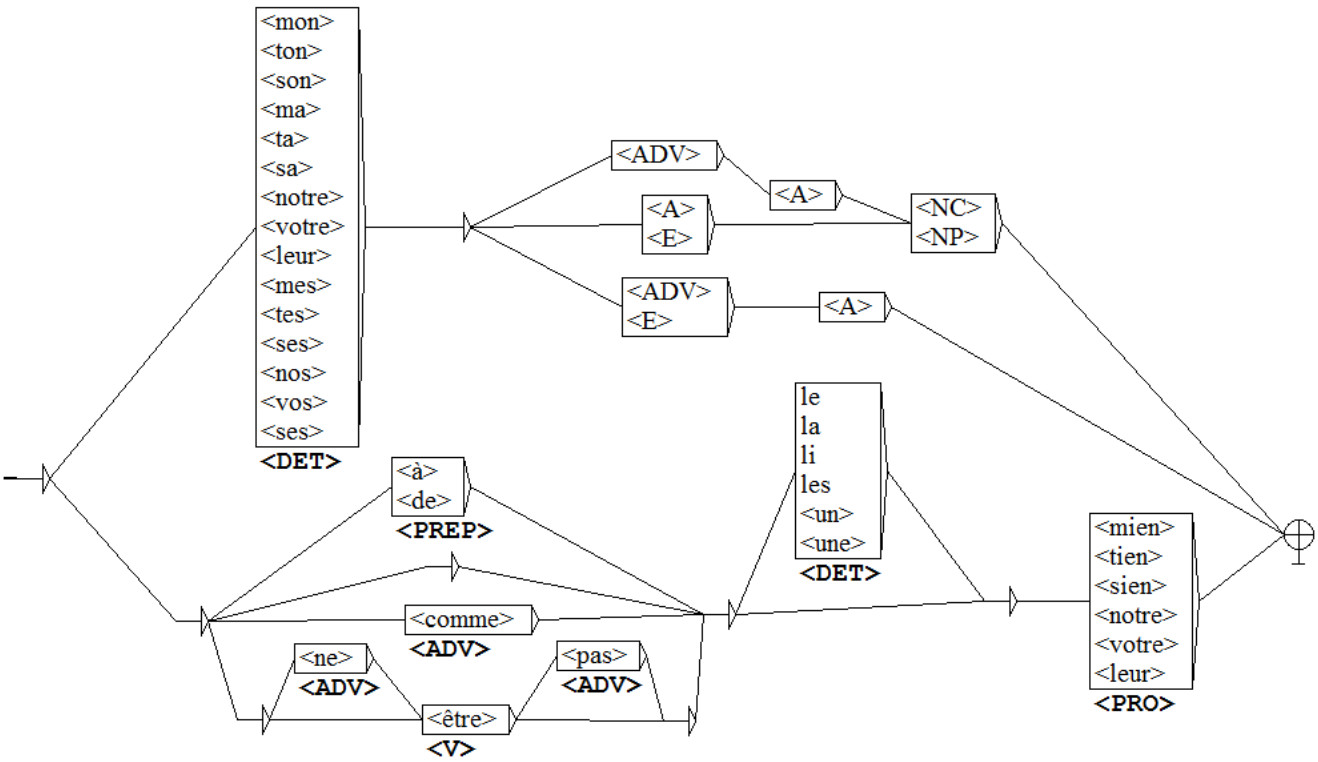


Figure 118. Grammaire de désambiguïsation des pronoms et des adjectifs possessifs

Les pronoms possessifs « *mien* », « *tien* », « *sien* », « *notre* », « *votre* » et « *leur* » ainsi que leurs différentes variantes comme « *noz* », « *voz* », « *lor* », « *leurs* », « *mienne* »,

« *tienne* », « *tiens* » « *miens* » et « *siens* » sont souvent précédés par le verbe « *être* », par une préposition « *à* » ou « *de* », par un adverbe « *comment* » et ou par les articles définis « *le* », « *la* », « *li* » ou « *les* » et indéfinis « *un* » ou « *une* ».

3.5. Désambiguïsation des adjectifs/adverbes/pronoms indéfinis

Ces indéfinis regroupent divers termes de nature différente : adjectifs, adverbes et pronoms. Ils sont considérés comme des mots grammaticaux qui peuvent fonctionner comme déterminant ou pronom. Ils partagent en commun la possibilité de marquer une information sur l'existence, la quantité et l'identité (Marchello-Nizia, 2005). Notre grammaire de la figure 119 recense une liste non-exhaustive des indéfinis employés dans MEDITEXT. Leur fréquence d'utilisation est bien variable et elle peut être influencée, pour certains termes, par le type et le genre de texte, ainsi que par la situation géographique et chronologique de l'auteur.

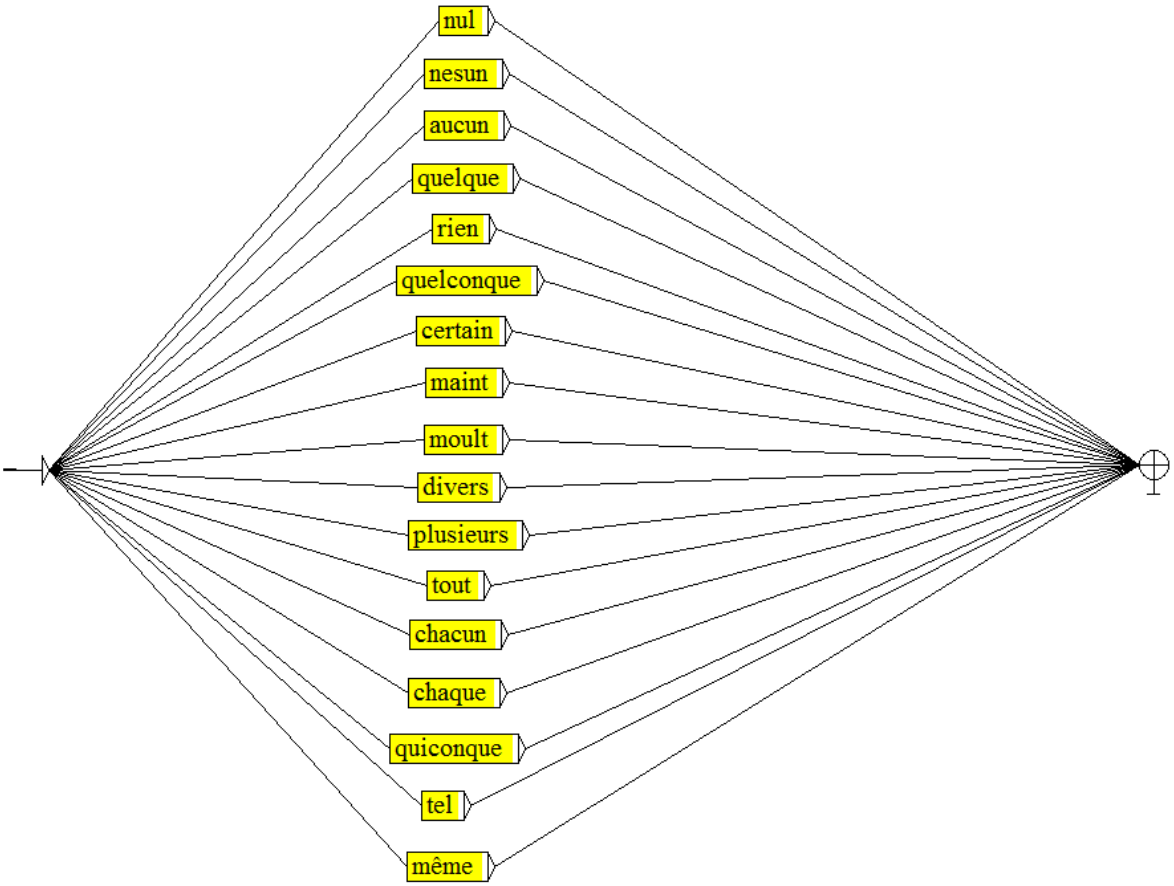


Figure 119. Graphe principal de désambiguïsation des adjectifs, pronoms et adverbes indéfinis

L'indéfini « *nul* », ainsi que les indéfinis « *nesun* » et « *aucun* », indiquent une quantité nulle et une existence niée sans préciser l'identité et ils sont généralement accompagnés de l'adverbe de négation « *ne* ». « *nul* » est fréquemment utilisé par rapport à « *nesun* » ou à

« *aucun* » et qui a de nombreuse variantes comme « *nule* », « *nuli* », « *null* », « *nulle* » et « *nuilly* ». Comme le montre la grammaire de la figure 120, il peut être désambiguïsé en lui attribuant l'étiquette d'un déterminant « *DET* » lorsqu'il est placé devant un substantif. C'est le cas le plus fréquent d'utilisation de cet indéfini qui couvre environ 57% des cas. Notre grammaire permet donc la reconnaissance des séquences comme « *nul serment* », « *nulle comparaison* » et « *nul miracle* ». Le substantif peut être précédé d'un groupe adjectival et/ou un déterminant afin de décrire des séquences comme « *nule petite peine* », « *nulles autres generacions* » et « *nulles grantz Charges* ». Cependant, lorsqu'il est suivi d'un verbe, « *nul* » est annoté comme un pronom « *PRO* ». Dans ce cas, notre grammaire permet de capturer diverses séquences comme par exemple « *nulz ne parle* », « *nul ne peche* », « *nul ne vouldroit* », « *nul que puissez* » et « *nully q'avoit* ».

Nous notons que, de même pour « *nesun* » et « *aucun* » qui sont désambiguïsés grâce à une grammaire avec une description des contextes presque identiques à celle du « *nul* ». Ces grammaires permettent donc de produire l'annotation « *DET* » lorsque « *nesun* » et « *aucun* » sont suivis d'un substantif précédé ou non d'un groupe adjectival et de générer l'annotation « *PRO* » lorsqu'ils sont placés devant un verbe.

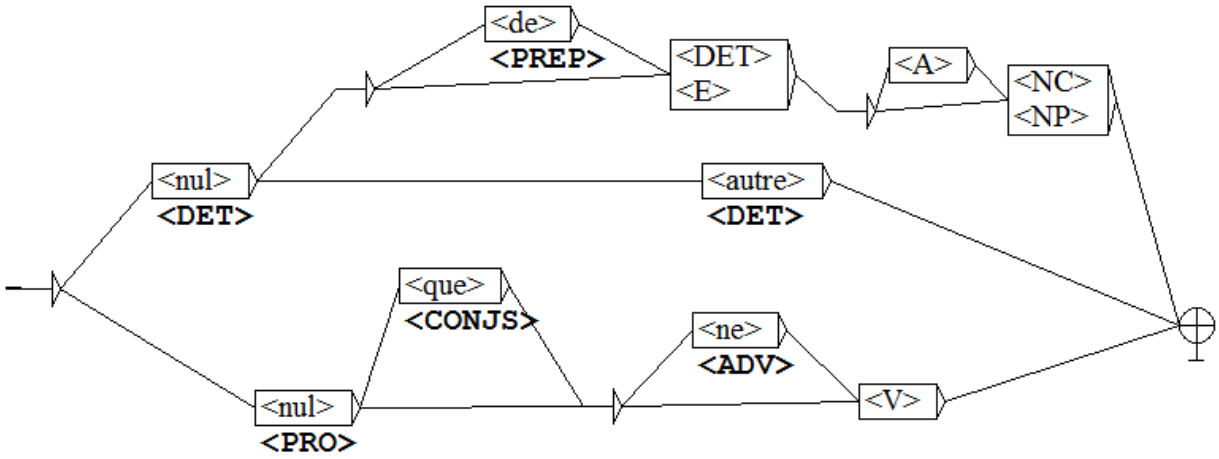


Figure 120. Grammaire de désambiguïsation de « *nul* »

L'indéfini « *quelque* » est peu fréquent dans MEDITEXT mais il est loin d'être rare. Il est déterminant d'un substantif. Comme le montre la grammaire de la figure ci-dessous, il est souvent précédé d'une préposition à savoir « *sans* », « *par* » et « *à* » qui sont les plus employées. Notre grammaire permet donc de désambiguïser « *quelque* » lorsqu'il apparait dans des séquences d'ALU comme par exemple « *quelque pierre* », « *quelque belle lance* », « *par quelque painture* », « *par quelque legiere passion* » et « *par quelque bonne operacion* ».

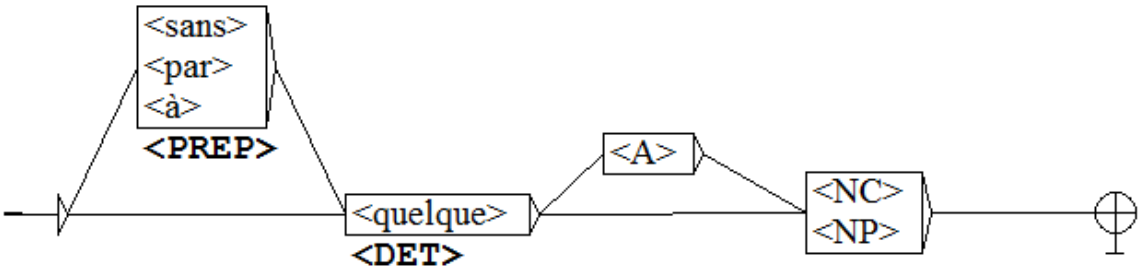


Figure 121. Grammaire de désambiguïsation de « *quelque* »

« *rien* » est un pronom courant dans MEDITEXT. Il est souvent accompagné de l'adverbe « *ne* » et d'un verbe. La grammaire de désambiguïsation se compose essentiellement de deux séries de règles : la première lorsque « *rien* » est placé après le verbe permettant ainsi de décrire des séquences comme « *pour ce que* ***conquester riens*** », « *je* ***n'entens riens*** », « *je* ***ne la despouille de riens*** », « *Elle* ***ne souffriroit pour riens*** » et « *se peut* ***nommer sans riens*** », la deuxième décrit les séquences où « *rien* » précède un verbe comme par exemple « *sans riens mangier* », « *riens n'est empeschee* », « *Mais le cuer* ***de riens n'y atouche*** », « *sans riens prendre* » et « *qui* ***pour riens ne tuent*** *mon filz* ».

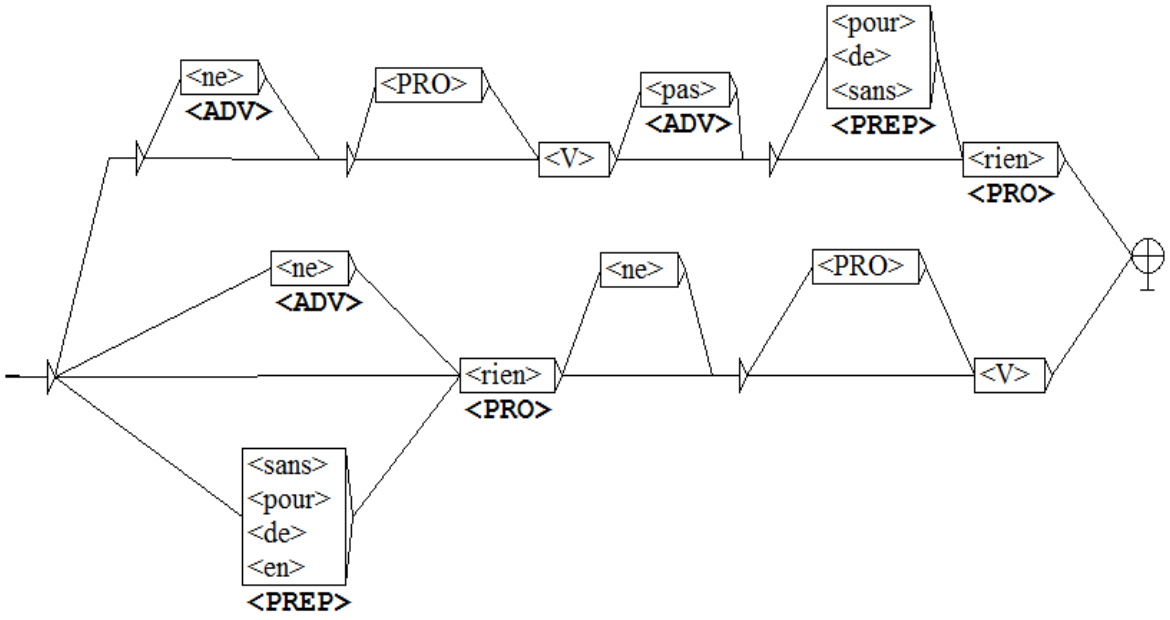


Figure 122. Grammaire de désambiguïsation de « *rien* »

« *certain* », « *maint* » et « *moult* » sont peu fréquents dans MEDITEXT et c'est presque la même description des contextes qui permet d'en lever l'ambiguïté. Par exemple, notre grammaire de la figure ci-dessous permet la désambiguïsation d'environ 71% des occurrences de « *certain* » attestées dans MEDITEXT. Cet indéfini s'emploie suivi d'un verbe qui peut être précédé ou non d'un adverbe de négation « *ne* » et/ou d'un pronom. Cette description de contexte permet donc de lever l'ambiguïté de « *certain* » en produisant l'étiquette pronom

« *PRO* » lorsqu'il apparait dans des séquences comme « *certain qui considere* », « *certain n'y estoit* » et « *certain le gouvernerent* ». Il peut être également annoté comme déterminant « DET » lorsqu'il se rencontre devant un substantif précédé ou non d'un adjectif et/ou d'un déterminant comme par exemple dans les expressions « *certaines chozes* » et « *certaines bonnes personnes* ».

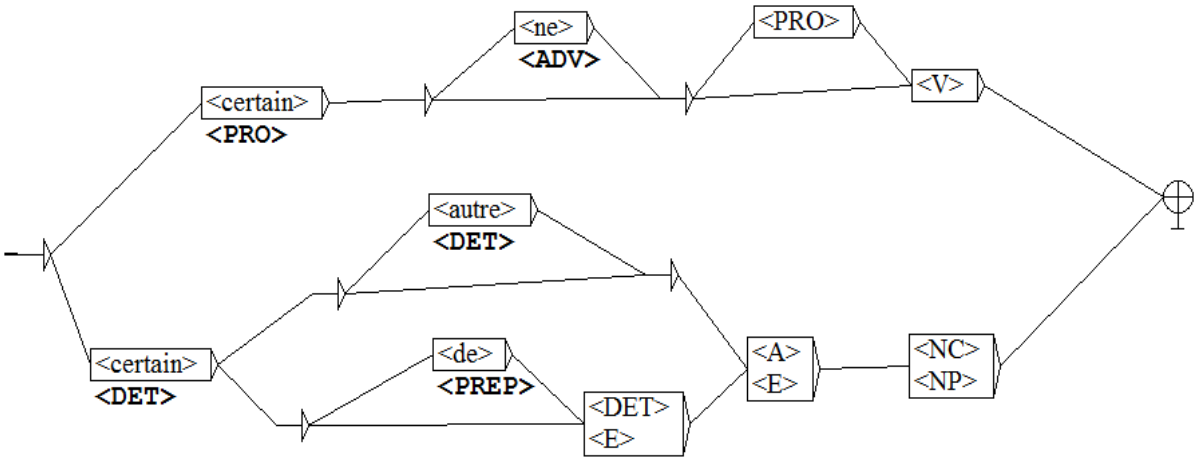


Figure 123. Grammaire de désambiguïsation de « *certain* »

L'indéfini « *tout* » est un terme exprimant la totalité d'éléments dont l'existence est affirmée, mais dont l'identité n'est pas définie (Marchello-Nizia, 2005). Il est fréquemment employé dans les différents textes du MEDITEXT quelles que soient la période et l'origine géographique de l'auteur ou le type et le genre du texte. Comme le montre la grammaire de la figure ci-dessous, il peut être annoté comme déterminant « *DET* » lorsqu'il est placé devant un substantif précédé ou non d'un déterminant et/ou d'un adjectif comme par exemple « *toutes choses* », « *toute creature* », « *tout le monde* » « ***toute mauvaise** fiction* », « ***toute bonne** police* » et « ***toute plaine** riviere* ». Comme il peut être suivi du « *autre* », de l'indéfini « *chacun* », du nom « *sorte* » pour former « *toute sorte* » ou d'un nombre cardinal précédé ou non d'un article. Il s'emploie également précédé de la préposition « *pour* ». Notre grammaire reconnait également des séquences comme « *tout vient* » et « *tout ne souffist* » qui permettent de désambiguïser « *tout* » en lui attribuant l'étiquette pronom « *PRO* » lorsqu'il est suivi d'un verbe précédé ou non de l'adverbe de négation « *ne* » et/ou un pronom. D'autres séquences sont reconnues par notre grammaire du pronom « *tout* » lorsqu'il se rencontre après l'adverbe « *comme* » ou la préposition « *en* » comme par exemple « *les aucuns **en tout**,* » « *qui est **en tout**,* »et « *non pas **en tous**,* ».

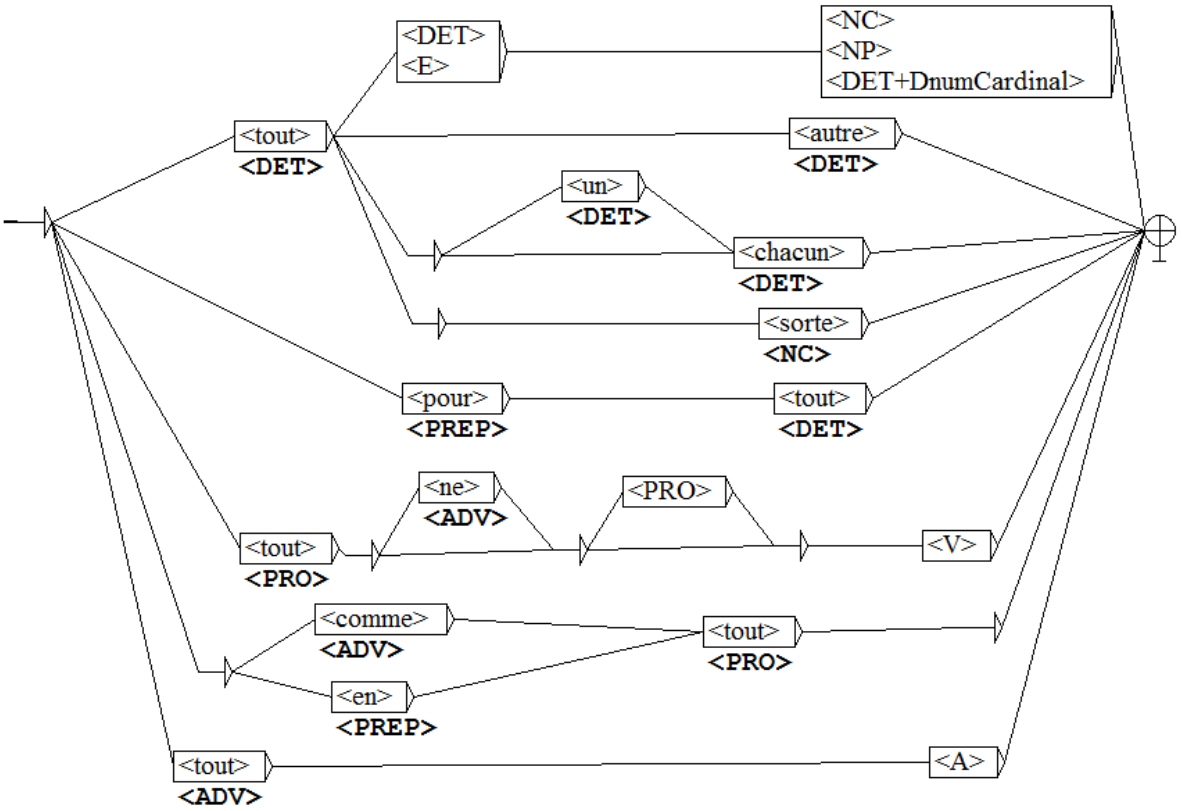


Figure 124. Grammaire de désambiguïsation de « *tout* »

Si « tout » est le seul à pouvoir exprimer la globalité, il existe en moyen français d'autres indéfinis utilisés pour exprimer la totalité d'éléments comme « *chacun* », « *chaque* » et « *quiconque* » qui ont des fréquences très variables. « *chacun* » reste de loin le plus courant et il n'est pas vraiment concurrencé par « *chaque* ». Ce dernier, il est très rare, on le rencontre au XVème siècle dans trois textes du MEDITEXT à savoir un discours et deux serments. Plus fréquent que « *chaque* » mais loin d'être courant, « *quiconque* » est utilisé le plus souvent au XIVème siècle avec, dans 63% des cas, la spécificité de pouvoir être utilisé comme un relatif indéfini. Nous constatons que la même description des contextes peut pratiquement être utilisée pour désambigüiser ces trois indéfinis. La grammaire de la figure ci-dessous permet donc de lever l'ambigüité de l'indéfini le plus fréquent « *chacun* » qui est déterminant devant un substantif précédé ou non d'un déterminant comme par exemple « *chascun jour* », « *Chascune personne* », « *chascun son dueil* », « *chascun aultre vice* » et « *chascune grande feste* ». Dans d'autres séquences lorsqu'il s'emploie devant un verbe « *chacun* » est pronom comme des ces exemples « *chascun sache* », « *chacun veoir* », « *chascun se doit* », « *chacun les puisse* », « *chacun qui auroit* » et « *chacun ne pourra* » et « *chascun ne les puet* ».

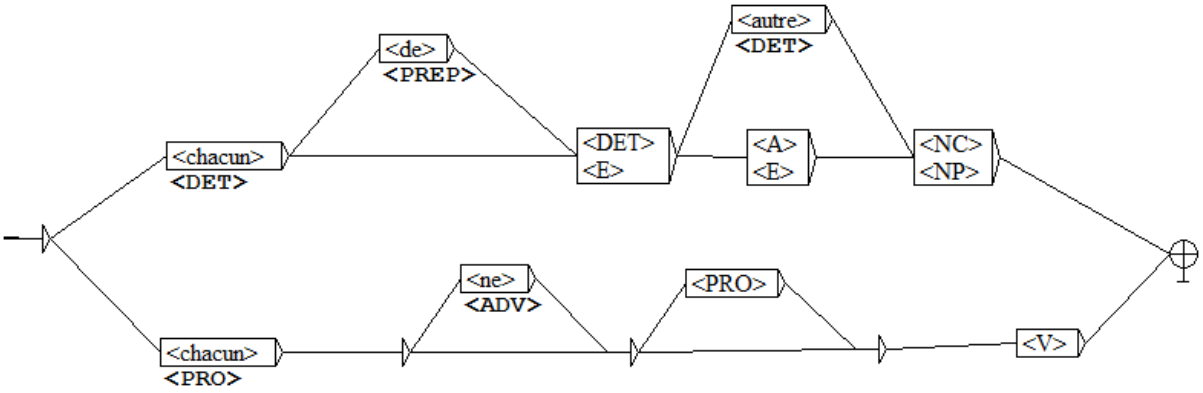


Figure 125. Grammaire de désambiguïsation de « *chacun* »

Comme en français moderne, « *tel* » et « *même* » marquent la similitude. De plus « *tel* » peut aussi marquer l'intensité. Il est bien fréquent et il possède une quarantaine de variantes orthographiques telles que « *tele* » « *tiele* » « *tiell* », « *teu* » et « *itel*». Notre grammaire recense les contextes de son apparition les plus courants comme « *telz enfans* », « *teles plusieurs loenges* », « *telle belle couronne* », « *telles nobles conditions* » et « *telz vos grans amis* » lorsqu'il est employé comme déterminant devant un substantif précédé ou non d'un groupe adjectival. Comme il peut dans ce contexte être accompagné d'une préposition « *en* » ou « *de* » comme dans les exemples « *telx de noz conseillers* » et « *telz de noz conseillers* », du l'adverbe « *comme* » comme dans les séquences « *telles comme ton serviteur* », « *tele comme notre bailli* » et « *telles comme iceux deniers* » ou d'un des déterminants « *aucun* », « *plusieurs* » et « *autre* » à titre d'exemple « *tele aucune foiz* », « *telz aultres saiges* » et « *telles plusieurs illusions* ». Notre grammaire décrit les contextes immédiats de « *tel* » lorsqu'il est pronom, le cas le plus fréquent lorsqu'il s'emploie devant un verbe précédé ou non d'un pronom et/ou l'adverbe de négation « *ne* » comme dans les séquences « *telles voyez* », « *telez sont ennemys* », « *tel me veulent* », « *telz ne desirent* » et « *tel ne se monstre* ». Le pronom « *tel* » peut également être accompagné du pronom relatif « *que* » dans ce cas notre grammaire reconnait des séquences comme « *tel que donne* », « *tel que m'oez* » et « *telle que n'avons* ».

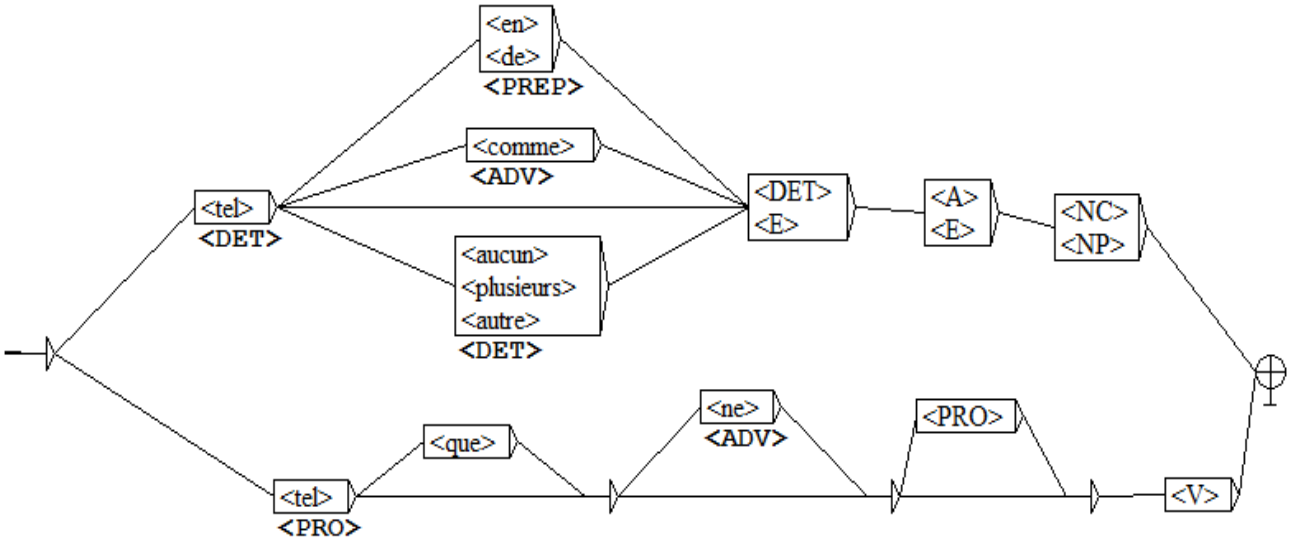


Figure 126. Grammaire de désambiguïsation de « *tel* »

La grammaire de désambiguïsation de « *même* » est moins riche en description des contextes que de « *tel* ». En effet, elle permet d'attribuer à « *même* » la catégorie déterminant « *ADV* » lorsqu'elle s'emploie devant un substantif précédé ou non d'un groupe adjectival permettant ainsi la reconnaissance des séquences comme « *mesme lieu* », « *mesme matiere* », « *meisme le pape* », « *mesmez de commun langage*» et « *mesmes de pire mort* ». Ces séquences peuvent être accompagnées des prépositions « *en* » et « *de* » comme par exemple « *mesmes en mariage* », « *mesme en ce monde* » et « *mesmes de leur bouche* ».

Notre grammaire permet également de désambiguïser « *même* » en l'annotant comme « *pronom* » lorsqu'il est suivi d'un verbe précédé ou non d'un pronom à titre d'exemple « *mesmes plourez* », « *mesme l'envelopa* », «*meismes ne vouroye* », « *mesmes ne la pourroient* » et « *mesmes ne se aiment* ». Ces séquences peuvent être accompagnées d'un pronom relatif à savoir « *que* », « *qui* » et « *quand* » comme le montre ces exemples « *mesmes qui veons* », « *mesmes qui seroit* », « *mesme qui doit* », « *mesmes que batailler* » et « *meismes quant il sera boute* ».

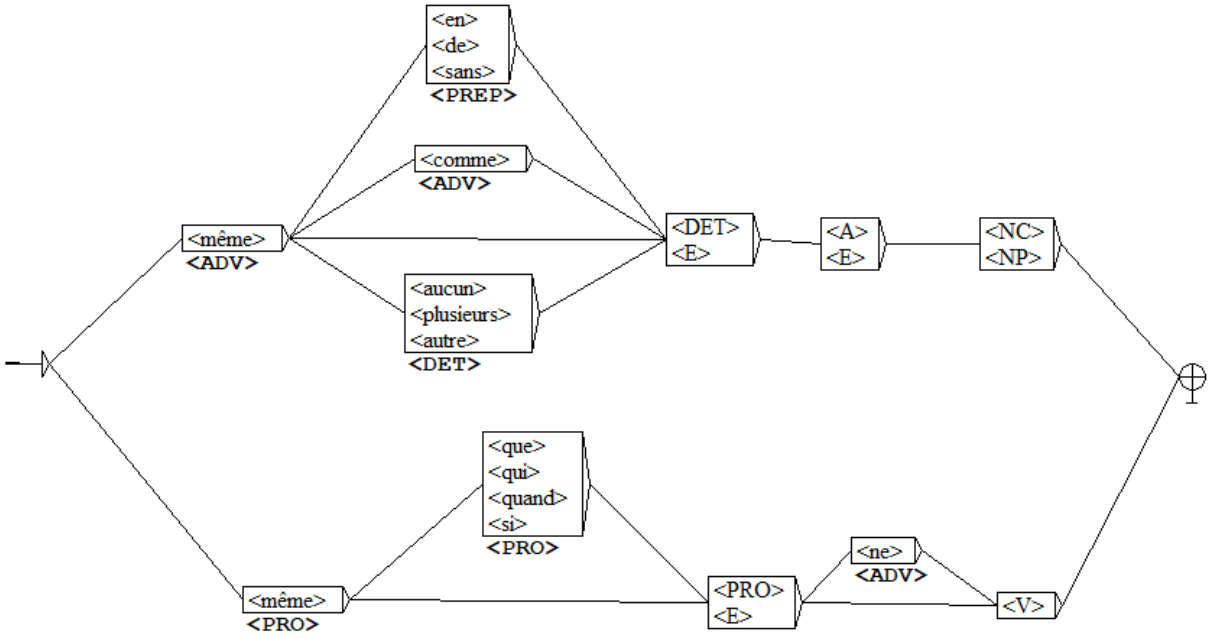


Figure 127. Grammaire de désambiguïsation de « *même* »

4. Evaluation

Afin de garantir de garantir une forte représentativité du corpus de test ou d'évaluation, ce dernier a été constitué à partir des extraits des textes en vers et en prose de différents origines, dates, auteurs et genres. En utilisant, les interfaces fournies par PALM, ce corpus a été annoté semi-automatiquement. Il a été par la suite annoté automatiquement par notre méthode symbolique d'étiquetage morphosyntaxique afin d'évaluer sa performance. Comme le montre le tableau suivant, les métriques d'évaluation affichent un taux de performance (F-Mesure) d'environ 90% avec un rappel de 89% et une précision de 91%. Nous constatons ainsi que notre méthode affiche des performances relativement proches d'un étiqueteur d'une langue standardisée.

Pour mieux évaluer notre système, nous l'avons comparé à un étiqueteur probabiliste « *Treetagger* » que nous avons mis en place à l'aide d'un corpus d'apprentissage constitué de textes extraits du « MEDITEXT » d'environ 200.000 formes. Une comparaison des métriques d'évaluation montre que notre méthode symbolique affiche de meilleurs résultats que l'étiqueteur probabiliste « *Treetagger* », illustrant ainsi son efficacité.

Métriques	**Méthode symbolique**	**Treetagger**

Rappel	89%	82%
Précision	91%	84%
F-Mesure	89,98	82,98

Tableau 13. Evaluation du système d'étiquetage morphosyntaxique

5. Conclusion

Dans ce chapitre, nous avons présenté une méthode symbolique pour l'annotation morphosyntaxique de textes en moyen français. Le processus d'étiquetage morphosyntaxique est composé essentiellement d'une phase de segmentation, d'une phase d'étiquetage a priori et d'une méthode de désambiguïsation. Pour notre méthode symbolique, les phases de la segmentation et de l'étiquetage a priori sont assurées par l'analyse lexicale (décrite au chapitre 5). Par conséquent, plusieurs segmentations d'une séquence de caractères sont proposées auxquelles toutes les étiquettes morphosyntaxiques sont associées. Notre méthode de désambiguïsation repose sur « les grammaires de levée d'ambigüité ». Ces dernières décrivent des contextes « caractéristiques » des formes fréquentes et ambigües. Cette description des contextes est facilitée par une analyse textométrique pour laquelle nous avons utilisé les cooccurrences, les segments répétés et les concordances.

Une série de grammaires locales ont été donc développées permettant ainsi de désambiguïser des mots grammaticaux fréquents ainsi que des formes qui apparaissent dans leurs contextes immédiats. Ces grammaires ont réduit considérablement la taille de la TAS en ne gardant que les annotations pertinentes selon le contexte. En effet, notre méthode affiche de bons résultats relativement proches des méthodes d'étiquetage morphosyntaxique des langues standardisées. L'étiquetage morphosyntaxique a permis de désambiguïser efficacement environ 90% des formes de texte en n'utilisant que des contextes caractéristiques avec un voisinage qui ne dépasse pas quatre ALU. D'où l'utilité d'appliquer notre méthode symbolique avant des processus ultérieurs d'analyses syntaxiques ou sémantiques comme la reconnaissance des entités nommées.

Chapitre 7

Reconnaissance des entités nommées

1. Introduction

La reconnaissance des entités nommées (REN) a été définie par la conférence MUC comme la tâche d'identification et de catégorisation des noms d'entités ENAMEX, des expressions numériques NUMEX et des expressions temporelles TIMEX. Ces entités sont présentes dans tout type de textes, peu importe leurs domaines (Ehrmann, 2008). Leurs reconnaissances sont indispensables pour tout système qui cherche à identifier automatiquement les informations pertinentes contenues dans un texte. Dans un processus TAL, la reconnaissance d'entités nommées est une tâche de l'analyse syntactico-sémantique (Mesfar, 2005) qui concerne des expressions linguistiques au sein des syntagmes nominaux (Ehrmann, 2008). Ces expressions ne contiennent pas de prédications verbales et ne nécessitent pas une analyse syntaxique préalable (Nouvel, 2012).

Il existe peu de travaux sur la reconnaissance d'entités nommées en langues vernaculaires et aucun d'entre eux n'est consacré au moyen français. Par la suite, nous proposons une méthode de REN dite « multiple », mise en œuvre grâce aux diverses ressources linguistiques conçues à l'aide de la plateforme linguistique NooJ.

2. Typologie d'entités nommées

La typologie des entités nommées est la hiérarchie ou l'ontologie sur laquelle notre système se base pour identifier et catégoriser les différentes entités. Plusieurs typologies ont été définies mais la plus utilisée est celle proposée lors du MUC-7. Nous considérons qu'avec quelques modifications, cette dernière pourrait répondre aux spécificités des entités nommées existantes dans des textes en moyen français. En effet, la typologie adoptée organise donc les types sous forme hiérarchique. Cette hiérarchie, composée de deux niveaux, contient 3 types d'expressions :

- ENAMEX : représente des expressions qui réfèrent à des personnes, des lieux et des institutions.
- NUMEX : représente l'ensemble des expressions numériques qui sont constituées des prix décrivant des quantités monétaires, des mesures permettant la

quantification à l'aide des unités de mesures et les expressions de poids contenant les unités de masses et de volumes.

- TIMEX : représente l'ensemble des expressions temporelles qui sont constituées des expressions de dates, des expressions d'horaire donnant des indications sur l'heure, des expressions d'âges décrivant l'âge d'un animé ou d'un évènement et des expressions de mesure de temps permettant l'identification d'une durée.

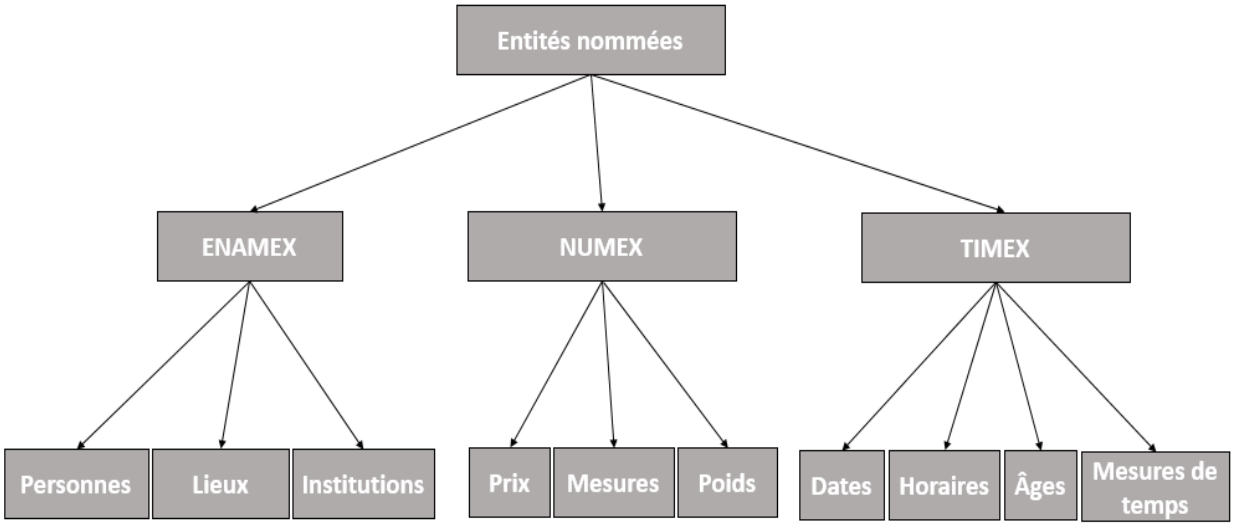


Figure 128. Typologies d'entités nommées

3. Méthode multiple pour la reconnaissance des entités nommées

La méthode proposée pour la reconnaissance des entités nommées se dit « multiple » parce qu'elle décrit des phénomènes de plusieurs niveaux d'analyses linguistiques, depuis l'analyse lexicale jusqu'à l'analyse syntaxique locale. Elle fait appel aux grammaires génératives (Chomsky, 1957) sans faire recours à plusieurs formalismes linguistiques. En effet, l'utilisation d'un formalisme universel et unique tel que *Head-driven Phrase Structure Grammar* (HPSG) (Pollard & Sag, 1994) capable de traiter tous les phénomènes de tous les niveaux linguistiques décrits dans cette méthode s'avère particulièrement inefficace et lourde à mettre en place (Silberztein, 2015). Une autre alternative serait donc d'utiliser plusieurs formalismes. En effet, chaque phénomène linguistique pourrait être décrit par un formalisme tel que *Xerox Finite-State Tools* (XFST) adapté pour l'analyse morphologique et *Lexical Fonctionnal Grammar* (LFG) adapté pour l'analyse syntaxique. Cependant, ces différents formalismes sont incompatibles et la communication entre eux est quasi impossible.

La mise en place d'une méthode « multiple » a été rendue possible grâce à l'utilisation de la notation unifiée de la plateforme NooJ tout en disposant d'outils différents pour traiter des phénomènes des niveaux linguistiques différents (Silberztein, 2015). En effet, NooJ offre la possibilité de développer à l'aide des éditeurs adaptés des ressources linguistiques diverses en utilisant des formalismes unifiés comme les dictionnaires électroniques et les grammaires locales. La même notation est donc utilisée pour tous les types de grammaires de la hiérarchie de Chomsky. Les grammaires sont compilées et appliquées aux textes en employant plusieurs types d'analyseurs qui produisent des résultats sous forme d'annotations rangées dans la structure d'annotation du texte (TAS). Cette dernière constitue le format pivot de communication entre les différents analyseurs de la plateforme NooJ, permettant ainsi l'analyse des phénomènes de niveaux linguistiques différents à savoir morphologique, lexical, syntaxique et sémantique.

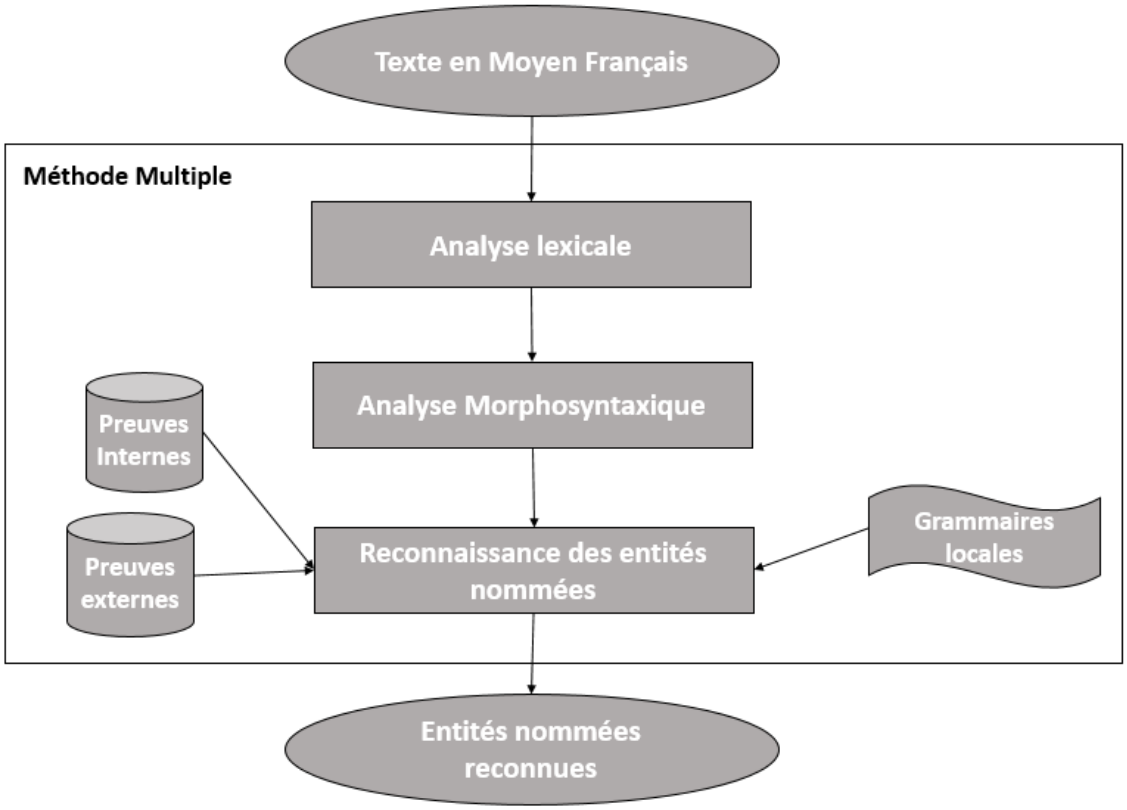


Figure 129. Système de reconnaissance des entités nommées

Comme l'illustre la figure 129, notre système de reconnaissance des entités nommées permet une analyse multi-niveau des textes en moyen français. L'analyse lexicale (décrite au chapitre 5) permet l'identification des ALU d'un texte et leur associée un ensemble d'informations linguistiques telles que leurs parties du discours et leurs lemmes. Ces différentes informations sont sauvegardées dans la TAS et elles sont ensuite utilisées par

l'analyseur morphosyntaxique (décrit au chapitre 6) afin de lever l'ambiguïté morpho-syntaxique. Par conséquent, les annotations jugées non pertinentes sont supprimées de la TAS afin de ne garder que les annotations appropriées aux contextes. Puis, nous faisons appel aux dictionnaires électroniques développés pour s'atteler à la reconnaissance des entités nommées qui permettent d'annoter les différentes ALU avec des traits sémantiques. En effet, ces dictionnaires que nous avons appelé « dictionnaires des entités nommées » permettent l'identification des preuves internes et externes (McDonald, 1996). Rappelons que : les preuves internes sont les éléments du vocabulaire simples ou composés, qui jouent le rôle de marqueurs lexicaux permettant la constitution des entités nommées et les preuves externes sont des éléments des contextes immédiats des entités nommées qui permettent leurs catégorisation. Enfin, toutes ces informations à savoir les descriptions morphosyntaxiques, lemmes et les traits sémantiques, qui constituent les preuves internes et externes, sont utilisées par nos grammaires locales afin de modéliser et de regrouper un ensemble de règles non-contextuelles. Ecrites sous forme de transducteurs à l'aide d'un éditeur graphique de la plateforme NooJ, ces règles décrivent les différentes combinaisons des séquences d'ALU ainsi que leurs contextes immédiats permettant ainsi l'identification et la catégorisation des entités nommées. Par la suite, nous revenons en détails sur les dictionnaires et les grammaires développés permettant la mise en place de notre méthode.

3.1. Dictionnaires des entités nommées

A cause de l'absence d'une orthographe standardisée en moyen français, les noms propres tels que les noms des lieux, les prénoms, les noms de personnes et les noms des institutions et les noms communs, tels que les professions et les titres, sont de nature instable. Bien qu'ils annotent chaque nom avec des traits sémantiques qui aident à l'identification des preuves internes et externes, les dictionnaires des entités nommées permettent d'assurer une étape préalable de normalisation des noms. En effet, ils procèdent à cette étape nécessaire de normalisation en recensant les différentes variantes d'un même nom en lui associant sa forme canonique moderne. Comme le montre le tableau 14, les annotations produites suite à l'application des dictionnaires, seront sauvegardées dans la structure d'annotation du texte TAS dans le but d'être utilisées par les grammaires locales afin de modéliser des séquences d'ALU permettant l'identification et la catégorisation des entités nommées.

Grammaires locales	Dictionnaires des entités nommées
Personnes	Prénoms, Patronymes, métiers, titres, noms de lieux, noms composés de personnes
Lieux	Noms de lieux, titres, Prénoms, Patronymes, métiers, titres, noms composés de personnes
Institutions	Noms des institutions, noms de lieux, titres, métiers

Tableau 14. Correspondances entre grammaires locales et dictionnaires des entités nommées

3.1.1. Dictionnaire des prénoms

Bien qu'étymologiquement « prénom » signifie « avant le nom », jusqu'à la fin du Moyen Âge, le prénom est utilisé comme nom principal identifiant une personne d'une façon exclusive. En effet, les premières traces écrites de l'apparition des patronymes remontent au XIIème siècle. Les prénoms français au Moyen Âge proviennent principalement de l'Antiquité, de la Bible et de l'étranger. Selon la période, nous constatons qu'il y a des prénoms utilisés plus fréquemment que d'autres. Comme tous les noms propres du Moyen Âge, les prénoms n'étaient pas stables. Ils subissent une évolution morphologique et leurs orthographes ne sont pas standardisées. Il existe plusieurs variantes pour chaque prénom comme le montre l'exemple suivant concernant le prénom « *Paul* ». Notre dictionnaire des prénoms contient environ 3000 entrées. Chaque entrée se compose du prénom tel qu'il apparaît dans le texte, de la forme standardisée du prénom, d'une partie du discours à savoir nom propre « *NP* » et d'un trait sémantique « *Prénom* ».

paul,NP+ Prénom
paol,paul,NP+ Prénom
paolo,paul,NP+ Prénom
paoul,paul,NP+ Prénom
paullus,paul,NP+ Prénom
paulus,paul,NP+ Prénom
pol,paul,NP+ Prénom
pollus,paul,NP+ Prénom
pols,paul,NP+ Prénom
poul,paul,NP+ Prénom

3.1.2. Dictionnaire des patronymes

Avec la croissance démographique de la population au Moyen Âge, plusieurs personnes portent le même prénom, par conséquent le prénom ne suffit plus à identifier une personne. On associe au prénom, un patronyme qui est issu d'un ou plusieurs surnoms individuels

propres à chaque personne. Notre dictionnaire contient environ 2000 surnoms qui peuvent faire référence à un métier, une origine géographique, un lieu de sa résidence, un nom du chef de famille ou à un sobriquet qui est généralement une singularité liée au physique ou au caractère.

Chaque entrée de notre dictionnaire des patronymes se compose de la forme standard du surnom, de ses variantes orthographiques, d'une partie du discours « *NP* » qui désigne un nom propre et d'un trait sémantique « *Nom* ». A titre d'exemple, nous donnons quelques variantes du surnom « *marchand* » :

marchand, NP+Nom
maarchaund,marchand,NP+Nom
marcaans,marchand,NP+Nom
marcande,marchand,NP+Nom
marcans,marchand,NP+Nom
marcaunz,marchand,NP+Nom
marceans,marchand,NP+Nom
marcens,marchand,NP+Nom
marchaant,marchand,NP+Nom

3.1.3. Dictionnaire des métiers

A partir de l'exploration des textes de notre corpus et du livre des métiers d'Etienne Boileau (Boileau, 1879), recueil de l'ensemble des métiers existant à Paris au XIIIe siècle, nous avons recensé une liste des métiers afin de construire un dictionnaire spécialisé. Notre dictionnaire contient environ 110 métiers et occupations qui se pratiquaient au Moyen Âge. Chaque entrée contient tous les paradigmes, les variantes, et des informations linguistiques, à savoir la partie du discours « *NC* » qui désigne un nom commun et le trait sémantique « *Profession* », comme le montre l'exemple ci-dessous de l'entrée « *paysan*».

paysan,NC+Profession
paisans,paysan,NC+Profession
paisant,paysan,NC+Profession
paissant,paysan,NC+Profession
paissauns,paysan,NC+Profession
paysans,paysan,NC+Profession
paysant,paysan,NC+Profession
paysanz,paysan,NC+Profession
païsant,paysan,NC+Profession
païssant,paysan,NC+Profession
peisanz,paysan,NC+Profession
peissant,paysan,NC+Profession
peissaunt,paysan,NC+Profession
peisuns,paysan,NC+Profession
pesant,paysan,NC+Profession
pesaunt,paysan,NC+Profession

3.1.4. Dictionnaires des titres

Notre dictionnaire des titres contient environ 60 titres parmi lesquels les titres de noblesses, correspondant à l'exercice de fonctions juridiques et/ou militaire attachées à un territoire donné, des titres religieux et des titres attribués à des officiers au service du souverain, des nobles ou des religieux. Comme le montre l'exemple ci-dessous de l'entrée « *empereur*», chaque entrée est composée d'une liste des variantes orthographique (environ 30 variantes pour l'entrée « *empereur* »), des paradigmes, d'une partie du discours « *NC* » qui désigne un nom commun et d'un trait sémantique « *Titre* ».

empereur,NC+Titre
amperere,empereur,NC+Titre
amperour,empereur,NC+Titre
amperur,empereur,NC+Titre
aumperur,empereur,NC+Titre
aunpirere,empereur,NC+Titre
empareur,empereur,NC+Titre
emperadour,empereur,NC+Titre
emperaire,empereur,NC+Titre
empereere,empereur,NC+Titre
empereior,empereur,NC+Titre
empereire,empereur,NC+Titre
empereor,empereur,NC+Titre
empereour,empereur,NC+Titre
emperer,empereur,NC+Titre
emperere,empereur,NC+Titre
empereres,empereur,NC+Titre
empereris,empereur,NC+Titre
empererz,empereur,NC+Titre
emperesses,empereur,NC+Titre
empereurs,empereur,NC+Titre
emperiere,empereur,NC+Titre

3.1.5. Dictionnaire des noms composés de personnes

Nous avons recensé une liste d'environ 300 mots composés qui font référence à des personnages décrits par les textes de notre corpus. Ce dictionnaire permet d'identifier ces personnes systématiquement et de lever l'ambiguïté grâce au modificateur « UNAMB » prédéfinie par la plateforme NooJ.

guillaume de castillon,NP+Personne+UNAMB
guillaume de dormans,NP+Personne+UNAMB
guillaume de gennes,NP+Personne+UNAMB
guillaume de graville,NP+Personne+UNAMB
guillaume de meleun,NP+Personne+UNAMB
guillaume de normandie,NP+Personne+UNAMB
guillaume le roux,NP+Personne+UNAMB

guillaume de melle,NP+Personne+UNAMB
guillaume le conquérant,NP+Personne+UNAMB

3.1.6. Dictionnaire des noms de lieux

Les noms de lieux ont plusieurs origines. Nous constatons la présence des noms de lieux d'origine grecque, gallo-romaine, germanique, franque et celtique. Les noms de lieux venus de l'antiquité décrivent généralement une des caractéristiques de la terre. Mais la grande majorité des noms de lieux en moyen français doivent leurs origines, soit aux noms religieux, tels que les noms des instituts religieux, et aux noms des saints, soit à la vie féodale avec ses différents aspects et surtout aux noms des familles nobles qui ont passé leurs noms aux terres sur lesquelles elles exerçaient leurs pouvoirs.

L'orthographe des noms de lieux en moyen français n'est pas standardisée, nous remarquons la présence de plusieurs variantes orthographiques possibles pour le même nom, comme l'illustre l'exemple ci-dessous de l'entrée « *italie* ».

Nous avons construit un dictionnaire contenant environ 1000 noms de lieux tels que les noms de pays, de villes et des domaines. Chaque entrée de ce dictionnaire est composée d'une liste des variantes orthographiques, d'une partie de discours « *NP* » et d'un trait sémantique « *Lieu* ».

italie,NP+Lieu
itaile,italie,NP+Lieu
itaille,italie,NP+Lieu
itallie,italie,NP+Lieu
itayle,italie,NP+Lieu
itaylle,italie,NP+Lieu
ytaile,italie,NP+Lieu
ytaille,italie,NP+Lieu
ytalie,italie,NP+Lieu
ytalye,italie,NP+Lieu

3.1.7. Dictionnaire des institutions

En explorant les textes de notre corpus en utilisant des méthodes d'analyse textométrique, nous avons recensé une liste des noms de divers instituts religieux, tels que ceux des églises et les abbayes, des instituts politiques, tels que les châteaux des rois, et des instituts sociaux, tels que les hôpitaux et les universités.

Comme le montrent les exemples ci-dessous, chaque entrée de notre dictionnaire est constituée du type d'institut suivi du nom donné à l'institut, d'une partie du discours (toujours un nom propre), d'un trait sémantique « Organisation » et du modificateur « UNAMB »

proposé par NooJ pour désambigüiser la séquence des ALU en question en lui attribuant l'annotation du dictionnaire.

église saint-benoît,NP+Organisation+UNAMB
abbaye de marcheroux,NP+Organisation+UNAMB
abbaye saint-nicolas de marcheroux,NP+Organisation+UNAMB
abbaye marchiennes,NP+Organisation+UNAMB
abbaye notre-dame de molosmes,NP+Organisation+UNAMB
château de Montgeoffroy,NP+Organisation+UNAMB
abbaye de montcetz,NP+Organisation+UNAMB
abbaye de montiers,NP+Organisation+UNAMB
abbaye de moreaucourt,NP+Organisation+UNAMB
abbaye de morimond,NP+Organisation+UNAMB
abbaye de moyenmoutier,NP+Organisation+UNAMB
abbaye de mozac,NP+Organisation+UNAMB
abbaye de mureau,NP+Organisation+UNAMB
château de ménars,NP+Organisation+UNAMB
abbaye de noirlac,NP+Organisation+UNAMB

3.2. Grammaires locales pour la reconnaissance des entités nommées

Les grammaires locales sont des grammaires hors contextes qui servent à localiser des phénomènes locaux dans le texte et à attester leur appartenance à des classes telles que « les entités nommées ». En effet, une grammaire locale est équivalente à un réseau de transition récursif RTNs qui consiste à définir un ensemble d'états et un ensemble de transitions d'un état à un autre en traitant chaque état non-terminal comme un éventuel appel à d'autres réseaux. Ces réseaux permettent ainsi la reconnaissance d'un ensemble fini de séquences décrit par un langage.

Dans la notation NooJ, les grammaires locales sont modélisées sous forme de graphes. Ces graphes contiennent un ensemble de nœuds incluant un nœud initial, un nœud terminal et des nœuds intermédiaires qui peuvent appeler des sous-graphes indépendants ou qui représentent des ensembles d'occurrences du langage naturel en faisant appel à des dictionnaires de mots simples et composés. Une telle structure offre la possibilité de formaliser plusieurs phénomènes linguistiques d'un même niveau d'analyse, ou de niveaux d'analyses différents, et de les imbriquer au sein d'un même graphe. Elle donne ainsi l'avantage de pouvoir formaliser, selon le contexte, des concepts et des phénomènes linguistiques complexes d'une façon indépendante, de produire des annotations selon le niveau d'analyse et de faire une analyse transformationnelle, tout en utilisant une architecture qui facilite la maintenance et la réutilisation des sous-graphes.

Les analyseurs de NooJ compilent le graphe principal en faisant appel aux sous-graphes et aux différents dictionnaires utilisés par les grammaires. Ensuite, ils ajoutent les résultats du

processus de compilation à la structure d'annotations de texte TAS, afin de localiser et de catégoriser d'une façon pertinente et très précise des phénomènes locaux dans le texte.

3.2.1. Grammaire de reconnaissance des personnes

Nous présentons la grammaire locale qui permet d'identifier des expressions accompagnées ou non d'un nom propre référant à des personnes. Cette grammaire lexicalisée fait appel aux dictionnaires des entités nommées présentés dans la section 3.1 plus précisément le dictionnaire des prénoms, le dictionnaire des patronymes, le dictionnaire des métiers, le dictionnaire des titres et le dictionnaire des noms composés de personnes. Ces dictionnaires ajoutent à la TAS des annotations sémantiques qui peuvent être des preuves internes et externes. Notre grammaire locale utilise donc ces annotations ainsi que les descriptions morphosyntaxiques et les lemmes afin de modéliser les différentes séquences d'ALU ainsi que leurs contextes immédiats dans le but de discerner les entités nommées de type personne. Elle est constituée de quatre sous-graphes principaux. Chaque sous-graphe modélise un ensemble d'expressions qui peuvent être regroupé selon leurs preuves internes et externes. A titre d'exemple, le sous-graphe « *par_métiers* » modélise toutes les séquences contenants les métiers et les professions, et le sous-graphe « *par_titres* » modélise toutes les expressions contenant des titres. Les séquences reconnues par les quatre sous-graphes sont annotées par l'étiquette <*ENAMEX+PERS*>, produite par notre grammaire. En effet, dans le graphe principal, l'étiquette <*ENAMEX+PERS*> permet d'annoter toutes les séquences repérées par les trois sous-graphes « *par_métiers* », « *par_nomEtprénom* » et « *par_titres* ». L'annotation des séquences reconnues par le sous-graphe « *par_verbes* » sera détaillée lors de la description de ce dernier.

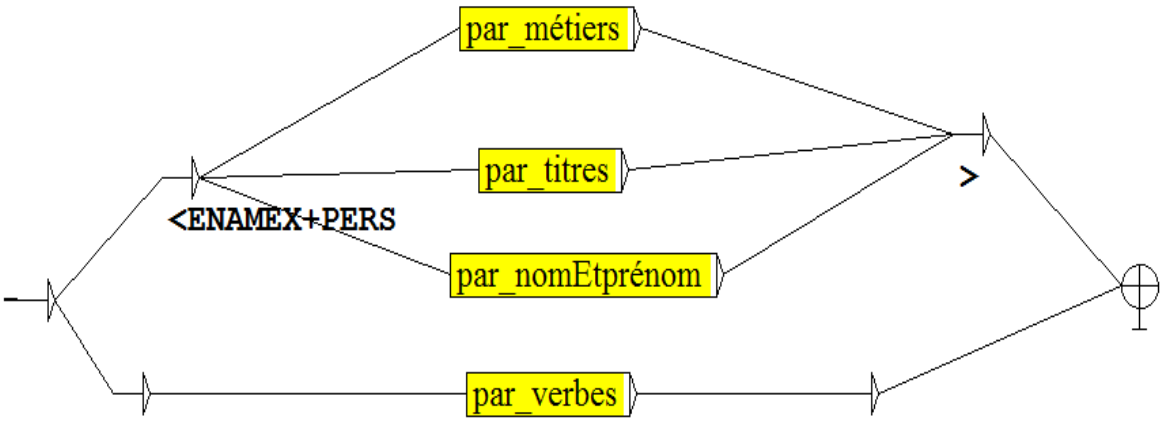


Figure 130. Grammaire locale de reconnaissance des personnes

Ces quatre sous-graphes appelés par le graphe principal font appel à plusieurs autres sous-graphes à savoir :

- Le sous-graphe « *liens_de_parenté* » permet de modéliser une preuve interne à savoir les liens de parentés qui existent entre les individus. Il permet ainsi de reconnaître des séquences comme « *le fils de juge Edward* ».

- Le sous-graphe « *institutions* » permet de recenser des types d'institutions du Moyen Âge qui sont essentiellement utilisées comme preuves internes permettant la reconnaissance des séquences qui font référence à des personnes comme « *le conseiller du* ***palais*** ».

- Le sous-graphe « *nomEtprénom* », présenté à la figure 131, permet la reconnaissance des noms propres des personnes tels « *John Chambon* ».

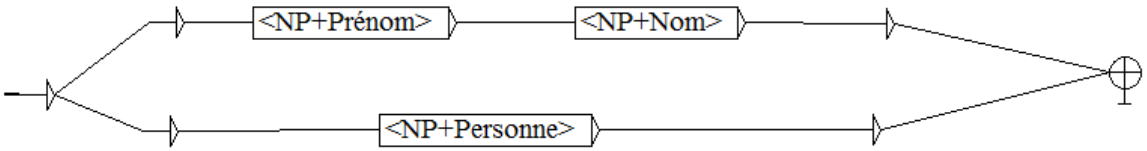


Figure 131. Sous-graphe pour la reconnaissance des noms propres de personne

- Comme l'illustre la figure 132, le sous-graphe « *lieux* » permet de recenser une liste de lieux qui pourrait faire partie d'une séquence d'ALU faisant référence à une personne.

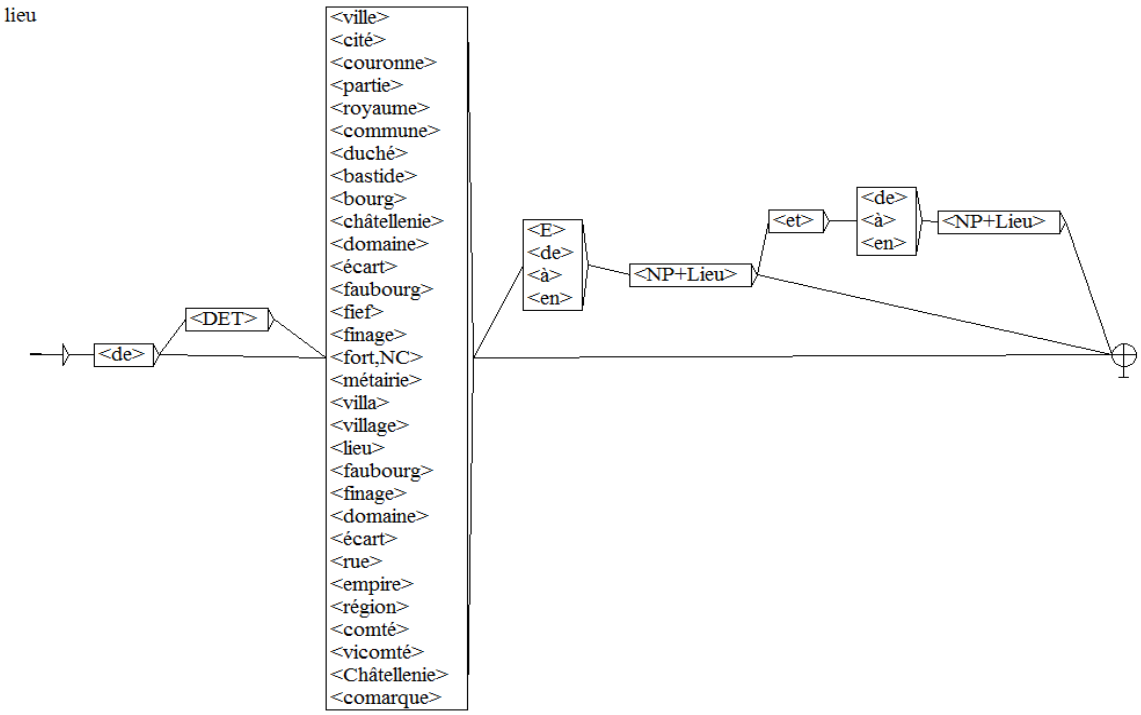


Figure 132. Sous-graphe « *lieux* »

La figure 133 montre le sous-graphe « *par_métiers* » qui permet d'identifier plusieurs séquences d'ALU contenant des mots déclencheurs, de métier ou de profession, et de les catégoriser en entités nommées de type « personne ». Cette grammaire contient plusieurs règles syntactico-sémantiques. Certaines de ces règles sont simples et peuvent être modélisées à l'aide d'une simple expression rationnelle comme les séquences « *docteur Augustin* » ou « *conseillier Josselin du boys* » constituées d'un métier suivi d'un patronyme. Cependant, si ces mêmes séquences sont précédées par un lien de parenté pour reconnaître des séquences comme par exemple «*neveu du prevost Edward* », les expressions rationnelles ne suffisent plus car il faut faire appel au sous graphe « *liens_de_parenté* ». Les règles deviennent plus complexes lorsqu'elles combinent des séquences reconnues par les sous-graphes « *liens_de_parenté*», « *nomEtprénom* » et « *institutions* » et des ALU ayant des traits sémantiques comme « *Profession* », « *Titre* » et « *Lieu* ». A titre d'exemple, une règle composée d'un métier suivi d'une préposition suivie d'un titre suivi de la préposition « *de* » suivie d'un nom de personne reconnu par le sous graphe « *nomEtprénom* » est modélisé par cette grammaire et peut reconnaitre des expressions comme *« procureur de Saint Marcellin »*. Un deuxième exemple est celui d'une règle composée d'un métier suivi d'une préposition, d'un titre, d'une deuxième préposition, d'une institution reconnue par le graphe « *institutions* », d'une troisième préposition et d'un nom de lieu qui reconnaît plusieurs séquences comme «*docteur du grant maistre d'ostel de France* ».

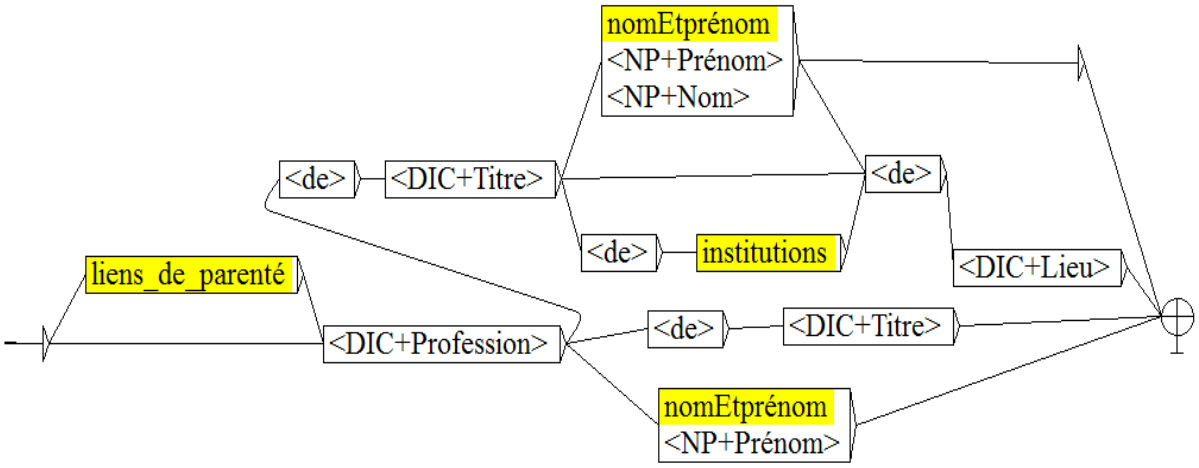


Figure 133. Sous-graphe pour la reconnaissance des personnes par leur profession

De même que pour les métiers, un sous-graphe « *par_titres* », représenté dans la figure 134, a été élaboré. Puisque nous travaillons sur un corpus constitué essentiellement des textes politiques, le nombre de séquences, qui font référence à des personnes, contenant des titres est très important. Ce sous-graphe permet la reconnaissance de plus de 40% des séquences

reconnues par le graphe général. En plus des sous-graphes « *liens_de_parenté*», « *nomEtprénom* » et « *institutions* » et les nombreuses annotations sémantiques produites par les dictionnaires des entités nommées, le sous-graphe « *lieux* », qui permet la reconnaissance des séquences d'ALU constituées essentiellement d'un type de lieu suivi d'un nom de lieu, a été utilisé. L'utilisation de ces nombreuses ressources a permis la modélisation d'une centaine de règles de reconnaissance. Cette grammaire permet, par exemple, d'extraire toute les séquences constituées d'un titre, précédé ou non d'un lien de parenté, suivi d'un nom de personne comme « *sainct Jehan Baptiste*» ou « *pere sainct Augustin*», d'un titre précédé, ou non, par un lien de parenté suivi d'une préposition, suivie d'un nom de lieu *« seigneur de Neufchastel*» ou « *frere du conte de Bloys* ». D'autres règles, qui ont des expressions rationnelles plus simples, ont été écrites c'est le cas de *<DIC+Titre> <NP+Prénom> <de> <DIC+Lieu>*. Elle permet la reconnaissance des séquences comme « *Monseigneur Phelippe de Bourgoingne* » ou encore cette règle qui a l'expression rationnelle *<DIC+Titre> <le> <A> <DIC+Titre> <NP+Prénom>* qui permet de capturer des motifs comme « *Seignur le bon Counte Dampmartin »*. D'autres séquences reconnues nécessitent une modélisation plus sophistiquée qui fait appel à plusieurs sous-graphes et plusieurs traits sémantiques tels que « *bourgeois de nostre ville de Poictiers* », « *roys du royaume de Gaule* », « *saincte Trinité de Vendosme* », « *seigneurs du royaume de France* », « *son frere monseigneur Philippe de Navarre* » et « *son oncle empereur d'Alemaingne* ».

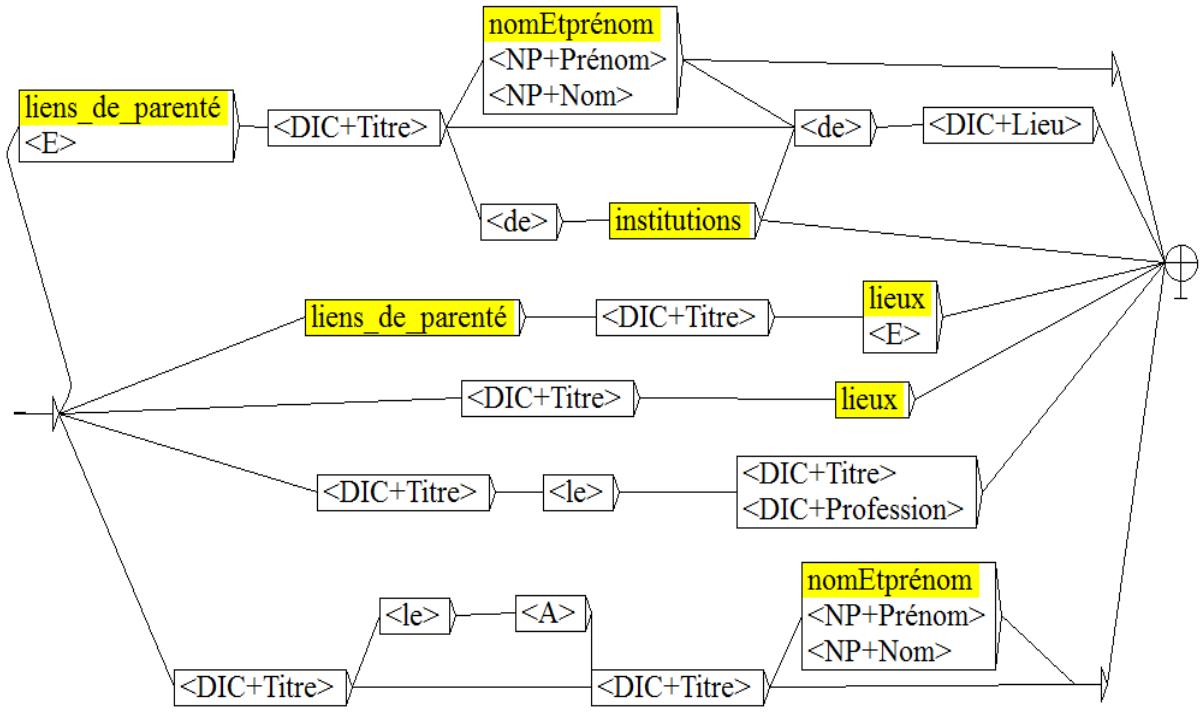


Figure 134. Sous-graphe « *par_titres* »

Les prénoms et les patronymes sont recensés par les dictionnaires des entités nommées et ils sont associés respectivement aux traits sémantiques « *Prénom* » et « *Nom* ». Comme le montre la figure 135, ils sont utilisés au sein du sous-graphe « *par_nomEtprénom*» pour repérer et catégoriser des entités nommées de type « Personne ». Bien que cette grammaire ne contienne pas plusieurs règles, elle permet la reconnaissance d'un nombre important de séquences. Elle donne la possibilité d'extraire principalement les noms de personne comme « *Estienne Marcel*» qui pourrait, ou non, être précédé par un lien de parenté comme « *filz de Mathieu Bassal*» ou « *filz Symon Pourcelet*». Cette grammaire contient des règles complexes permettant la reconnaissance de n'importe quelle séquence constituée d'un nom de personne composé d'un prénom et d'un patronyme ou d'un simple prénom précédé ou non par un lien de parenté reconnu par le sous graphe « *liens_de_parenté*» et suivi par un nom de lieu comme « *Denis de Secile*» ou par un adjectif comme « *Job le juste* » ou « *Alexandre le grant* ».

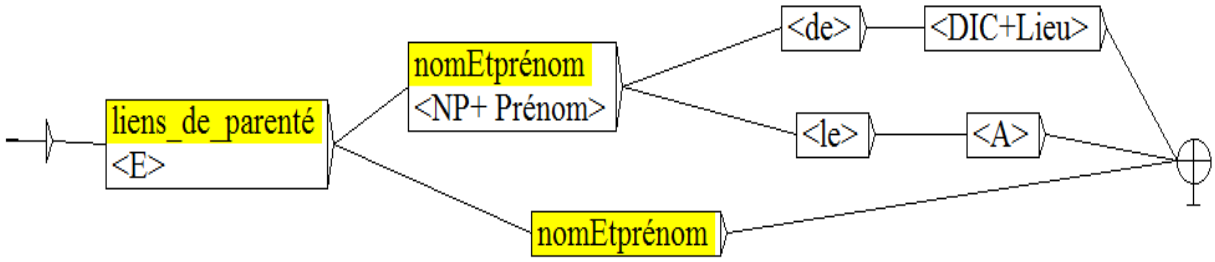


Figure 135. Sous-graphe « *par_nomEtprénom* »

Notre graphe principal contient un quatrième sous-graphe « *par_verbes* ». En réalité, ce sous-graphe est constitué des règles syntactico-sémantiques que nous n'avions pas pu insérer dans les trois sous-graphes « *par_métiers* », « *par_titres* » et « *nomEtprénom* ». Cette grammaire utilise des verbes comme mots déclencheurs comme « *appeler* », « *nommer* » et « *surnommer* ». Ces verbes peuvent être suivis d'un patronyme comme « *disoit Platon* », d'un prénom comme « *nommé Adam* » ou d'un prénom suivi d'un patronyme comme « *disoit Judas Machabeus* ». Ces expressions peuvent contenir un titre ou un nom de métier après le verbe et avant le prénom et/ou le patronyme comme « *appelle maistre Gile Haneboit* », « *nommé pape Clement* » et « *dit Roi Edward* ». Cette grammaire, qui contient un petit nombre de règle, reconnaît plus de deux milles séquences pertinentes. Seuls les titres, les prénoms, les noms ou les professions, qui constituent les séquences retrouvées par la grammaire, sont annotés par l'étiquette <*ENAMEX+PERS*>.

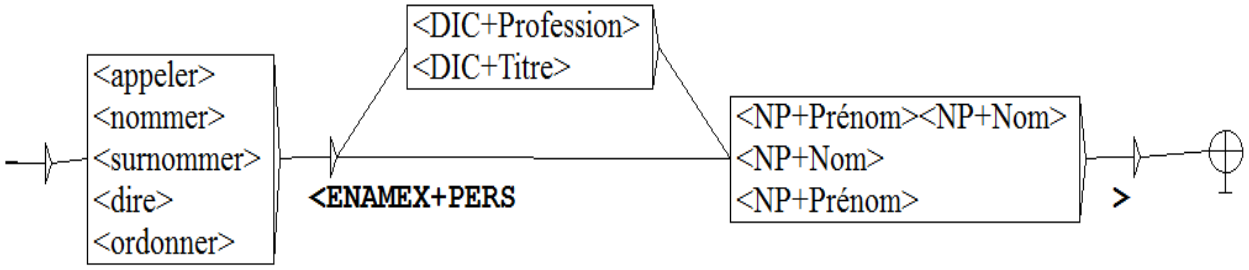


Figure 136. Sous-graphe « par_verbes »

3.2.2. Grammaire de reconnaissance des lieux

L'entité nommée « lieu » désigne principalement les toponymes administratifs comme les pays, les villes, les villages et les toponymes géographiques comme les montagnes, les fleuves. Notre grammaire, qui permet la reconnaissance des lieux existant dans MEDITEXT, distingue trois types d'expressions pour désigner un lieu : celles contenant un nom propre de lieu comme preuves internes sont formalisées à travers le sous-graphe « *par_location* », les expressions contenant un nom de personne comme preuves internes sont représentées dans le sous-graphe « *par_personnes* » et enfin celles reconnues grâce à leurs contextes immédiats sont décrites dans le sous-graphe « *par_déclencheurs* ».

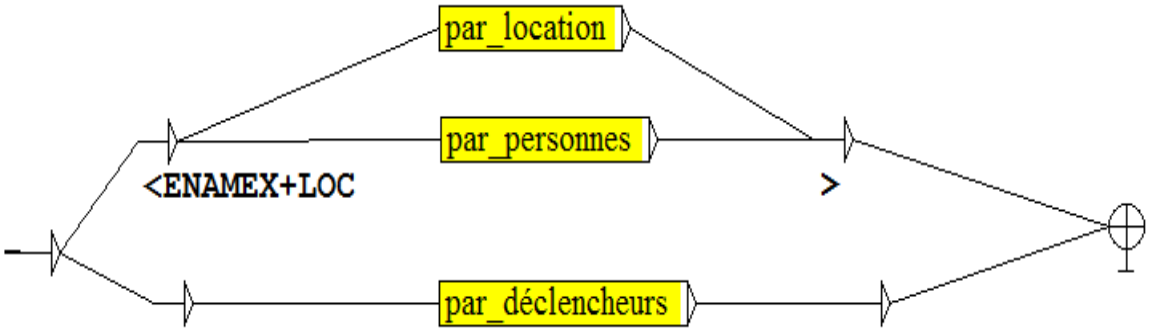


Figure 137. Grammaire de reconnaissance des entités nommées de type « lieu »

Les trois sous-graphes qui constituent le graphe global et qui seront détaillés ultérieurement, font appels à des graphes dont nous présentons les fonctions :

- Le sous-graphe « *type_localisation* » reconnait tous les mots déclencheurs qui désignent le type de localisation comme « *cité* », « *ville* » et « *région* ». Ils sont utilisés comme des preuves internes et ils peuvent être suivis ou non d'un nom de lieu.

Figure 138. Sous-graphe « *type_localisation* »

- Le sous-graphe « *institutions* » identifie une liste de type d'institution comme « *université* », « *église* », « *château* » utilisés comme preuves externes pour décrire le contexte gauche des expressions qui désignent une location.
- Le sous-graphe « *description* » sert à reconnaitre les séquences qui décrivent un lieu via des adjectifs et des adverbes. Les séquences reconnues par ce graphe constituent des preuves internes des expressions référant à un « Lieu » comme « *plusieurs belles villes de France* », « *notre belle ville* » et « *un grand nombre de ville du royaume* ».
- Le sous-graphe « *nomEtprénom* » reconnaît les noms propres de personnes. Les séquences reconnues par cette grammaire peuvent être utilisées comme preuves internes ou externes, selon le contexte.

Par la suite, nous représentons les trois sous-graphes qui constituent le graphe global, afin de décrire les différentes règles de reconnaissance.

Le sous-graphe « *par_location* » formalise donc toutes les règles qui permettent d'identifier des séquences contenant un nom de lieu comme preuve interne. On distingue des règles qui permettent la reconnaissance d'un nom de lieu suivi ou non d'un adjectif tels que « *le Havre* » ou « *notre belle Paris* ». D'autres règles identifient des séquences constituées

d'un marqueur lexical reconnu par le sous-graphe « *type_localisation* » qui désigne le type de lieu suivi d'un nom de lieu comme « *ville de Poictiers* » ou « *cité d'Amiens*». Ces dernières peuvent aussi contenir un adjectif, par exemple « *bonne ville de Compiengne* ».

Des séquences constituées de plusieurs noms de lieu ont été formalisées comme « *Saint-Ouen de Rouen* » ou « *Napples à Romme* » ou «*la ville de Clermont en Auvergne*». Notre sous-graphe contient également des règles sophistiquées qui combinent des séquences reconnues par le sous-graphe « *description* », des séquences repérées par le sous-graphe « *type_localisation* » et des ALU ayant le trait sémantique « Lieu », afin d'identifier des séquences comme « *les parties de Guyen* », « *pluseurs grosses villes de Picardie*» ou « *pluseurs des communes de Tournay* ».

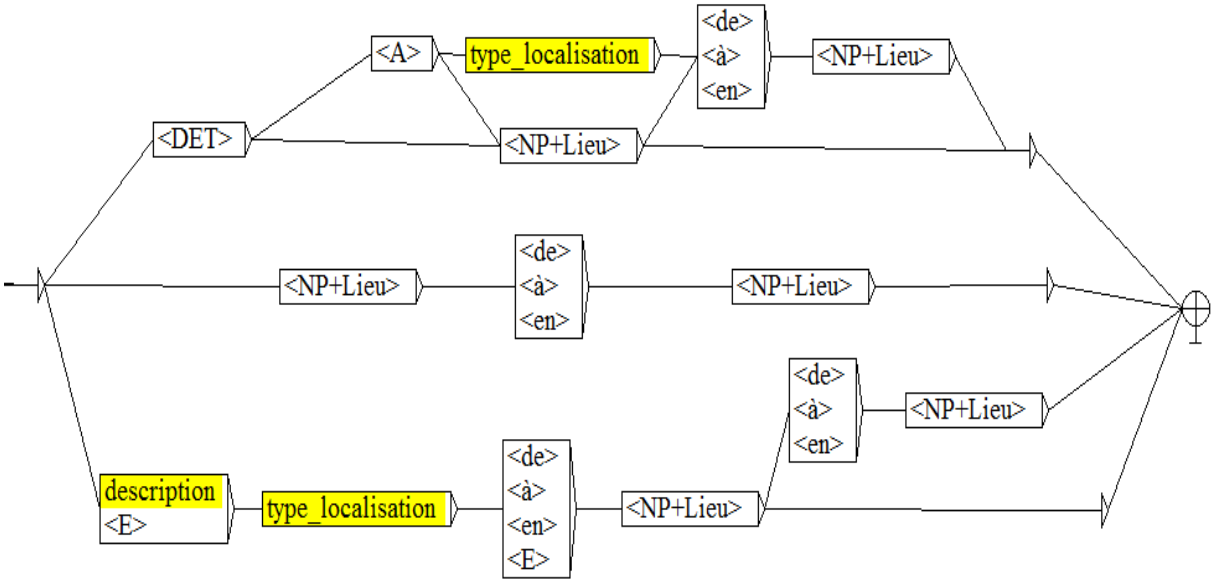


Figure 139. Sous-graphe « *par_location* »

Les expressions, référant à un lieu, constituées d'un type de lieu accompagné d'un nom de personne sont très fréquemment utilisées dans MEDITEXT. La reconnaissance de ce type d'expressions est assurée par notre sous-graphe « *par_personnes* ». Il contient un ensemble de règles permettant la reconnaissance des séquences constituées principalement d'un type de lieu, précédé ou non par une séquence reconnue par le sous-graphe « *description* » qui décrit le lieu concerné, et un nom de personne ou un titre ou une profession telle que « *la partie du roy d'Angleterre* » ou « *la grand terre du roy de France*». D'autres règles sont développées pour capturer des expressions plus complexes. C'est le cas d'une règle ayant une expression rationnelle constituée d'un type de lieu reconnu par le sous-graphe déclencheur précédé ou non d'une description et suivi d'un lieu, lui-même suivi par un nom de personne. De telles règles reconnaissent des expressions comme « *Domain en France de seignur Bertrand* ».

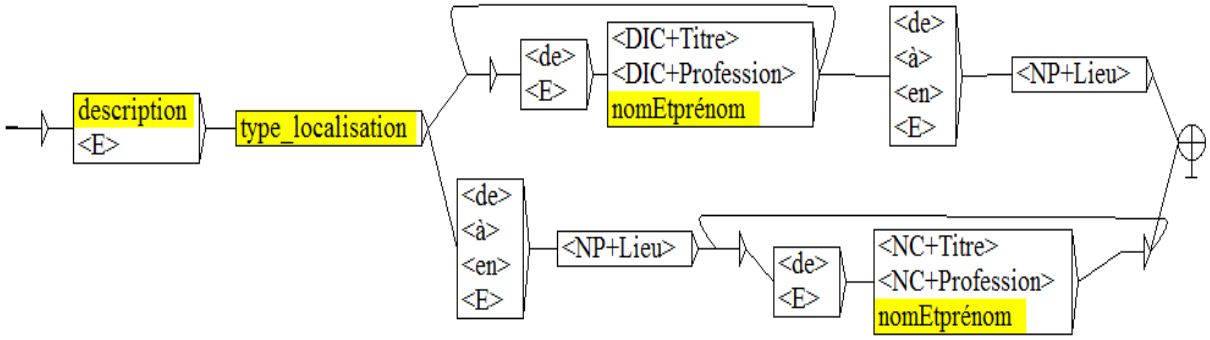


Figure 140. Sous-graphe « *par_personnes* »

Le troisième sous-graphe compris dans notre graphe global est le sous-graphe « *par_déclencheurs* » qui reconnait les noms de lieu annoté par le trait sémantique « Lieu » ou les types de location reconnus par le sous-graphe « *type_localisation* ». Ces différentes ALU sont catégorisés à l'aide de la description de leur contexte gauche. En effet, ces contextes immdiats sont modélisés en utilisant des preuves externes comme les titres, les professions et les noms propres d'une personne ainsi que des verbes comme « *venir* », « *aller* » et « *nommer* », des noms communs comme « *peuple* » et « *voyage* » et les séquences repérées par le sous graphe « *institutions* ».

Le sous-graphe « *par_déclencheurs* » permet d'identifier des séquences comme « *duché de Normendie* », « *empereur d'Alemaingne* », « *eglise dudit Chateaubriant* » et « *envoya a Cartage*». Mais seulement les noms et les types de lieu seront annotés comme des locations par l'étiquette *<ENAMEX+LOC>*.

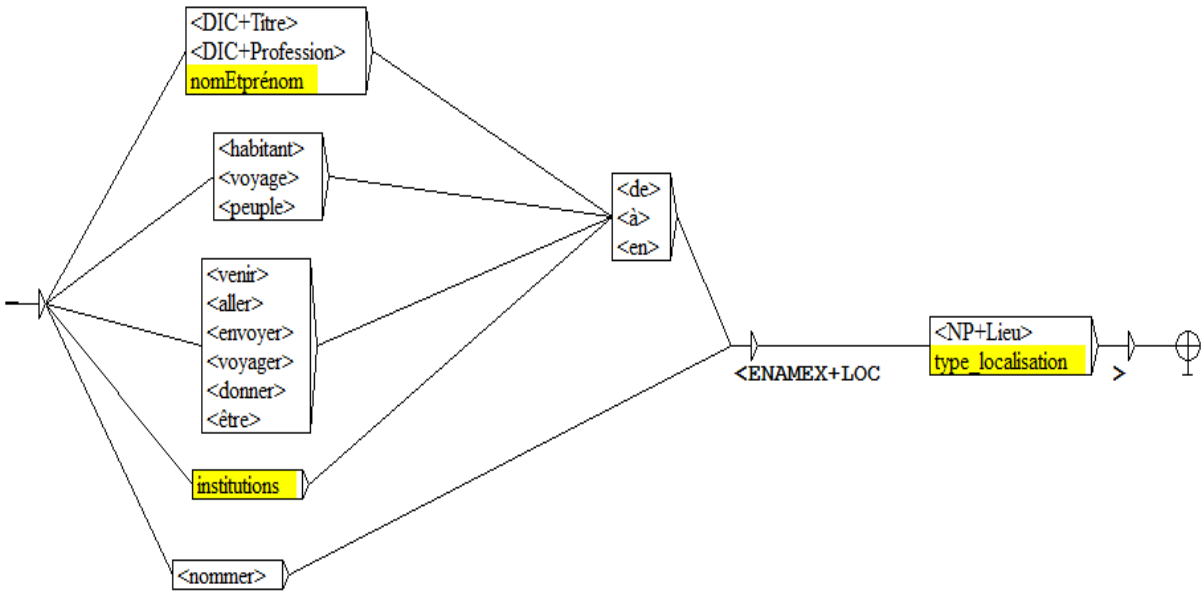


Figure 141. Sous-graphe « *par_déclencheurs* »

3.2.3. Grammaire de reconnaissance des institutions

L'entité nommée « institutions » couvre les noms des organisations de nature publique, coutumière ou privée qui permettent de gouverner et d'administrer le royaume. Nous avons développé un graphe principal capable de reconnaître l'ensemble des institutions royales, seigneuriales ou provençales dans tous les secteurs religieux, politique, universitaire et santé. L'entité nommée « institution » est composée d'un type d'institution et d'un nom d'institution. Les différents types d'institutions sont identifiés par notre sous-graphe « *type_institutiton* » dans lequel nous avons recensé les types d'organisations tels que « *université* », « *palais* », « *hôpital* », « *église* » et « *cathédrale* ». Les noms d'institutions sont multiples et divers. Ils peuvent faire référence à une personne par son nom, par son titre ou par son métier, ainsi « *eglise dudit maistre René* », « *monastere de Monseigneur Saint Denis* » ou « *ostel du procureur* ». Ils peuvent aussi prendre des noms des lieux tels que « *l'église d'Angiers* », « *Parlement de Gloucestre* » ou « *notre chastellet de paris* ». A partir du MEDITEXT, nous avons développé un dictionnaire des noms d'institutions existant ayant le trait sémantique « Organisation », et un deuxième dictionnaire dont chaque entrée est constituée d'une séquence composée d'un type et d'un nom d'organisation, ayant le trait sémantique « Organisation+Attestée ». Ces deux dictionnaires nous permettent de reconnaître des séquences diverses comme « *abbaye notre dame de sénanque* », « *chapelle aux planche* » et « *château de sallenôves* ». Toutes les séquences reconnues par le graphe principal sont annotées par l'étiquette <*ENAMEX+INSTITUTION*>.

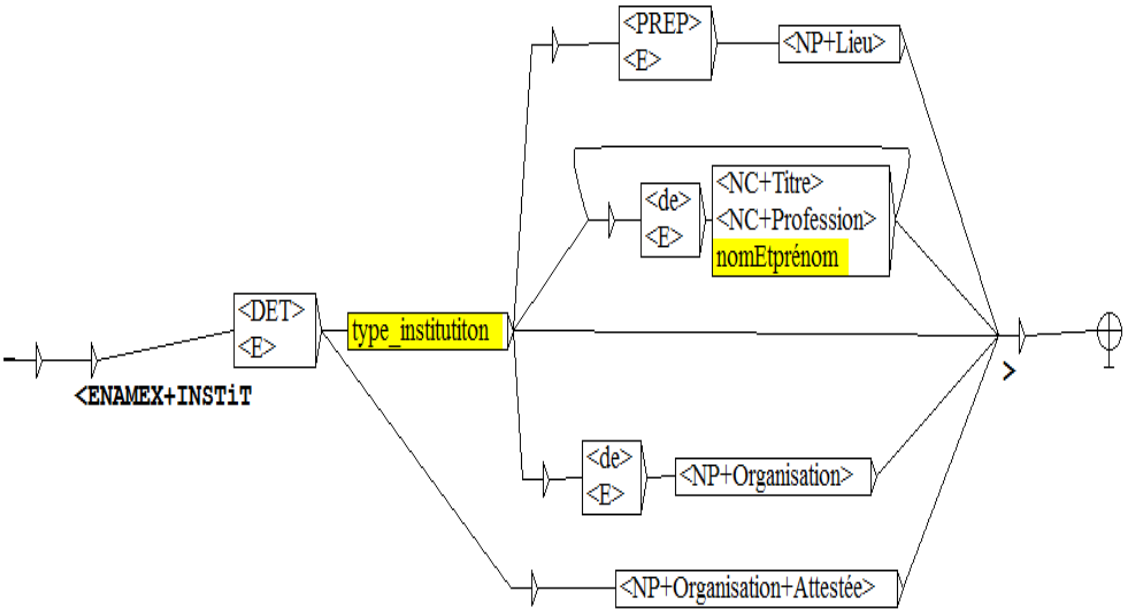


Figure 142. Graphe de reconnaissance des institutions

3.2.4. Grammaire de reconnaissance des entités numériques

Les expressions numériques ne sont pas associées directement à un objet mental mais elles forment une représentation numérique par composition d'éléments et sans faire appel à des inférences logiques (Nouvel, 2012). Nous constatons, en examinant les diverses formes des expressions numériques existantes dans MEDITEXT, qu'elles sont constituées principalement d'une valeur permettant de quantifier, qui pourrait être un déterminant numérique ou un chiffre arabe, suivie d'une unité de mesure, de monnaie ou d'une unité de masse ou de volume. La reconnaissance des nombres en chiffres arabes et romains et des déterminants numériques est une tâche d'analyse lexicale décrite au chapitre 5 indispensable pour identifier les expressions numériques ainsi que d'autres expressions linguistiques de la syntaxe locale. En effet, l'analyse lexicale a permis le développement des grammaires locales permettant la reconnaissance des nombres en chiffres romains et en chiffres arabes et des mots simples et composés qui correspondent aux nombres cardinaux ou ordinaux. Ces grammaires tiennent compte de la spécificité des nombres et des déterminants numériques en moyen français, à titre d'exemple les différentes variantes orthographiques des nombres cardinaux et ordinaux et des mots issus du latin. Les annotations produites par ces grammaires sont utilisées pour la reconnaissance des trois entités numériques à savoir les expressions monétaires qui correspondent au sous-graphe « prix », les expressions de mesure qui correspondent au sous-graphe « Mesure » et les expressions de poids qui correspondent au sous-graphe « poids ».

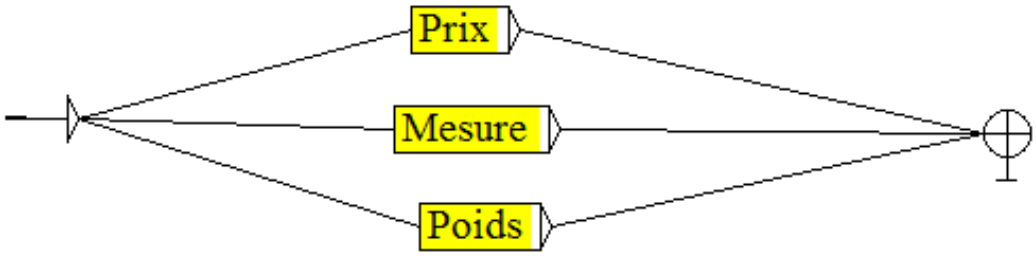


Figure 143. Graphe principal de reconnaissance des entités numériques

Grâce à des analyses textométriques, des expressions monétaires diverses ont été recensées. Comme l'illustre le graphe de la figure 144, ces expressions sont composées essentiellement d'un nombre composé des chiffres arabes ou d'un déterminant numérique ou d'une séquence d'ALU, qui désigne une quantité, décrite par le sous-graphe « description », suivie d'une unité monétaire.

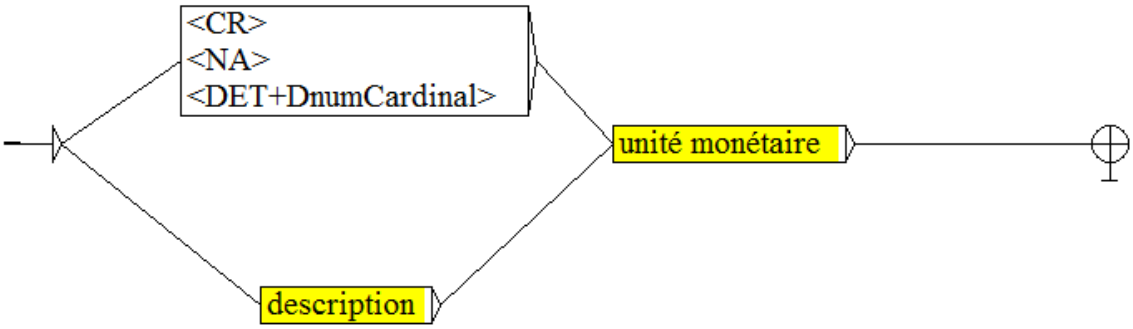


Figure 144. Sous-graphe « prix »

Le sous-graphe « unité monétaire », présenté ci-dessous, est utilisé pour le sous-graphe « prix » pour dénombrer les différentes unités monétaires.

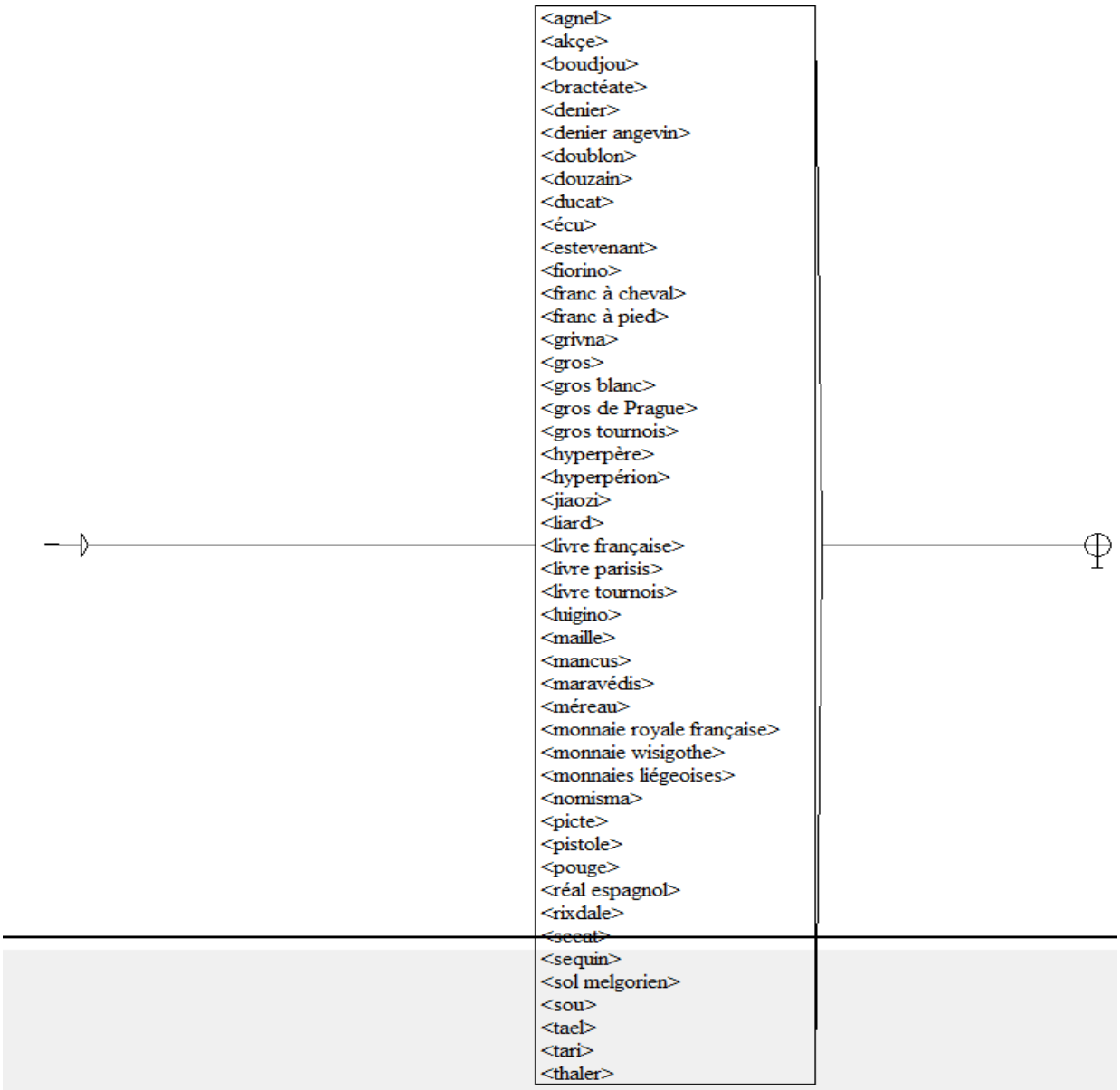


Figure 145. Sous-graphe « unité monétaire »

Comme les expressions monétaires, les expressions de mesure sont composées d'une valeur numérique suivie d'une unité de mesure. Les unités de mesures utilisées au Moyen Âge sont différentes de celles utilisées de nos jours. En effet, le système métrique n'a été adopté qu'en 1790 par l'assemblée nationale constituante afin de créer un système de mesure stable, simple et uniforme. Au Moyen Âge, les unités de mesures sont relatives au corps humain tel que « *pouce* », « *paume* », « *palme* », « *empan* » et « *pied* », « brassé », « coudée » et « poignée ». Ces dernières sont reconnues grâce au sous-graphe « unité de mesure ».

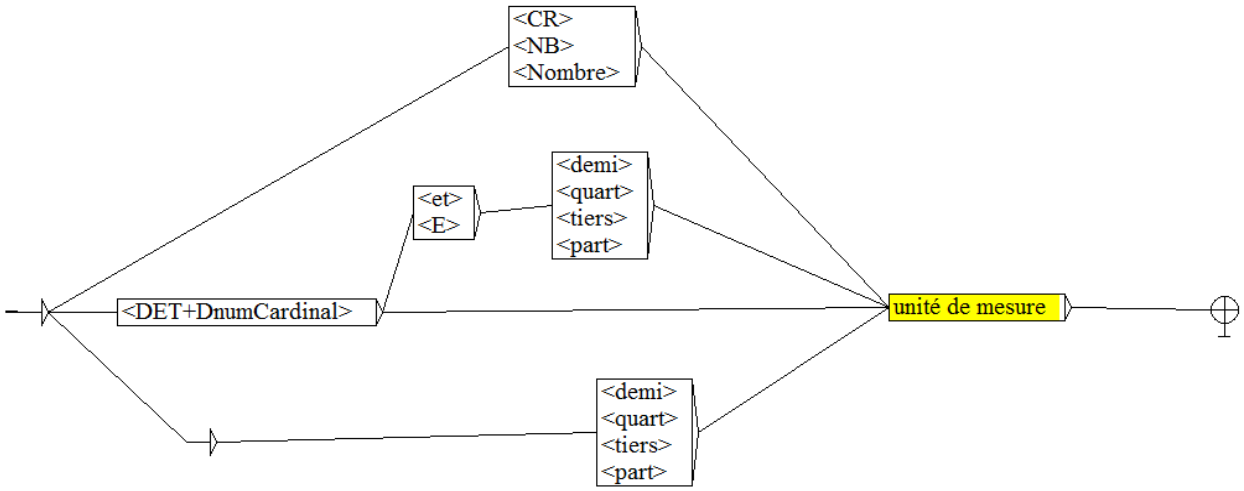


Figure 146. Sous-graphe « Mesure »

Afin de modéliser les expressions de poids, nous avons répertorié les unités de masses et de volumes utilisées sur le territoire français durant la période allant du XIVe au XVIe siècle. Ces dernières ne sont pas très nombreuses. Comme le montre la figure 147, elles sont généralement précédées par une valeur numérique exprimée par un chiffre arabe ou romain ou encore un déterminant numérique, formant ainsi les expressions de poids.

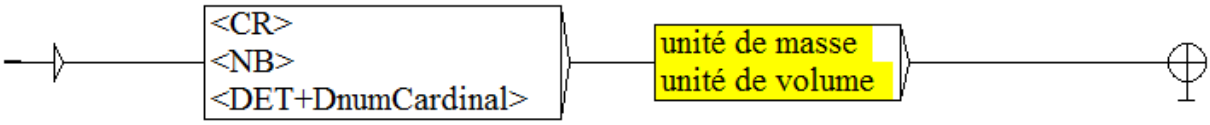


Figure 147. Sous-graphe « Poids »

3.2.5. Grammaire de reconnaissance des entités temporelles

Notre corpus, rappelons-le, est constitué principalement des textes politiques issus de rois, de parlements et de certains personnages de la vie politiques française au Moyen Âge. Ce type de texte fait référence à plusieurs événements. Ainsi, un grand nombre des expressions temporelles sont utilisées afin de situer les différents événements dans un espace temporel. Ces expressions temporelles « TIMEX » sont composées de quatre types d'expressions, à savoir les expressions de dates, horaires, mesures de temps et âge.

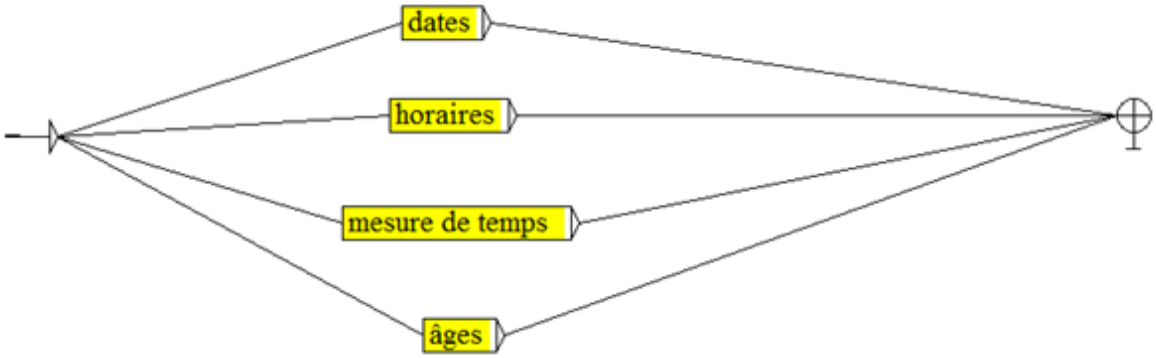


Figure 148. Graphe « TIMEX »

Les expressions de dates en moyen français sont nombreuses, diverse et variées. Nous pouvons les classer en cinq sous-catégories :

- Les dates avec chiffres arabes comme « le lundi 2 mars 1540 » : c'est le système utilisé dans les textes en français moderne. Nous notons que dans nos textes la présence du jour et de l'année est facultative. Par conséquence, notre sous-graphe « le lundi 2 mars » reconnait des expressions comme « le 2 mars 1460 » et « le 2 mars ».
- Les dates avec des chiffres romains : « le lundi VII jours de mois 1245». Pour identifier ce type de dates, nous utilisons les annotations produites lors de l'analyse lexicale par la grammaire de reconnaissance des chiffres romains. De même que pour les dates avec chiffre arabe, la présence du jour ainsi que la mention de l'année est facultative.
- Les dates en toutes lettres : « deuxième jour du mois de juillet ». Bien que cette sous-catégorie de date ressemble aux dates avec des chiffres arabes et celles avec des chiffres romains, elle regroupe des expressions plus diverses et plus variées. A cause de l'absence d'une norme précise dans l'écriture des dates en lettres, la structure syntaxique des expressions varie d'un auteur à l'autre, d'un texte à un autre et parfois au sein d'un même texte. Cette sous-catégorie contient donc des expressions de dates qui sont rarement écrites avec des chiffres. En effet, le moyen français est avant tout une langue parlée et cet ensemble des expressions couvrent les dates transcrites de l'orale qui peuvent être diverses au contraire des expressions contenant des nombres qui obligent l'auteur à respecter certaines normes. Nous citons comme exemples de ces expressions : « *quatriesme jour d'avril derrenier passé*», « *le second jour de ce present moys de janvier*» *et « septiesme jour de juillet de l'annee passee ».*
- Les dates avec mois ou saison : « *en mais 1550* ». Elles sont très fréquentes dans MEDITEXT. Elles permettent de situer des faits à l'aide essentiellement des mois ou

des saisons suivies ou non de l'année généralement mentionnée en chiffres arabes. Elles permettent ainsi d'identifier diverses dates telles que « *en hiver 1205* », « *au mois mais* » et « *en juin 1412* ».

- Les dates qui font référence à un événement : « avant pâque ». Cet ensemble d'expressions exprime le temps au moyen d'une conjonction ou préposition suivie du nom d'événement marquant comme une cérémonie religieuse, une cérémonie royale et des guerres. Le temps décrit peut avoir un rapport de simultanéité, d'antériorité ou de postériorité avec l'événement marquant auquel on fait référence. Nous avons répertorié toutes les conjonctions et les prépositions qui expriment la simultanéité. Citons parmi les usitées « *durant* », « *pendant* » et « *lors de* ». Pour l'antériorité « *avant* » et « *jusqu'à* » et la postériorité « *après* », « *depuis* » et « *dès* ». Ces différents marqueurs de temps ont été associés à une liste d'événements recensés à partir de MEDITEXT qui atteste de ces guerres et de ces cérémonies religieuses et royales.

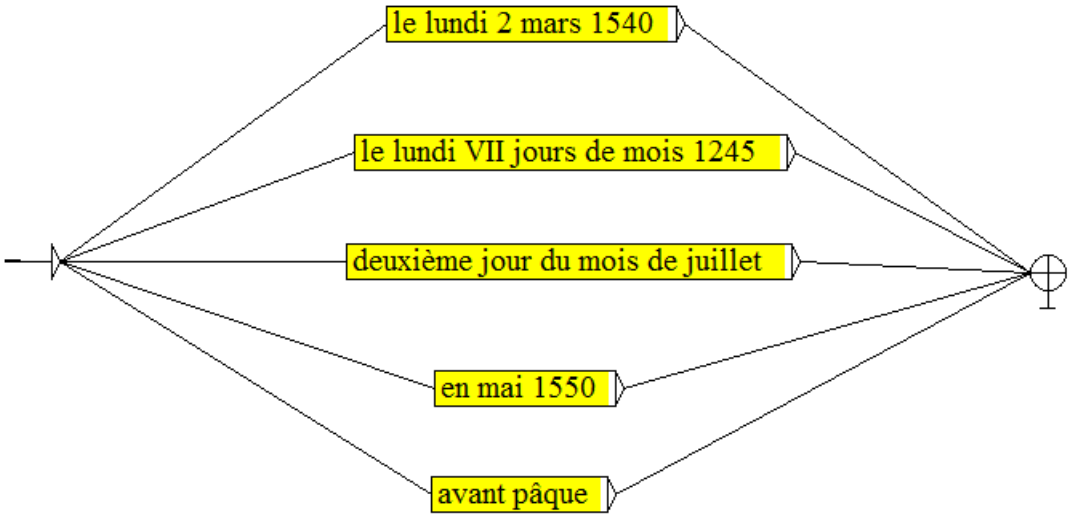


Figure 149. Sous-graphe « dates »

Le sous-graphe « horaires » permet la reconnaissance des différentes expressions d'horaire donnant des indications sur l'heure. Il produit l'étiquette *<ADV+Horaire>* décrivant un adverbe *« ADV »* ayant un trait sémantique *« Horaire »* permettant d'annoter les séquences d'ALU reconnues par les deux sous-graphes « *à dix heures* » et « *à l'aube* ».

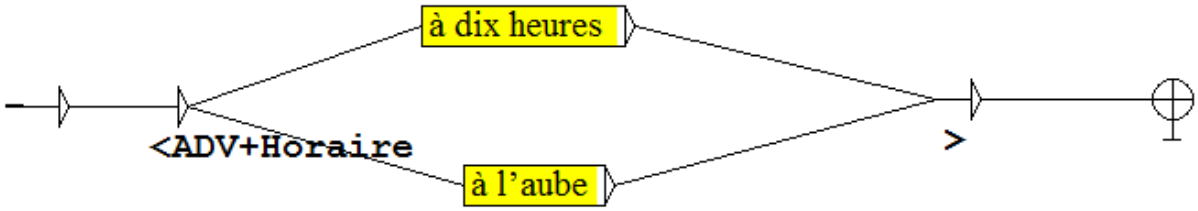


Figure 150. Sous-graphe « horaires »

Au contraire du français moderne, nous constatons qu'en moyen français, on utilise rarement les chiffres pour exprimer l'heure : par exemple, l'expression « *à dix heures* » est orthographiée en toutes lettres, nous constations que son équivalent en chiffre par exemple « *à 10 heures* » est inutilisé dans MEDITEXT. Comme le montre le sous-graphe « à dix heure » de la figure 151, nous avons élaboré toutes les possibilités de combinaisons de l'heure comme « *huyt heures du matin*», « *neuf heures de nuyt*» et « *une heure apres midi* ». Dans le but de prendre en considération la spécificité du moyen français et de MEDITEXT, un lexique spécifique sur les systèmes solaire et lunaire a été recensé permettant au sous-graphe « à l'aube » d'extraire des expressions comme « *l'heure de coucher de soleil* » ou « *à l'aube* ».

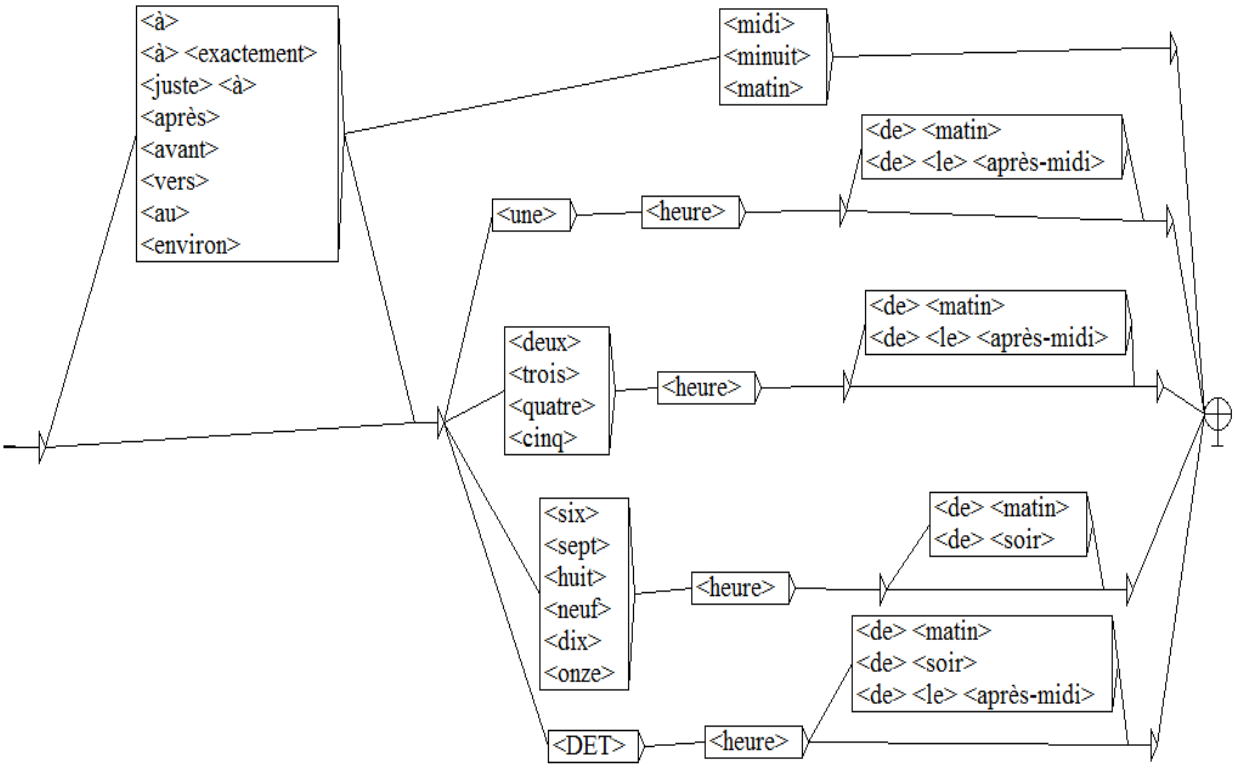


Figure 151. Sous-graphe « à dix heures »

Les expressions de mesure de temps permettent d'identifier la durée. D'abord, nous avons commencé par recenser les différents termes exprimant la durée tels que « *pendant* », « *depuis* », « *de… à ..* », « *depuis… à …* », « *à partir de* » et « *au bout de* ». Comme le montre la figure 152, ces termes nécessaires à la reconnaissance des expressions de mesure de temps

sont suivis des séquences reconnus par le sous-graphe « horaires » (voir figure 150) ou le sous-graphe « dates » (voir figure 149) afin de décrire les bornes de la durée à identifier.

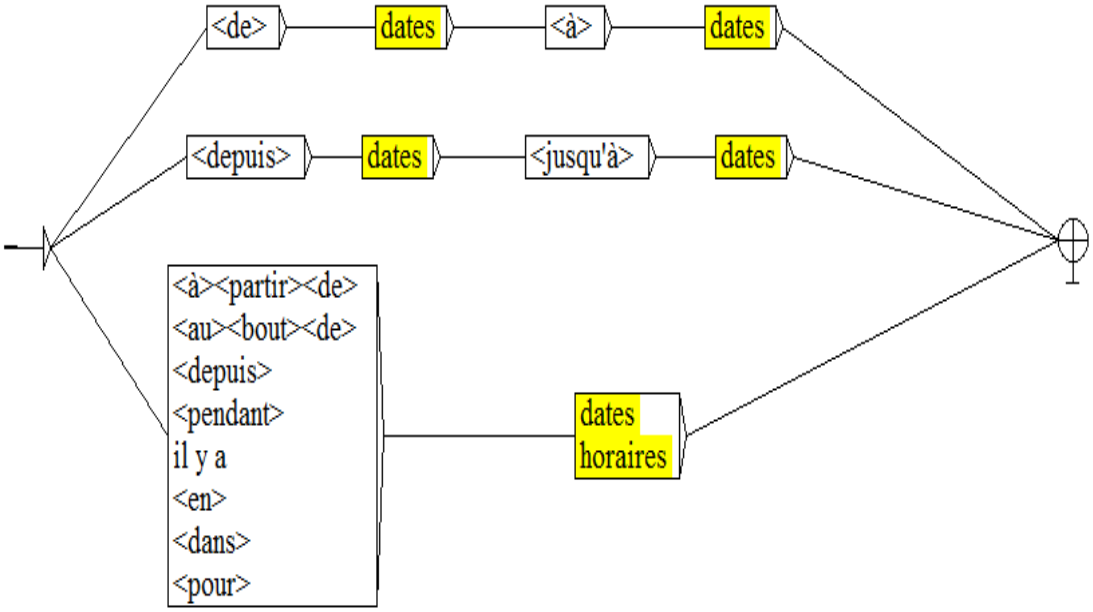


Figure 152. Sous-graphe « Mesure de temps »

Les expressions d'âges cherchent à décrire l'âge d'un animé ou d'un évènement. Elles sont composées d'un déterminant numérique suivi d'un marqueur lexical désignant l'unité des calculs qui est généralement l'année mais ce peut être la décennie ou le siècle. Nous distinguons deux types d'expressions d'âges : *(i)* celles qui sont constituées avec les nombres cardinaux et *(ii)* celles qui sont constituées avec les nombres ordinaux. Nous avons donc utilisé notre grammaire d'extraction des déterminants numériques et nous avons dressé une liste des marqueurs lexicaux qui peuvent être une unité de calcul temporel comme les saisons, l'année, la décennie et le siècle.

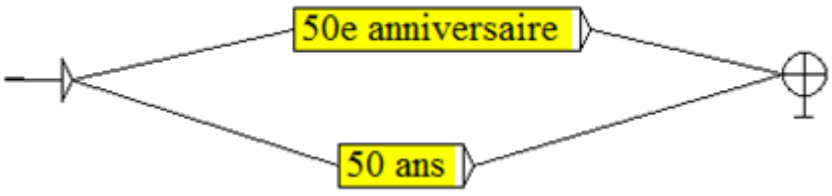


Figure 153. Sous-graphe « âges »

4. Évaluation

Nous avons mis en place un système d'évaluation des performances pour notre méthode de reconnaissance d'entités nommées. En effet, nous avons constitué un corpus de test différent du corpus de référence que nous avons annoté manuellement en identifiant et en catégorisant en langage XML les expressions linguistiques qui font référence à des entités

nommées à l'aide d'une interface en ligne que nous avons mis en place. Puis, ce même corpus de test a été annoté avec notre méthode « multiple » de reconnaissance des entités nommées. Finalement, notre système d'évaluation permet d'effectuer des calculs de performance en donnant la possibilité d'examiner les différences entre l'annotation manuelle et l'annotation automatique.

Notre système d'évaluation adopte donc aux entités nommées les métriques d'évaluation développées par le projet Cranfield (Cleverdon et al, 1966) qui sont associées au bruit et au silence documentaire : la précision, le rappel et la F-mesure.

- Rappel (R) est défini par le nombre d'entités pertinentes retrouvées par rapport à leur nombre total. Il s'oppose au silence et il permet de mesurer la capacité du système à restituer l'ensemble des entités pertinentes (Baccini et al, 2010).
  R= Nombre d'entités pertinentes retrouvées/Nombre total des entités pertinentes

- Précision (P) est définie par le nombre d'entités pertinentes retrouvées par rapport au nombre total des entités retrouvés. Elle permet de mesurer la capacité du système à ne restituer que des entités pertinentes (Baccini et al, 2010). Toutes les entités non-pertinentes retrouvées constituent le bruit de système.
  P = Nombre d'entités pertinentes retrouvées/Nombre total des entités retrouvées

- F-Mesure nommée aussi F-score. Elle permet de calculer une moyenne combinant la précision et le rappel qui favorise les systèmes qui ont des valeurs de rappel et précision homogènes.
  F-Mesure = 2*P*R/P+R

	Rappel	Précision	F-mesure
Personnes	81%	82%	81.5%
Lieux	77%	83%	79.9%
Institutions	79%	82%	80.47%
NUMEX	82%	83%	82.5%
TIMEX	83%	85%	83.99%

Tableau 15. Evaluation des entités nommées

Nous constatons que les performances de notre système REN sont encourageantes et elles présentent un intérêt indéniable pour une langue non standardisée. Cependant, notre système souffre de quelques imperfections qui ont été explorées, les causes en sont :

- la présence d'un nombre de variantes orthographiques, non prises en comptes par nos dictionnaires électroniques et par nos règles morphologiques,

- l'instabilité des structures syntaxiques des expressions linguistiques qui désignent certaines entités nommées,
- la présence du vocabulaire issu d'autres langues étrangères comme le Moyen Anglais et le latin.

5. Conclusion

Nous avons présenté une méthode dite « multiple » pour la reconnaissance des entités nommées. Elle décrit plusieurs niveaux d'analyses linguistiques allant du niveau lexical à celui de la syntactico-sémantique. La mise en place d'une telle méthode est rendue possible grâce à l'utilisation de la notation unifiée de la plateforme NooJ. A cause de la nature instable des noms propres et des noms communs en moyen français, nous avons dû procéder à l'identification et à la normalisation des preuves internes et externes. C'est la raison pour laquelle, des dictionnaires des entités nommées ont été développés afin d'annoter chaque nom avec un trait sémantique. Ces annotations ont été utilisées par nos grammaires locales pour modéliser les différentes séquences d'ALU qui font référence à des entités nommées. Les résultats obtenus affichent un taux de reconnaissance très encourageant pour une langue ayant une grammaire et une syntaxe de nature changeante et ayant une orthographe non-standardisée.

Il paraît que, pour augmenter la qualité des résultats, il est nécessaire d'enrichir nos ressources linguistiques en analysant des textes de type et des aspects géographiques et chronologiques différents de celui de MEDITEXT qui est constitué essentiellement des textes politiques. En effet, les dictionnaires électroniques peuvent être enrichis par des nouvelles entrées lexicales, des nouvelles variantes orthographiques et des règles de flexions et de dérivations provenant des textes de type différent tels que les textes littéraires et théâtreux et de périodes et de provenances différentes. Les grammaires locales peuvent également intégrer des nouvelles règles reposant sur d'autres contextes permettant ainsi une mesure plus fine du rappel.

Conclusion générale

L'objectif du présent travail est de mettre en place un système d'analyse multi-niveaux qui permet d'enrichir des textes en moyen français par des annotations pour les rendre exploitables afin de répondre à des besoins applicatifs. La méthode retenue consiste à traiter plusieurs phénomènes linguistiques en cascade, à savoir morphologiques, lexicaux, syntaxiques et sémantiques. Son intérêt vient du fait que les résultats fournis par un niveau d'analyse peuvent servir pour le niveau suivant. D'autant qu'il reste possible pour ce dernier niveau, de pouvoir en modifier ou valider les résultats fournis par les niveaux linguistiques qui précédents. Par exemple, les grammaires locales de reconnaissance des entités nommées exploitent les annotations produites par les grammaires de désambigüisation tout en pouvant également corriger ou valider ces annotations. Cette communication entre les niveaux linguistiques différents s'est rendue possible grâce à la TAS qui peut assurer l'interopérabilité des données.

A partir du corpus de référence MEDITEXT, un processus à trois étapes a été conçu de manière à permettre un enrichissement incrémental du dictionnaire en moyen français. En effet, à partir d'une liste des formes les plus fréquentes de MEDITEXT, l'interrogation des ressources linguistiques à savoir le dictionnaire du moyen français DMF et l'*Anglo-normand dictionnary* permet d'associer les différentes formes attestées à des entrées lexicales définies par ces ressources. Ensuite, une méthode itérative, qui exploite l'annotation semi-automatique des textes, a été mise en place afin enrichir le dictionnaire. Finalement, l'association des règles de flexions et de dérivation aux entrées lexicales ainsi que le recours au compilateur des dictionnaires NooJ, ont généré automatiquement des formes fléchies qui enrichissent et améliorent la couverture du dictionnaire.

Ce dictionnaire compilé est traité par un analyseur lexical afin d'annoter chaque forme du texte avec des informations linguistiques, essentiellement un lemme et une partie du discours. Toutefois la description de certains phénomènes linguistiques lexicaux se révèle nécessaire pour l'identification de toutes les ALU. Cet analyseur comporte trois opérations essentielles à savoir les analyses typographiques, la reconnaissance des mots simples et la reconnaissance des mots composés. Les analyses typographiques apportent une solution aux problèmes liés à l'utilisation des caractères, plus précisément les signes de ponctuations, l'apostrophe et les nombres en chiffres arabes et romains. Quant aux mots simples, l'ensemble de traitements

appliqués a rendu possible l'analyse de phénomènes qui leur sont spécifiques, la contraction, l'agglutination et la désagglutination. Finalement, le recours aux techniques permettant le regroupement des séquences des ALU simples à servi à décrire les mots composés de la langue standard.

Ensuite, une série de grammaires locales ont donc été développées permettant la désambiguïsation des mots grammaticaux ainsi que des formes qui apparaissent dans leurs contextes immédiats. Ces grammaires décrivent les contextes caractéristiques des formes fréquentes et ambigües que sont les déterminants-pronoms « *le* », « *la* », « *li* », « *l'* » *et* « *les* », les articles indéfinis « *un* », « une », « *de* » et « *des* », les articles partitifs « *de* » et « *des* », les pronoms et adjectifs démonstratifs, les adjectifs et pronoms possessifs et les déterminants-pronoms indéfinis. Cette description des contextes est effectuée par une analyse textométrique qui utilise les cooccurrences, les segments répétés et les concordances. Ces grammaires ont réduit considérablement la taille de la TAS en ne gardant que les annotations pertinentes selon le contexte. Il ressort qu'il est utile, voire nécessaire, d'appliquer les grammaires de désambiguïsation avant les grammaires de la reconnaissance des entités nommées.

De plus des annotations fournies par les analyses lexicale et morphosyntaxique, des annotations sémantiques ont été ajoutées à la TAS par la prise en compte des dictionnaires électroniques dits *dictionnaires des entités nommées*. En effet, ces annotations correspondent aux preuves internes et externes permettant l'identification et la catégorisation des entités nommées. Ecrites sous formes de transducteurs, les grammaires locales ont ainsi utilisé les différentes annotations de la TAS. Ce faisant, elles ont modélisé et regroupé un ensemble de règles non-contextuelles qui décrivent des séquences d'ALU ainsi que leurs contextes. Les séquences d'ALU ainsi reconnues font référence tant aux noms de personne, de lieu et d'institution qu'à des expressions temporelles comme les dates, les horaires, les mesures de temps et l'âge ainsi que les expressions numériques que sont les prix, les mesures et les poids.

Notre système d'analyse multi-niveaux a été intégré dans PALM en vue de son utilisation par des collègues historiens du laboratoire LAMOP pour la constitution de corpus annotés en moyen français. La mise en production de notre système permettra son amélioration en envisageons des éventuels ajouts et développements de nouvelles règles et de nouvelles grammaires pour enrichir notre analyseur lexical, notre module d'étiquetage morphosyntaxique et notre méthode de reconnaissance d'entités nommées. Il est également possible d'intégrer à notre système d'autres modules qui permettent de répondre à d'autres

besoins tels qu'un module d'analyse des relations entre entités et un module d'analyse syntaxique permettant l'application d'une grammaire syntagmatique ou une grammaire de dépendance.

Informatiquement, notre système peut être extensible pour admettre un système d'analyse massive de gros volume des données en exploitant des architectures avec parallélisme massif et un écosystème de calculs distribués.

Annexes

Annexe A

Tableau de description du jeu d'étiquettes

Abréviation	**Partie de discours**	**Description**
ADV	Adverbe	Adverbe général Adverbe de négation (*ne, pas, goutte, mie, point, onques*) Adverbe interrogative (*quand, comment, où*)
A	Adjectif	Adjectif qualificatif Adjectif comparatif (*meilleur, pire*) Adjectif superlatif (*meilleur*) Adjectif indéfini (*une autre nef*) Adjectif possessif (*contre .i. suen voisin ; le vostre mandement*)
NP	Nom propre	aaron, adam, aix-en-provence, alfred-le-grand, Allemagne,…
NC	Nom commun	allié, ambassade, amende, ampoule, pigeon,...
PRO	Pronom	Pronom personnel/réfléchi (direct, indirect, tonique): *(je, me, moi, lui, leur)* Pronom relatif (*que, qui, donc [=dont], quoi, où, lequel…)* Pronom impersonnel (ex. *il semble que…*) Pronom adverbial (*en, y*) Pronom possessif (*je y lesseré le mien* ; *tien, suen*) Pronom démonstratif (*cist, cil, cestui, icelui…*) Pronom indéfini (*en [=on], ung, nul, nully, aucun, autre*…) Pronom interrogatif (*qui, quoi*)
V	Verbe	Indicatif présent (*je voys [=vais]*) Indicatif imparfait (*ele aloit*) Indicatif future (*je lairrai, il venra*) Indicatif passé simple (*ce fu*) Conditionnel présent (j'aroie, nous finirons) Subjectif présent (*que nous voisons [=allions], que vous feissiés*) Subjonctif imparfait (*il allast, je*

		voulsisse) Impératif (*soions paisible, bien viengnez*) Infinitif Participe passé Participe présent auxiliaire (*pooir, voloir, devoir*)
CONJC	Conjonction de coordination	*mais, ou, et, donc, or, ni, car,…*
CONJS	Conjonction de subordination	*que, si*, qui,…
DET	Déterminant	Déterminant défini (*li, le, la, les*) Déterminant non défini (*une bele demoiselle*) Déterminant démonstratif (*cel hom* ; *en ceste nuit ; cestuy lieu ; celui, iceluy, celle*) Déterminant possessif (*s'amour, s'atente* ; *mon, ma, nostre, vostre…*) Déterminant indéfini (*chascun jour, chascune creature, trestout*) Déterminant interrogatif (*quele aventure vos a ça amenez ?*) Déterminant composé (*ledit, ladite*) Déterminant relatif (*pour quelle amour [= pour l'amour de qui]*)
INTJ	Interjection	*ha, hé, o, a, aa,…*
NUMORD	Nombre ordinal	chiffre ordinal (adj., pronom) (*li quartz jors [adj]* ; *au nueviesme [=pronom]*)
NUMCARD	Nombre cardinal	chiffre cardinal (adj., pronom) (*après ces ii vertuz [adj]* ; *ii de ses neveuz[pronom]*)

Annexe B

Les opérateurs spéciaux

opérateur	Désignation
<WF>	Une suite de caractères à laquelle on associe une forme
<L>	Passer le curseur à gauche
<R>	Passer le curseur à droite
<LOW>	Une suite de caractères à laquelle on associe une forme miniscule
<W>	Un caractère en miniscule
<UPP>	Une suite de caractères à laquelle on associe une forme majuscule
<U>	Un caractère en majuscule
<CAP>	Une suite de caractères à laquelle on associe une forme dont seulement le premier caractère caractère est en majuscule
<NB>	Une suite de chiffres
<D>	Un chiffre
<P>	Un délimiteur
<^>	Marque de début d'une suite de caractères
<$>	Marque de la fin d'une suite de caractères
<V>	Une voyelle

Annexe C

Liste des textes Français du MEDITEXT

I. Royaume de France et pays bourguignons

1. Traités politiques

- XIIIe siècle

FrP60TP13Fr3	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy (2/6)
FrP60TP13Fr4	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy (3/6)
FrP60TP13Fr5	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy (4/6)
FrP60TP13Fr6	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy (5/6)
FrP60TP13Fr7	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy (6/6)
FrP60TP13Fr8	Gilles de Rome, Li livres du gouvernement des rois, trad. fr. de Henri de Gauchy de "De Regimine Principum" (1/6)

- XIVe siècle

FrP60TP14Fr1	Jean de Salisbury, Policraticus, trad. de Denis Foulechat (1372), extrait du livre VII

- XVe siècle

FrP60TP15Fr1	Alain Chartier, Le Quadrilogue invectif
FrP60TP15Fr2	Avis à Yolande d’Aragon
FrP60TP15Fr3	Le débat des hérauts d'armes de France et d'Angleterre
FrP20TP15Fr1	Jean JUVENAL DES URSINS, A, A, A, Nescio Loqui [2/2]
FrP20TP15Fr3	Jean JUVENAL DES URSINS, A, A, A, Nescio Loqui [1/2]
FrP20TP15Fr2	Jean JUVENAL DES URSINS, Audite celi
FrP21TP15Fr1	icéron, trad. Laurent de Premierfait, Le livre de Tulle d’Amistié [extraits]

2. Textes historiques

- *XIVe siècle*

FrP50HI14Fr1	Les grandes chroniques de France. Chronique des règnes de Jean II et de Charles V, t. 1 [1350-1364] (1/2)
FrP50HI14Fr2	Les grandes chroniques de France. Chronique des règnes de Jean II et de Charles V, t. 1 [1350-1364] (2/2)
FrP50HI14Fr3	Chroniques des quatre premiers Valois (1327-1393), p. 1-100.
FrV50HI14Fr1	Jean Creton, Histoire du roy d'Angleterre Richard, 1399 [1/2]
FrV50HI14Fr2	Jean Creton, Histoire du roy d'Angleterre Richard, 1399 [2/2]

- *XVe siècle*

FrP50HI15Fr1	Gilles le Bouvier, Les Chroniques du roi Charles VII [extraits]
FrP50HI15Fr3	Registre de la famille de Gennes [1499-1562]
FrP50HI15Fr4	Guillaume Cousinot, La geste des nobles françois [extrait: sur le traité de Troyes]
FrP50HI15Fr5	Histoire de Charles VI [extrait: sur le traité de Troyes]
FrP50HI15Fr6	Chronique anonyme du roi Charles VI [extrait: sur le traité de Troyes]
FrP50HI15Fr7	recueil de petite chronique / France et Angleterre [fr 5059] [extrait: sur le traité de Troyes]
FrP50HI15Fr8	Chronologie universelle jusqu'a la mort de Charles VII [extrait: sur le traité de Troyes]
FrP50HI15Fr9	Cronicques abregee de l'histoire de France [extrait: sur le traité de Troyes]
FrP50HI15Fr10	Summa Cronicorum Francie [extrait: sur le traité de Troyes]
FrP50HI15Fr11	Chronique des rois de France jusqu'à la mort de Louis XI [extrait: sur le traité de Troyes]
FrP50HI15Fr12	Grandes Chroniques de France (annees 1368-1461) [extrait: sur le traité de Troyes]
FrP50HI15Fr13	Pierre Cochon, Chronique Normande [extrait: sur le traité de Troyes]
FrP50HI15Fr14	Journal de Clément de Fauquembergue
FrP50HI15Fr15	Mémoires [de Pierre de Fenin]
FrP50HI15Fr16	J. Lefevre de Saint-Remy, Chronique [extrait sur le traité de Troyes]
FrP50HI15Fr17	Enguerrand de Monstrelet, Chronique [extrait sur le traité de Troyes]

FrP50HI15Fr18	Jean de Wavrin, Recueil de Croniques et anchiennes histoiries [extrait sur le traité de Troyes]
FrP50HI15Fr19	Journal d'un Bourgeois de Paris [extrait: sur le traité de Troyes]
FrP50HI15Fr20	Petite chronique allant de 1400 a 1467 (BnF ms Fr 5365) [extrait: sur le traité de Troyes]
FrP50HI15Fr21	Chronique francaise : chronologie des rois de France jusqu'a Charles VI (BnF ms fr 10468) [extrait: sur le traité de Troyes]
FrV50HI15Fr1	La Complaincte des bons françois, trad. Robert Blondel [extraits]

3. Actes et Lettres

- XIVe siècle

FrP26AL14Fr1	Ordonnance de Jean II [30 mars 1357 (n.s.)]
FrP26AL14Fr2	Ordonnance de Charles VI sur les Juifs [mars 1385 (n.s.)]
FrP26AL14Fr3	Ordonnance de Charles VI [9 fév. 1388 (n.s.)]
FrP26AL14Fr4	Ordonnance de Charles VI [5 fév. 1389 (n.s.)]
FrP26AL14Fr5	Ordonnance de Charles VI [28 fév. 1389 (n.s.)]
FrP26AL14Fr6	Ordonnance de Charles VI [28 fév. 1389 (n.s.)] (2)
FrP26AL14Fr7	Ordonnance de Charles VI [1 mars 1389 (n.s.)]
FrP26AL14Fr8	Ordonnance de Charles VI, 2 juillet 1388
FrP26AL14Fr9	Ordonnance de Charles VI [9 fév. 1389 (n.s.)]
FrP26AL14Fr15	Ordonnance de Charles VI [28 déc. 1388]
FrP26AL14Fr16	Ordonnance de Charles VI [28 jan. 1390 (n.s.)]
FrP26AL14Fr19	Lettres de rémission de 1358-9, publiées par S. Luce dans l"Histoire de la Jacquerie"
FrP26AL14Fr20	Lettre de rémission, JJ 231, no. 61 [1488]
FrP26AL14Fr21	7 Lettres de rémission de 1358 dans Archives nationales, JJ86
FrP26AL14Fr22	Dossier concernant Etienne Marcel à partir d'Archives Nationales, JJ86 : 6 lettres de rémission et 8 lettres de donation, d'abolition et de confiscation.
FrP26AL14Fr23	Dossiers concernant Etienne Marcel à partir d'Archives Nationales, JJ90 : 1 Lettre de rémission et 4 lettres de donation
FrP26AL14Fr25	30 Lettres de rémission dans Archives Nationales, JJ120 (1381-1382)

FrP26AL14Fr26	30 Lettres de rémission dans Archives Nationales, JJ127 [1385]
FrP26AL14Fr28	30 Lettres de rémission dans Archives Nationales, JJ143 [1392]
FrP26AL14Fr29	30 Lettres de rémission dans Archives Nationales, JJ151 [1396]
FrP26AL14Fr31	7 lettres de rémission accordées pour cause de folie, JJ77-JJ146 [1349-1394]
FrP26AL14Fr32	Lettre d'abolition pour la ville d'Amiens, sept. 1358
FrP26AL14Fr33	7 Lettres de donation, d'abolition et de confiscation concernant Etienne Marcel, dans Archives Nationales, JJ86
FrP26AL14Fr34	4 Lettres de donation, concernant Etienne Marcel, dans Archives Nationales JJ90
FrP26AL14Fr35	Résumé des plaidoiries du 4 mai 1377 dans le procès intenté par Hugues Bernier (dossier Etienne Marcel)
FrP26AL14Fr36	6 Lettres de rémission de 1382-1383 [thème: insurrections urbaines en Normandie]
FrP26AL14Fr46	Choix de pièces inédites relatives au règne de Charles VI [33 lettres de rémission]
FrP26AL14Fr47	2 lettres de rémission : JJ119 (1381). Choix de pièces inédites relatives au règne de Charles VI [1/17]
FrP26AL14Fr48	4 lettres de rémission : JJ120 (1382). Choix de pièces inédites relatives au règne de Charles VI [2/17]
FrP26AL14Fr49	4 lettres de rémission : JJ122 (1383). Choix de pièces inédites relatives au règne de Charles VI [3/17]
FrP26AL14Fr50	3 lettres de rémission : JJ123 (1383-1384). Choix de pièces inédites relatives au règne de Charles VI [4/17]
FrP26AL14Fr51	1 lettre de rémission : JJ128 (1385). Choix de pièces inédites relatives au règne de Charles VI [5/17]
FrP26AL14Fr52	3 lettres de rémission : JJ132 (1388). Choix de pièces inédites relatives au règne de Charles VI [6/17]
FrP26AL14Fr53	1 lettre de rémission : JJ136 (1389). Choix de pièces inédites relatives au règne de Charles VI [7/17]
FrP26AL14Fr10	2 lettres de rémission : JJ153 (1397-1398). Choix de pièces inédites relatives au règne de Charles VI [8/17]
FrP60AL14Fr2	Dossier de lettres concernant Etienne Marcel
FrP60AL14Fr3	4 Lettres d'Etienne Marcel [1357- 1358] [JJ86 (pièces 67, 94, 76) + JJ90-193]

FrP60AL14Fr4	4 Lettres d'Etienne Marcel [1357- 1358] [JJ86 (pièces 67, 94, 76) + JJ90-193]
FrP60AL14Fr5	Journal des états réunis à Paris au mois d'octobre 1356

- *XVe siècle*

FrP26AL15Fr2	1 lettre de rémission : JJ159 (1404). Choix de pièces inédites relatives au règne de Charles VI [10/17]
FrP26AL15Fr3	Ordonnance de Charles VI [7 jan. 1408 (n.s.)]
FrP26AL15Fr4	Ordonnance de Charles VI [7 jan. 1401 (n.s.)]
FrP26AL15Fr5	Ordonnance de Charles VI, 20 oct. 1409
FrP26AL15Fr6	JJ217 : 139 lettres de rémission de 1487 (1/5)
FrP26AL15Fr7	JJ217 : 139 lettres de rémission de 1487 (2/5)
FrP26AL15Fr8	JJ217 : 139 lettres de rémission de 1487 (3/5)
FrP26AL15Fr9	JJ217 : 139 lettres de rémission de 1487 (4/5)
FrP26AL15Fr11	30 Lettres de rémission dans Archives Nationales, JJ155 [1400]
FrP26AL15Fr12	30 Lettres de rémissions dans Archives Nationales, JJ160 [1405]
FrP26AL15Fr13	30 Lettres de rémission dans Archives Nationales, JJ165 [1410/12]
FrP26AL15Fr14	30 Lettres de rémissions dans Archives Nationales, JJ169 [1415/1417]
FrP26AL15Fr15	5 Lettres de rémission accordées pour cause de folie, JJ169-JJ175 [1415-1426]
FrP26AL15Fr16	L'Ordonnance cabochienne (26-27 mai 1413)
FrP26AL15Fr20	2 lettres de rémission : JJ161 (1407). Choix de pièces inédites relatives au règne de Charles VI [11/17]
FrP26AL15Fr21	1 lettres de rémission : JJ164 (1409). Choix de pièces inédites relatives au règne de Charles VI [12/17]
FrP26AL15Fr22	1 lettre de rémission : JJ165 (1411). Choix de pièces inédites relatives au règne de Charles VI [13/17]
FrP26AL15Fr23	1 lettre de rémission : JJ166 ? (mai 1412). Choix de pièces inédites relatives au règne de Charles VI [14/17]
FrP26AL15Fr24	1 lettre de rémission : JJ169 (1415). Choix de pièces inédites relatives au règne de Charles VI [15/17]
FrP26AL15Fr25	3 lettres de rémission : JJ170 (1415-1418). Choix de pièces inédites relatives au règne de Charles VI [16/17]
FrP26AL15Fr26	1 lettre de rémission : JJ160 (1403). Choix de pièces inédites relatives au règne de Charles VI [17/17]

FrP26AL15Fr28	Lettres de rémission: Paris pendant la domination anglaise (1420-1436)
FrP26AL15Fr29	JJ171 (1420-1421) : lettres de rémission 1-14, extrait de "Paris pendant la domination anglaise (1420-1436)" [1/5]
FrP26AL15Fr31	JJ172 (1422-1424) : lettres de rémission 15-32, extrait de "Paris pendant domination anglaise (1420-1436)" [2/5)
FrP26AL15Fr32	JJ173 (1424-1427) : lettres de rémission 33-61, extrait de "Paris pendant domination anglaise (1420-1436)" [3/5]
FrP26AL15Fr33	JJ174 (1428-1430) : lettres de rémission 62-69, extrait de "Paris pendant la domination anglaise (1420-1436)" [4/5]
FrP26AL15Fr35	JJ175 (1430-1435) : lettres de rémission 70-80, extrait de "Paris pendant la domination anglaise (1420-1436)" [5/5]
FrP26AL15Fr36	JJ217 : 139 lettres de rémission de 1487 (5/5)
FrP26AL15Fr39	Ordonnance de Charles VI, 20 oct. 1409

- *XVIe siècle*

FrP26AL16Fr3	L'ordonnance de Villers-Cotterêts [ordonnance d'août 1539]

4. Discours

- *XVe s.*

FrP60D15Fr1	Jean Gerson, Proposition devant le Sénat dans le fait de Ch. de Samoisy.
FrP60D15Fr2	Jean Gerson, Discours au roi contre le prévôt Guillaume de Tignonville
FrP60D15Fr3	Jean Gerson, Discours pour la paix de l'Eglise et l'union des Grecs, 1409
FrP60D15Fr4	Jean Gerson, Discours au roi contre Jean Petit, 1413
FrP60D15Fr5	Jean Gerson, Discours au roi pour la réconciliation
FrP60D15Fr6	Jean Gerson, Discours au roi 'Vivat rex' pour la réformation du royaume
FrP20D15Fr1	Jean Tinctor, Invectives contre la secte de vauderie (traduction francaise du Tractatus contra sectam vaudensium de Jean Tinctor)

5. Sermons

- *XIVe s.*

FrP20S14Fr1	Jean Gerson, Pour la Pentecôte
FrP20S14Fr3	Jean Gerson, Neuf considérations (ou avis pour la conduite chrétienne)

FrP20S14Fr4	Jean Gerson, La montaigne de contemplation
FrP20S14Fr5	Jean Gerson (?), Briesve introduction en la sainte foy crestienne
FrP20S14Fr7	Jean Gerson, Les Vertus et l'enfant Jésus
FrP20S14Fr8	Jean Gerson, Piteuse complainte
FrP20S14Fr9	Jean Gerson, Discours sur l'excellence de la virginité
FrP20S14Fr10	Jean Gerson, Pour l'Epiphanie
FrP20S14Fr11	Jean Gerson, Pour la fête de l'Annonciation
FrP20S14Fr12	Jean Gerson, Pour le dimanche des Rameaux [1395]
FrP20S14Fr13	Jean Gerson, Pour la fête de saint Michel [1393]
FrP20S14Fr14	Jean Gerson, Pour la fête de la Sainte Trinité [1396]
FrP20S14Fr15	Jean Gerson, Pour la fête de saint Pierre et saint Paul [1392]
FrP20S14Fr16	Jean Gerson, Pour le dimanche de Pâques [1394]
FrP20S14Fr19	Jean Gerson, Pour 1er mercredi des Cendres
FrP20S14Fr20	Jean Gerson, Pour la fête de la Toussaint
FrP20S14Fr21	Jean Gerson, Pour la Purification de la Ste Vierge

- *XVe siècle*

FrP20S15Fr1	Jean Gerson, Sermon de la Passion (collation)
FrP20S15Fr2	Jean Gerson, Pour le 3e dimanche de Carême, [1402]
FrP20S15Fr3	Jean Gerson, Pour le 2e dimanche de l'Avent [De la luxure] [1402]
FrP20S15Fr4	Jean Gerson, Mémoire contre le péché de blasphème
FrP20S15Fr5	Jean Gerson, Exhortation générale pour la fete de la Desponsation Notre Dame
FrP20S15Fr6	Jean Gerson, Onze ordonnances
FrP20S15Fr7	Jean Gerson, Considérations sur St Joseph
FrP20S15Fr8	Jean Gerson, Aultres considérations sur St Joseph
FrP20S15Fr9	Jean Gerson, Les douze degrés d'humilité
FrP20S15Fr11	Jean Gerson, Canticordum du pèlerin ; ou Conférences spirituelles
FrP20S15Fr12	Jean Gerson, Doctrine contre conscience trop scrupuleuse
FrP20S15Fr13	Jean Gerson, Ung petit traictie pour traire a moralite toute la passion de J.C.N.S.

FrP20S15Fr14	Jean Gerson, Dialogue spirituel
FrP20S15Fr15	Jean Gerson, Le miroir de l'âme
FrP20S15Fr16	Jean Gerson, La mendicité spirituelle ou Le secret parlement, ou Le truant
FrP20S15Fr17	Jean Gerson, Méditation sur l'Ascension
FrP20S15Fr18	Jean Gerson, Examen de conscience selon les sept péchés mortels
FrP20S15Fr19	Jean Gerson, Pour qu'on refrène sa langue ; ou Petit livre contre détraction
FrP20S15Fr20	Jean Gerson, Remède contre les tentations de blasphème
FrP20S15Fr21	Jean Gerson, Sermon de la Passion
FrP20S15Fr22	Jean Gerson, Pour le jour des morts
FrP20S15Fr24	Jean Gerson, Pour le mercredi des Cendres
FrP20S15Fr27	Jean Gerson, Pour le dimanche des Rameaux [1402]
FrP20S15Fr28	Jean Gerson, Pour le jour de Pâques [1402]
FrP20S15Fr29	Jean Gerson, Pour le jour de la Pentecôte [1401]
FrP20S15Fr30	Jean Gerson, Pour la commémoraison des défunts
FrP20S15Fr31	Jean Gerson, Pour la fête du Saint-Sacrement [1402]
FrP20S15Fr32	Jean Gerson, Discours pour recommander l'Hôtel-Dieu
FrP20S15Fr33	Jean Gerson, Pour le 4e dimanche de Carême [1402]
FrP20S15Fr34	Jean Gerson, Pour la fête de la Nativité de saint Jean Baptiste [1403]
FrP20S15Fr35	Jean Gerson, Pour le 1er dimanche de Carême [1402]
FrP20S15Fr36	Jean Gerson, Pour le 3e dimanche après Pâques [1402]
FrP20S15Fr37	Jean Gerson, Poenitemini [De la gourmandise] (collation)
FrP20S15Fr38	Jean Gerson, Pour le 3e dimanche de l'Avent [De la luxure]
FrP20S15Fr39	Jean Gerson, Pour le 4e dimanche de l'Avent et Vigile de Noël [De la luxure] [1402]
FrP20S15Fr40	Jean Gerson, Pour le dimanche dans l'octave de Noël [De la chasteté]
FrP20S15Fr41	Jean Gerson, De la chasteté (collation)
FrP20S15Fr42	Jean Gerson, De la chasteté conjugale
FrP20S15Fr43	Jean Gerson, De la chasteté conjugale (collation) [1403]
FrP20S15Fr44	Jean Gerson, De l'avarice [1403]

FrP20S15Fr45	Jean Gerson, De la paresse
FrP20S15Fr46	Jean Gerson, De la paresse (collation) [1403]
FrP20S15Fr47	Jean Gerson, De la colère [1403]
FrP20S15Fr48	Jean Gerson, De l'envie [1403]
FrP20S15Fr49	Jean Gerson, De l'orgueil [pour le 2e dimanche de Carême] [1403] oeuvre 381
FrP20S15Fr50	Jean Gerson, De l'orgueil [pour le 2e dimanche de Carême] [1403] oeuvre 382
FrP20S15Fr54	Jean Gerson, De l'orgueil [pour le 3e dimanche de Carême] [1403] oeuvre 383
FrP20S15Fr58	Jean Gerson, Discours sur le fait des Mendiants
FrP20S15Fr59	Jean Gerson, Pour la commémoraison des défunts
FrP20S15Fr60	Jean Gerson, Pour la fête de la Ste Trinité
FrP20S15Fr61	Jean Gerson, Pour la fête de l'Immaculée Conception
FrP20S15Fr62	Jean Gerson, Pour le mercredi de la 4e semaine de Carême
FrP20S15Fr63	Jean Gerson, Pour la fête de la sainte Trinité [1402]
FrP20S15Fr64	Jean Gerson, Poenitemini [De la gourmandise]
FrP20S15Fr65	Jean Gerson, Pour la fête de Noël [1402] oeuvre 351
FrP20S15Fr66	Jean Gerson, Pour la fête de Noël [1402 ou 1396] oeuvre 385
FrP20S15Fr69	Jean Gerson, De l'orgueil [pour le 1er dimanche de Carême] [1403] oeuvre 381
FrP20S15Fr73	Jean Gerson, Pour la fête de saint Antoine. oeuvre 345
FrP20S15Fr74	Jean Gerson, Pour la fête de Saint Antoine [jan 1403] oeuvre 347
FrP20S15Fr75	Jean Gerson, Pour la fête de Saint Antoine [1403 ou 1396)] oeuvre 384
FrP60S15Fr1	Jean Gerson, Discours sur l'union de l'Eglise et la paix (Discours sur l’unité de l’église)
FrP60S15Fr2	Jean Gerson, Réponse aux critiques contre la proposition de l'Université pour la paix, oct. 1413

6. Poème politique

- XIVe siècle

FrV60PP14Fr1	Complainte sur la bataille de Poitiers
FrV60PP14Fr2	Eustache Deschamps, 8 ballades (contre la cour) (1383-1404)
FrV60PP14Fr4	Eustache Deschamps, 1 rondeau et 1 ballade [thème : état, nation, clocher], [1383, après 1398]
FrV60PP14Fr5	Eustache Deschamps, 5 ballades [thème : « J'ay veu les temps desordonnez »]
FrV60PP14Fr6	Eustache Deschamps, 2 ballades [thème : « une vision de l'histoire »]

- *XVe siècle*

FrV60PP15Fr1	Ballade contre les Anglais (1429)
FrV60PP15Fr2	Ballade du sacre de Reims
FrV60PP15Fr3	Une ballade au roi Charles VII
FrV60PP15Fr4	Une ballade sur la reprise de Paris par les Français, le 13 avril 1436
FrV60PP15Fr5	Le songe véritable (pamphlet politique d'un Parisien du XVe s.) (1/2)
FrV60PP15Fr6	Le songe véritable (pamphlet politique d'un Parisien du XVe s.) (2/2)
FrV60PP15Fr7	Jean Creton, Trois ballades politiques inédites

7. Traités moraux et religieux

- *XIVe siècle*

FrV20TMR14Fr1	Jean Le Bel, Li ars d'Amour, de vertu et de boneurté [extraits]
FrP20TMR14Fr1	Jean Gerson, Le jardin amoureux de l'ame dévote
FrP20TMR14Fr2	Jean Gerson, Brève manière de confession pour jeunes gens
FrP20TMR14Fr3	Philippe de Mézières, Le Songe du Vieil Pèlerin [prologue]
FrP20TMR14Fr6	Philippe de Mézières, Le Songe du Vieil Pelerin, Livre II
FrP20TMR14Fr7	Philippe de Mézières, Le Songe du Vieil Pelerin, Livre III (1/3)
FrP20TMR14Fr8	Philippe de Mézières, Le Songe du Vieil Pelerin, Livre III (2/3)
FrP20TMR14Fr10	Philippe de Mézières, Le Songe du Vieil Pelerin, Livre III (3/3)
FrP20TMR14Fr12	Jean Dupin, Le roman de Mandevie [extraits]
FrP60TMR14Fr1	Jean Gerson, Requête pour les condamnés à mort

- *XVe siècle*

FrP20TMR15Fr1	Jean Gerson, Notes sur la confession

FrP20TMR15Fr2	Jean Gerson, Proesme du traité de la consolation sur la mort de ses amis
FrP20TMR15Fr3	Jean Gerson, Douze considérations pour que soit exaucée la prière
FrP20TMR15Fr4	Jean Gerson, Douze considérations pour que soit exaucée la prière (suite)
FrP20TMR15Fr5	Jean Gerson, Testamentum peregrini tendentis in paradisum
FrP20TMR15Fr6	Jean Gerson, A.B.C. des simples gens, ou Alphabetum puerorum
FrP20TMR15Fr7	Jean Gerson, Contre le Roman de la Rose
FrP20TMR15Fr8	Jean Gerson, Traité de consolation sur la mort de ses amis
FrP20TMR15Fr9	Jean Gerson, Traité des diverses tentations de l'ennemi
FrP20TMR15Fr10	Jean Gerson, Contre les folles assertions des flatteurs
FrP20TMR15Fr12	Jean Gerson, Le prouffit de sçavoir quel est peche mortel et veniel
FrP20TMR15Fr13	Jean Gerson, La science de bien mourir ; ou La médecine de l'âme
FrP20TMR15Fr18	Jehan Henry, Livre de vie active de l'Hôtel Dieu de Paris, 1re partie
FrP20TMR15Fr19	Dialogue entre un Chevalier et Crestienté [extrait]
FrP20TMR15Fr21	Jean Gerson, Contre la fête des fous
FrP21TMR15Fr2	Boccace, Des cas des nobles hommes et femmes, trad. Laurent de Premierfait, [extraits]
FrV20TMR15Fr2	La complainte de François Guérin, Lyon, 1460
FrV25TMR15Fr1	Le livre de la deablerie [Prologue, table des matières]
FrV25TMR15Fr2	Le livre de la deablerie [Premier Livre]

8. Poèmes moraux ou religieux

-XIVe siècle

FrV20PMR14Fr1	Jean Gerson, Ballade 'Amour qui fait les seraphins ardoir'
FrV20PMR14Fr2	Jean Gerson, Pratique du psalterium mystique
FrV20PMR14Fr3	Guillaume de Digulleville, Le pèlerinage de vie humaine
FrV20PMR14Fr4	Guillaume de Diguleville, Le Roman de la fleur de lis

-XVe siècle

FrV20PMR15Fr1	Jean Gerson, Livret-proverbes pour escoliers

FrV20PMR15Fr3 Jean Gerson, Les dix commandemens de la loy

FrV20PMR15Fr4 Eloy d'Amerval, Le livre de la deablerie [Prologue + table des matières]

FrV20PMR15Fr5 Eloy d'Amerval, Le livre de la deablerie [Premier Livre]

FrV20PMR15Fr6 Complainte de l'âme damnée

FrV20PMR15Fr7 Philippe de Vitri, Le Chapel des trois fleurs de lis

9. Théâtre

- *XIVe siècle*

FrV20TH14Fr1 Jean Gerson, L'école de la conscience

FrV20TH14Fr4 Jean Gerson, L'école de la raison (suite) ; Complainte de la Conscience

FrV20TH14Fr7 Jean Gerson, L'école de la raison. oeuvre 304

- *XVe siècle*

FrP20TH15Fr1 Jean Gerson, Complainte des âmes du purgatoire ; ou Pour esmouvoir les vifs a prier pour les mors

FrV20TH15Fr1 Jean Gerson, Miroir de bonne vie

FrV20TH15Fr2 Jean Gerson, La dance macabre

FrV20TH15Fr3 Le Mystère du Viel Testament

rV20TH15Fr5 Le Cycle de Mystère des Premiers Martyrs

II. Îles Britanniques

1. Traités politiques

- *XIVe siècle*

FrP60TP14En2 Adam of Orleton, Apologia [partie française] (1/3)

FrP60TP14En3 Adam of Orleton, Apologia [partie française] (2/3)

FrP60TP14En4 Adam of Orleton, Apologia [partie française] (3/3)

2. Actes et Lettres

- *XIIIe siècle*

FrP60AL13En1 Magna Carta, version française

FrP60AL13En3	Lettre royale promettant que les plans du conseil pour la réforme seront promulgués, 18 oct. 1258
FrP60AL13En4	Promulgation d'une lettre royale pour contrôler les sheriffs, 20 oct.1258
FrP60AL13En5	Provisions of Westminster (administrative and political resolutions) oct. 1259
FrP60AL13En6	Ordonnances des barons déclarant que les réformes seront promulguées, 22 fév.-28 mars 1259
FrP60AL13En7	Monstraunces, 1297
FrP60AL13En8	Confirmatio Cartarum, 10 oct. et 5 nov. 1297
FrP60AL13En9	Articuli super Cartas, Lenten Parliamant, Westminster, mars 1300.

- *XIVe siècle*

FrV60AL14En1	Letter to Edward III, 1341 [Lettre envoyé à Edouard III, 1341]
FrP60AL14En2	Serment de couronnement d'Edouard II, 1308
FrP60AL14En3	Articles contre Piers Gaveston présentés par le comte de Lincoln, mars à avril 1308.
FrP60AL14En4	Articles de Stamford, 1309
FrP60AL14En5	The Ordinances of 1311
FrP60AL14En6	Les articles des Ordainers d'après les Annales de Londres (1311)
FrP60AL14En7	Lettres patentes royales du 16 mars 1311
FrP60AL14En8	Traité de Leake, 9 août 1318, confirmé par le Parlement de York, 20 oct. 1318
FrP60AL14En9	Discours d'ouverture du parlement de Westminster, 6 oct. 1320
FrP60AL14En10	Articles contre Hugh le Despenser le fils et Hugh le Despenser le père au Parlement, 15/7/1321
FrP60AL14En11	Statut d'York, 1322
FrP60AL14En12	Lettres I et II de la reine Isabelle, citées par Adam Orleton, 1334
FrP60AL14En13	Sir William Trussell, Jugement contre Hugh Despenser junior, 1 nov. 1326
FrP60AL14En14	Six articles of deposition against Edward II [Six articles de déposition contre Édouard II]
FrP60AL14En15	Jugement contre Roger Mortimer, 26 nov. 1330

3. Discours

- XIVe siècle

FrP60D14En1	Discours d'ouverture d'Édouard II au parlement de Londres, octobre 1324
FrP60D14En2	John Straford, Discours d'ouverture du Parlement à Westminster, 30 sept. 1331
FrP60D14En3	John Stratford, Discours en forme de prédication au Parlement, 16 mars 1332
FrP60D14En4	Geoffrey Scrope, Discours au Parlement, 16 mars 1332
FrP60D14En5	John Stratford, Discours d'ouverture du Parlement, 9 sept. 1332
FrP60D14En6	Geoffrey Scrope, Causes des Somons du Parlement à York, 4 déc. 1332
FrP60D14En7	Geoffrey Scrope, Prononciation des Causes des Somons du Parlement à York, 21 jan. 1333
FrP60D14En8	Discours d'ouverture (anonyme) pour Parlement, 1er avril 1340
FrP60D14En9	Discours d'ouverture (anonyme) pour Parlement, 12 juillet 1340
FrP60D14En10	Discours d'ouverture (anonyme) pour Parlement, 26 avril 1341
FrP60D14En11	Robert Parning, Discours d'ouverture du Parlement, 30 avril 1343
FrP60D14En12	Bartholomew Burghersh, Discours sur l'état de la guerre, devant Parlement, 30 avril 1343
FrP60D14En13	Robert Sadington, Discours d'ouverture du Parlement, 10 juin 1344
FrP60D14En14	Anonyme, Cause des semonces du Parlement, 13 sept. 1346
FrP60D14En15	Bartholomew Burghersh, Discours sur l'état de la guerre au Parlement, 13 sept. 1346
FrP60D14En16	William Thorpe, Cause des Somons du Parlement, 17 jan. 1348
FrP60D14En17	William Thorpe, Cause des Somons du Parlement, 1 avril 1348
FrP60D14En18	William de Shareshull, Cause des Somons du Parlement, 15 fév. 1351
FrP60D14En19	William Shareshull, Cause des Somons du Parlement, 17 jan. 1352
FrP60D14En20	Bartholomew Burghersh, Discours au Parlement sur l'état de la guerre, 17 janvier 1352
FrP60D14En21	William Shareshull, Cause des Somons du Parlement [Grand Conseil], 27 sept. 1353
FrP60D14En22	Bartholomew Burghersh, Discours au Parlement sur l'état de la guerre, 7 oct. 1353

FrP60D14En23	William Shareshull, Cause des Somons du Parlement [Grand Conseil], 30 avril 1354
FrP60D14En24	Bartholomew Burghersh, Discours au Parlement sur l'état de la guerre
FrP60D14En25	Walter de Manny, Cause des Somons du Parlement, 25 nov. 1355
FrP60D14En26	Henry Green, Cause des Somons du Parlement, 14 oct. 1362 (prononcé en anglais)
FrP60D14En27	Simon de Langham, Causes des Sommons du Parlement (en anglais), 9 oct. 1363
FrP60D14En28	Simon de Langham (?), Causes "en especial" du Parlement
FrP60D14En29	Simon de Langham, Causes des Sommons du Parlement, 4 mai 1366
FrP60D14En30	Simon de Langham, Causes des Sommons du Parlement en especial, 5 mai 1366
FrP60D14En31	Simon Langham, Causes des Sommons du Parlement, 4 mai 1368
FrP60D14En32	Simon Langham, Suite des causes des Sommons, 5 mai 1368
FrP60D14En33	Simon Langham, archevêque de Canterbury, Discours aux Lords, 8 mai 1368
FrP60D14En34	William de Wykeham, Causes des Sommons du Parlement, 3 juin 1369
FrP60D14En35	William de Wykeham, Causes des Sommons du Parlement, 24 fév. 1371
FrP60D14En36	John Knyvet, Cause des Somons du Parlement, 5 nov. 1372
FrP60D14En37	Guy Brian, Discours aux Lords en la Chambre Blanche, 5 novembre 1372
FrP60D14En38	Guy Brian, Discours aux Lords, en la Chambre Blanche, 6 nov. 1372
FrP60D14En39	John Knyvet, Cause des Somons du Parlement, 22 nov. 1373
FrP60D14En40	John Knyvet, Cause des semonces du Parlement, 29 avril 1376
FrP60D14En41	Robert Ashton, Complément au discours d'Adam de Houghton, 28 jan. 1377
FrP60D14En42	Richard le Scrope, Rappel des cause des semonces, 15 oct.1377
FrP60D14En43	Protestation de Jean de Gand, duc de Lancastre, 15 oct. 1377
FrP60D14En44	Richard le Scrope, Cause des semonces du Parlement , 22 oct. 1378
FrP60D14En45	Richard le Scrope, Réponse aux Communes, oct. 1378
FrP60D14En46	Richard le Scrope, 2e réponse au Speaker, oct. 1378
FrP60D14En47	Simon Sudbury, Discours en défence de l'abbé de Westminster, 1377

FrP60D14En48	Richard le Scrope, Cause des Semonces du Parlement, 27 avril 1379
FrP60D14En49	Richard le Scrope, Première déclaration des causes des semonces, 17 jan. 1380
FrP60D14En50	Richard le Scrope, Deuxième déclaration des causes des sémonces, 1380
FrP60D14En51	Hugh Segrave, Discours d'ouverture du Parlement, 13 nov. 1381
FrP60D14En52	Richard le Scrope, Les causes des semonces, 8 mai 1382
FrP60D14En53	Robert Braybroke, Causes des semonces du Parlement, 7 oct. 1382
FrP60D14En54	John Gilbert, Deuxième discours d'ouverture, 9 oct. 1382
FrP60D14En55	Robert Braybroke, Discours d'ouverture du Parlement, 24 fév. 1383
FrP60D14En56	Michael de la Pole, Cause des semonces du Parlement, 27 oct. 1383
FrP60D14En57	Henry Despenser, Défense contre ses accusateurs, 24 novembre 1383
FrP60D14En58	Michael de la Pole, Deuxième intervention au Parlement contre Henry Despenser, 1383
FrP60D14En59	Henry Despenser, Deuxième réponse à ses accusateurs, 24 nov. 1383
FrP60D14En60	Michael de la Pole, Troisième intervention au Parlement contre Henry Despenser, 1383
FrP60D14En61	Michael de la Pole, Cause des semonces du Parlement, 5 mai 1384
FrP60D14En62	Michael de la Pole, Défense contre les accusations de John Cavendish, avril 1384
FrP60D14En63	Michael de la Pole, Cause des semonces du Parlement, 1 oct. 1386
FrP60D14En64	Michael de la Pole, Réponse à ses accusateurs, oct. 1386
FrP60D14En65	Thomas Arundel, Discours d'ouverture au Parlement, 3 fév. 1388
FrP60D14En66	William Wykeham, Discours d'ouverture, 17 jan. 1390
FrP60D14En67	William Wykeham, Discours, 12 novembre 1390
FrP60D14En68	Thomas Arundel, Cause del somons du Parlement, 3 nov. 1391
FrP60D14En69	Thomas Arundel, Cause del somons du Parlement, 21 jan. 1393
FrP60D14En70	Thomas Arundel, Cause des somons du Parlement, 28 jan. 1394
FrP60D14En71	Richard Fitzalan, comte d'Arundel, jan. 1394
FrP60D14En72	Richard II, réponse à Richard Fitzalan, comte d'Arundel, jan. 1394.
FrP60D14En73	Thomas Arundel, Cause des somons du Parlement, 28 jan. 1395

FrP60D14En74	Edmund Stafford, Discours d'ouverture du Parlement, 22 jan. 1397
FrP60D14En75	Richard II, Réponse aux Communes, jeudi 25 jan. 1397
FrP60D14En76	Edmund Stafford, Discours à la demande du roi, 2 février 1397
FrP60D14En77	Richard II, Discours aux Communes, 25 jan. 1397
FrP60D14En78	Richard II, Réponse aux Communes, 20 sept. 1397
FrP60D14En79	William Edington, Discours de clôture, 13 nov. 1362
FrP60D14En80	Simon Sudbury, Cause des semonces du Parlement à Northampton, 8 nov. 1380
FrP60D14En81	Michael de la Pole, Première intervention au Parlement contre Henry Despenser, 1383
FrP60D14En82	Discours 1-9 pendant le Good Parliament (1376), d'après l'Anonimalle Chronicle
FrP60D14En83	Peter de la Mare, Discours du Speaker, 1376
FrP60D14En84	Peter de la Mare, Discours du Speaker, oct. 1377
FrP60D14En85	James Pickering, Discours du Speaker, oct. 1378
FrP60D14En86	James Pickering, Discours du Speaker en réponse à Scrope, 1378
FrP60D14En87	John de Gildesburgh, Discours du Speaker, janvier 1380.
FrP60D14En88	John de Gildesburgh, Speaker du Parlement, nov. 1380
FrP60D14En89	Richard Waldegrave, Discours du Speaker, nov. 1381
FrP60D14En90	Anonymous Speaker, Accusations against Michael de la Pole
FrP60D14En91	James Pickering, Discours du Speaker, mars 1383
FrP60D14En92	John Bussy, Intervention à propos de l'expédition projetée en Lombardie, 25 janvier 1397
FrP60D14En93	John Bussy, Discours du Speaker, 18 sept. 1397
FrP60D14En94	John Bussy (?), Discours pour la suppression du pardon accordé aux Appellants, 18 sept. 1397
FrP60D14En95	John Bussy (?), Discours accusant les Appellants de trahison, 20 sept. 1397

- *XVe siècle*

FrP60D15En1	William Thirning, Cause des semonces du Parlement, 21 jan. 1401
FrP60D15En2	Henry IV, Discours en réponse aux Communes, 6 juin 1418
FrP60D15En3	Henry IV, Discours de remerciement à la fin du Parlement, 2 déc. 1407

FrP60D15En4	Henry IV, Réponse au Speaker, 28 jan. 1410
FrP60D15En5	Discours du Chancelier Thomas Beaufort, 3 nov. 1411
FrP60D15En6	Henry, prince de Galles, Discours au roi, 30 nov. 1411
FrP60D15En7	Henry Beaufort, Discours d'ouverture de la deuxième session du Parlement, 1416
FrP60D15En8	Adam Forster, Discours sur le sort des prisonniers de guerre écossais et français, 20 oct. 1402
FrP60D15En9	Henry, prince de Galles, Discours de soutien au duc d'York, 2 déc. 1407
FrP60D15En10	Arnold Savage, Protestation du Speaker, 22 jan. 1401
FrP60D15En11	Arnold Savage, Discours du Speaker, 25 jan. 1401
FrP60D15En12	Arnold Savage, Discours du Speaker sur la rébellion du pays de Galles, 21 fév. 1401
FrP60D15En13	Arnold Savage, Discours du Speaker à la fin du Parlement, [10 mars] 1401
FrP60D15En14	Arnold Savage, Discours du Speaker, Parlement de jan. 1404
FrP60D15En15	Henry Retford, Discours du Speaker, 16 oct. 1402
FrP60D15En16	John Tiptoft, Discours I du Speaker, 23 mars 1406
FrP60D15En17	John Tiptoft, Discours II du Speaker, 3 avril 1406
FrP60D15En18	John Tiptoft, Discours III du Speaker, 24 mai 1406
FrP60D15En19	John Tiptoft, Discours IV du Speaker, 7 juin 1406
FrP60D15En20	Thomas Chaucer, Discours du Speaker, 9 nov. 1407
FrP60D15En21	Thomas Chaucer, Discours II du Speaker, 14 nov. 1407
FrP60D15En23	Thomas Chaucer, Discours III du Speaker, 2 déc. 1407
FrP60D15En24	Walter Stourton, Discours du Speaker, 22 mai 1413

4. Sermons

- XIVe siècle

FrP60S14En1	Simon de Langham, Sermon et causes des Sommons du Parlement (en anglais), 20 janvier 1365
FrP60S14En2	Adam Houghton, Sermon et cause des somons du Parlement, 28 janvier 1377
FrP60S14En3	Sermon d'ouverture de Simon Sudbury, 14 octobre 1377

FrP60S14En4 Adam Houghton, Sermon d'ouverture du Parlement, 23 octobre 1378

FrP60S14En5 William Courtenay, Sermon pour l'ouverture du Parlement, 'Rex convenire fecit consilium', 9 nov. 1381

FrP60S14En6 Edmund Stafford, Sermon d'ouverture sur 'Rex unus erit omnibus', 17 sept. 1397

FrP60S14En7 Thomas Arundel, Sermon d'ouverture sur 'Incumbit nobis ordinare pro regno', 6 oct. 1399

FrP60S14En8 Simon Sudbury, Sermon au Parlement, "Unus erit pastor noster", oct. 1378

- *XVe siècle*

FrP60S15En1 Edmund Stafford, Sermon d'ouverture sur 'Pax multa diligentibus legem', 2 oct. 1402

FrP60S15En2 Henry Beaufort, Sermon d'ouverture sur 'Multitudo sapientium', 1404

FrP60S15En3 Henry Beaufort, Sermon d'ouverture sur 'Rex vocavit seniors terrae', 1404

FrP60S15En4 Thomas Langley, Sermon d'ouverture sur 'Multorum consilia requiruntur in magnis', 1 mars 1406

FrP60S15En5 Thomas Arundel, Sermon d'ouverture sur 'Regem honorificate', 1407

FrP60S15En6 Henry Beaufort, Sermon d'ouverture sur 'Decet nos implere omnem justiciam', 27 jan.1410

FrP60S15En7 Henry Beaufort, Sermon d'ouverture sur 'Ante omnem actum consilium stabile', 1413

FrP60S15En8 Henry Beaufort, Sermon d'ouverture sur 'Posuit cor suum ad investigand' leges', 30 avril 1414

FrP60S15En9 Henry Beaufort, Sermon d'ouverture sur 'Deum tempus habemus operemur bonum', 19 nov. 1414

FrP60S15En10 Henry Beaufort, Sermon d'ouverture sur 'Sicut et ipse fecit nobis ita et nos ei faciamus', 4 nov. 1415

FrP60S15En11 Henry Beaufort, Sermon d'ouverture sur 'Iniciavit vobis viam', 16 mars 1416

FrP60S15En12 Henry Beaufort, Sermon d'ouverture sur 'Operam detis ut quietis sit', 19 oct. 1416

FrP60S15En13 Thomas Langley, Sermon d'ouverture sur 'Confortamini, viriliter agite, et gloriosi eritis', 16 nov. 1417

FrP60S15En14 Thomas Langley, Sermon d'ouverture sur 'Initium sapientie timor Domini', 16 octobre 1419

FrP60S15En15 Thomas Langley, Sermon d'ouverture sur 'Inivit David consilium', 2 déc. 1420

FrP60S15En16 Thomas Langley, Sermon d'ouverture sur 'Laudans invocabo Dominum', 2 mai 1421

FrP60S15En17 Thomas Langley, Sermon d'ouverture sur 'Que magis ipsarum est, et radix omnium bonorum', 1 déc. 1421

FrP60S15En18 Henry Chichele, Sermon d'ouverture sur 'Principes populi congregantur cum Deo', 9 nov. 1422

FrP60S15En19 Thomas Langley, Sermon d'ouverture sur 'Deum timete, reges honorificate', 20 oct. 1423

5. Poèmes politiques

- *XIIIe siècle*

FrV60PP13En2 Song of the barons

FrV60PP13En3 Lament of Simon de Monfort

FrV60PP13En4 Song of the Church

FrV60PP13En6 Sur les Etats du Monde

FrV60PP13En5 Thomas Turberville

FrV60PP13En7 Vulneratur karitas

- *XIVe siècle*

FrV60PP14En1 On the King's Breaking His Confirmation of Magna Carta

FrV60PP14En2 Trailbaston

FrV60PP14En3 Elegy on Edward I

FrV60PP14En4 Lament of Edward II

FrV60PP14En5 Against the King's Taxes

FrV60PP14En6 L'Ordre de Bel Ayse

FrV60PP14En7 Lettre du Prince des Envieux

FrV60PP14En8 On the Times

FrV60PP14En9 Ingratitude of the Great

6. Textes hagiographiques

- *XIVe siècle*

FrV20HAG14En1 Nicole Bozon, La Vie seinte Juliane virgine

FrV20HAG14En2 Nicole Bozon, La Vie seinte Margarete

FrV20HAG14En3 Nicole Bozon, La vie sein Martha

FrV20HAG14En4 Nicole Bozon, De seinte Elizabeth, fille le roy de Ungarie

FrV20HAG14En5 Nicole Bozon, La vie seinte Cristine virgine

FrV20HAG14En6 Nicole Bozon, La vie seinte Angneys

FrV20HAG14En7 Nicole Bozon, La vie seinte Agace virgine

FrV20HAG14En8 Nicole Bozon, La vie la Marie-Magdalene

7. Théâtre

-XIVe siècle

FrV20TH14En1 Nicole Bozon, La vie Seinte Lucie virgine

Annexe D

Plateforme d'analyse linguistique médiévale (PALM) : Analyse de textes en moyen français

Version 0.1

Manuel d'utilisation

1. Pourquoi PALM ?

1.1 Que fait PALM ?

PALM est une plateforme en ligne qui effectue un premier traitement de textes médiévaux afin qu'ils puissent être analysés avec des logiciels conçus pour l'analyse statistique et sémantique de textes en langues modernes.

Bien que PALM comprenne une bibliothèque numérique de textes médiévaux appelée MEDITEXT, cette dernière est fournie afin de permettre à l'utilisateur de construire des corpus pour un usage statistique et sémantique. Il faut insister sur le fait qu'il ne s'agit pas d'éditions en ligne de ces textes, dont certains sont dans une forme numérique brute. Les utilisateurs souhaitant citer ces textes doivent se tourner vers les éditions scientifiques proprement dites.

De manière plus spécifique, PALM offre des facilités pour l'annotation semi-automatique de corpus de textes par des descriptions morphosyntaxiques composés essentiellement d'une partie du discours et d'un lemme.

PALM a été développé pour un usage sur des textes en latin tardo-médiéval, anglais et français du nord de la France et de l'Angleterre, mais son architecture a été conçue afin de

permettre l'annotation et le développement de ressources pour des textes dans d'autres langues.

Les textes peuvent être téléchargés dans PALM sans balisage, ou à partir de fichier préparés en XML-TEI ou sous Word. La plateforme permet une exportation dans plusieurs formats adaptés à l'usage de logiciels tels que NooJ, Hyperbase, Lexico 3, Le Trameur, Analyse et TXM.

1.2 Pourquoi le faire ?

Les utilisateurs visés par PALM sont des historiens, littéraires ou philosophes qui souhaiteraient utiliser des outils informatiques disponibles pour l'analyse statistique et sémantique de corpus de textes de la fin du Moyen Âge mais qui n'ont pu le faire à cause de l'absence d'orthographe standardisée et de la présence de graphies variées dans leurs textes.

Pour les langues modernes, de nombreux outils digitaux existent pour assister le chercheur dans des tâches allant de la simple recherche lexicale, à travers les concordances par exemple, à l'usage d'outils statistiques allant de l'identification de collocations et de cooccurrences à des méthodes statistiques sophistiquées telles que l'analyse factorielle.
Sans PALM, un chercheur qui souhaite employer ces outils pour des textes médiévaux doit d'abord annoter son corpus manuellement, en rassemblant les différentes graphies et les formes fléchies.
PALM facilite grandement le processus d'annotation, le rendant aussi automatique que possible, mais offre également des facilités pour la correction manuelle, inévitable pour des textes en latin tardo-médiéval, en moyen français et en moyen anglais.

1.3 Comment PALM annote-t-elle ?

PALM annote…

(1) grâce aux ressources linguistiques implémentées : des dictionnaires électroniques et des « règles » développées manuellement.

(2) en offrant un environnement convivial dans lequel l'utilisateur peut corriger cette annotation et créer ainsi de nouvelles ressources linguistiques.

Les corpus de textes annotés dans PALM peuvent ensuite être exportés dans différents formats qui peuvent être employés dans les logiciels d'analyse textuel conçus pour des langues modernes standardisées, tels que TXM, Tramer, Lexico 3 et Hyperbase.

1.4 Pourquoi annoter les textes ?

L'annotation est utile, y compris pour les textes en langues modernes. Elle permet d'effectuer des analyses statistiques et de recenser toutes les flexions d'un verbe, par exemple, ce qui peut être très important dans les langues fléchies telles que le français.

L'annotation est encore plus importante pour le traitement des langues vernaculaires médiévales, du fait de l'absence d'orthographe standardisée dans ces langues.

L'annotation permet de regrouper toutes les variantes orthographiques d'un lemme particulier, afin d'effectuer des analyses statistiques autrement impossibles.

En moyen français, où la variation graphique est bien marquée, on rencontre un grand nombre de mots dont l'orthographe varie, particulièrement dans des contextes pragmatiques proches de la pratique juridique ou économique par exemple.

1.5 Quels textes peuvent-ils être traités par PALM ?

Les utilisateurs peuvent à la fois importer leurs propres textes dans PALM et utiliser la bibliothèque interne MEDITEXT.

1.6 Qu'est-ce que MEDITEXT?

MEDITEXT est un corpus textuel d'abord rassemblé sous la direction de Jean-Philippe Genet et Claude Gauvard. Il a été corrigé et augmenté dans le cadre du projet du European Research Council « Signs and States » de 2010 à 2014 (voir ci-dessous, paragraphe 1.8). Il est la base de la bibliothèque interne de PALM.

MEDITEXT, et par conséquent la bibliothèque interne de PALM, rassemble essentiellement des textes « politiques » d'origine française et d'origine anglaise : c'est-à-dire, soit des textes ayant trait à des événements politiques identifiés (discours ; lettres ; traités ; poèmes ; sermons ; chroniques) ; soit des textes consacrés de manière générale au bon et au mauvais gouvernement. Il rassemble également des textes gouvernementaux (proclamations, ordonnances) et des textes adressés au roi par ses sujets (cahiers de doléances ; requêtes ; lettres de rémission).

1.7 Comment se connecter à PALM ?

PALM est accessible sur internet à l'adresse < http://palm.huma-num.fr/PALM/ >.

Pour se connecter à PALM, il suffit de s'inscrire pour devenir « Utilisateur ». En effet, il y a trois statuts possibles dans PALM : « Utilisateur » (3), « Expert » (2) et « Administrateur » (1). La plupart des utilisateurs de PALM dispose donc d'un accès « Utilisateur ». Ils peuvent consulter et utiliser des textes de la bibliothèque ainsi que dans leur espace de travail personnel au niveau « Utilisateur » (niveau 3 – pour une explication sur les niveaux d'accès aux textes, voir ci-dessous, paragraphe 2.2.4).

Les « experts » sont généralement des membres de l'équipe de PALM, chargés de développer les ressources linguistiques nécessaires afin de nourrir et d'améliorer l'annotation automatique proposée par PALM. Ils peuvent voir les textes encodés avec un niveau d'accès restreint (niveau 2) de la bibliothèque et peuvent également voir les textes des niveaux 2 et 3 des autres utilisateurs. L'utilisateur a la possibilité de rendre son texte inaccessible aux experts (par exemple s'il prépare une édition, une thèse, etc.), dans ce cas il doit attribuer à son texte le droit d'accés le plus restreint (niveau 1).

Le statut d' « Administrateur » est réservé au développement technique de PALM.

2. MEDITEXT: la bibliothèque de PALM

Lorsqu'un utilisateur se connecte pour la première fois à PALM, il accéde à une courte description de sa bibliothèque interne, « MEDITEXT ». Il est ensuite possible de retourner à cette description en cliquant sur « Présentation » dans le menu déroulant « Bibliothèque ». Pour accéder à MEDITEXT, il faut cliquer sur « Consulter la bibliothèque » dans ce dernier menu.

Nous soulignons, toutefois, que PALM n'a pas pour objectif d'offrir des éditions numériques des textes inclus. Son but est de permettre à l'utilisateur de créer un corpus de textes, qui peut ensuite être traité avant d'être exporté pour une analyse conduite par un logiciel d'analyse textométrique par exemple.

2.1 Parcourir la bibliothèque

La bibliothèque de PALM comprend plus de 480 textes. Pour une liste complète des textes, voir l'annexe B de ce manuel. Il est possible de les parcourir par titre, langue, pays d'origine, par « période » (une période couvre un demi-siècle) ou par code si le code spécifique d'un texte est déjà connu.

La fonction « Rechercher » au sein de la bibliothèque est également prévue, mais elle n'est pas encore active.

2.2 Trouver des informations sur un texte de la bibliothèque

Pour trouver davantage d'informations sur un texte, il faut effectuer un clic droit sur ce dernier et sélectionner « Détails ». Un écran « Détails du texte » apparaît, délivrant les informations de base sur le texte.

2.2.1 Titre court

Le premier champ est un titre court proposé dans un format standardisé afin de faciliter l'identification. Il ne s'agit pas du « titre » du document au sens strict, mais plus d'un nom abrégé (incluant le nom de l'auteur) afin de permettre une localisation rapide dans la bibliothèque. Pour une identification plus précise et scientifique du texte, voir le champ « Édition » ci-dessous.

La langue par défaut pour un titre court est le français, sauf quand il est généralement connu sous un nom dans une langue différente. Si le titre est seulement issu d'une convention éditoriale, une traduction en français moderne est parfois suggérée entre crochets, lorsqu'elle aide à identifier le sujet d'un texte.

Pour les auteurs bien connus en France, les noms sont donnés en français. Les noms d'auteur alternatifs en anglais ou en latin sont fournis si besoin dans le champ « Auteur » ci-dessous.

Remarque : les textes longs sont divisés en plusieurs fichiers plus courts. Quand cela est possible, les subdivisions éditoriales ou auctoriales du texte sont suivis, mais la division est parfois nécessairement arbitraire.

Exemple de titres courts typiques :

Magna Carta
Gilles de Rome, De Regimine Principum, pt. II, bk. 2
Ranulph Higden, Polychronicon, vol. viii, p. 50-100.
Against the King's Taxes
Acceptation par Richard d'York du titre de Protecteur, 17 nov. 1455

Davantage de détails sur cette forme standardisée seront donnés ci-dessous au paragraphe 3.2.2. « Télécharger un nouveau texte : titre court ».

2.2.2 Annoté ?

Pour un texte de la bibliothèque, ce champ note s'il existe une version annoté par un expert.

2.2.3 Période

Le champ « Période » renseigne sur la date approximative d'un texte, afin de permettre de trouver plus aisément des textes de la même période. Chaque texte se voit assigner une période d'un demi-siècle. Lorsque la période de composition n'est connue que de manière approximative, ou lorsque la composition s'est étalée sur plusieurs années, la période la plus probable ou la plus significative a été retenue – ou sinon la période la plus ancienne. Ainsi, si l'on recherche des textes sur un certain nombre d'années, il est utile de rechercher également les périodes justes avant et justes après.

Remarque : il n'a pas été entrepris de vérification précise pour la datation des textes. À moins que ce ne soit spécifié, c'est la date utilisée dans l'édition qui a été reprise.

2.2.4 Niveaux d'accès

PALM contient des textes à trois niveaux d'accès : (3) « Utilisateur », qui peut être vu et utilisé par n'importe qui possédant un compte utilisateur ; (2) « Expert », qui peut seulement être vu par les utilisateurs avancés de PALM ; (1) « Administrateur », qui ne peut être vu que par l'administrateur système.

Du point de vue du simple utilisateur, seul les textes de niveau 3 apparaîtront dans la bibliothèque. Un utilisateur peut toutefois fixer le niveau d'accès de ses propres textes afin d'en restreindre l'accès aux autres utilisateurs du système (voir ci-dessous, paragraphe 3.2, « Télécharger un nouveau texte »).

2.2.5 Langue principale

Il s'agit de la langue de la majorité du texte, dans la mesure où les textes médiévaux contiennent souvent des phrases ou des passages entiers dans des langues différentes ou, dans des cas extrêmes, peuvent être en plusieurs langues.

2.2.6 Pays d'origine

Il s'agit, dans la mesure du possible, de donner l'origine d'un texte en français. Les termes 'France' et 'Angleterre', d'où proviennent la plupart des textes de la bibliothèque, sont employés pour les royaumes de la fin du Moyen Âge.

2.2.7 Auteur

Ce champ identifie les versions alternatives du nom d'un auteur, surtout quand celui-ci est connu dans différentes langues. Il identifie également les autres auteurs éventuels.

2.2.8 Édition

Ce champ a été conçu pour permettre à l'utilisateur d'identifier et de localiser l'édition (ou une autre source éventuelle, manuscrits compris) utilisée pour la numérisation. Le système de citation français a été suivi, bien que le titre et le nom de l'auteur soient cités dans la même langue que dans l'édition. En général, seul le lieu de publication est fourni, sans le nom de l'éditeur, sauf si davantage de précisions sont nécessaires pour identifier une édition. Pour des exemples de citations, voir ci-dessous, paragraphe 3.2.10 (« Télécharger un nouveau texte : Édition »).

2.2.9 Numérisé par…

Ce champ indique le nom de la personne (ou éventuellement des personnes) qui a numérisé, corrigé et téléchargé le texte dans PALM.

2.2.10 Date

Si le texte est daté, la date sera indiquée. Pour la forme employée, voir ci-dessous, paragraphe 3.2.13.

2.2.11 Notes

Ce champ est prévu pour les notes techniques ou les commentaires au cas où les autres champs ne suffisent pas (par exemple, sur la provenance d'un texte, sa datation, ou le manuscrit dont l'édition est issue).

2.2.12 Ajouter le texte

Après les informations sur le texte se trouve un bouton « Ajouter le texte » qui permet de le transférer dans l'espace de travail de l'utilisateur pour son traitement (par exemple, comme partie d'un corpus en vue d'une annotation ou d'une exportation (voir ci-dessous, chapitre 4).

2.3 Voir un texte de la bibliothèque

Il est possible d'accéder au texte lui-même en parcourant la bibliothèque, par un clic droit sur le titre qui donne accès au bouton « Voir le texte ». On peut ensuite parcourir le texte par courtes parties. L'objectif n'est pas d'offrir une édition numérique du texte, mais de permettre son examen avant de le sélectionner pour un traitement et une exportation.

2.4 Ajouter un texte de la bibliothèque à son espace de travail

Il y a plusieurs manières d'ajouter un texte de la bibliothèque à son espace de travail : soit faire un clic droit sur le texte à partir de l'écran « Consulter la bibliothèque » (menu « Bibliothèque ») et sélectionner « Ajouter un texte à votre espace de travail » ; soit cliquer sur le bouton « Ajouter un texte » en bas de l'écran « Détails d'un texte » ou de celui d'un texte lui-même.

Une fois le clic sur « Ajouter un texte » effectué, le texte est transféré dans l'espace de travail pour traitement, de la même manière que si un texte a été téléchargé

2.5 « Champs », « types de texte » et génération automatique de codes MEDITEXT

Lorsqu'un nouveau fichier texte est téléchargé dans PALM, un code est automatiquement généré, sur la base des informations fournies, dans l'ordre suivant :

Langue (En/Fr/La...) ; Prose ou vers ? (P/V) ; 'Champ' ; 'Type de texte' ; Siècle ; Pays d'origine (En/Fr/It...) ; Compteur

Cela produit un résultat comme par exemple : FrP20TMR14En17.

Le « champ » est un marqueur du contexte socio-littéraire de production d'un texte, tel que théorisé par Jean-Philippe Genet (à partir du concept de « champ » développé par Pierre Bourdieu) dans son ouvrage *La genèse de l'État moderne,* Paris, PUF, 2003. Les codes « Champ » utilisés dans PALM sont :

20 Religieux
21 Philosophique
22 Philologique (y compris l'enseignement, la rhétorique et la grammaire)
23 Scientifique
24 Médical
25 Littéraire
26 Juridique
27 Pratique (Vie quotidienne)
28 Musical
30 Administratif
50 Historique
60 Politique
00 Autres

Le « Type de texte » est une description plus artisanale du type de texte en question, très liée à la nature des sources présentes dans MEDITEXT.

AL Acte ou lettre
S Sermon
D Discours

HI Texte historique
TP Traité politique
TMR Traité moral ou religieux
PMR Poème moral ou religieux
PP Poème politique
HAG Hagiographie
TH Théâtre
TST Traité scientifique ou technique

Le compteur permet de distinguer entre les différents fichiers qui pourraient avoir des codes identiques. Il est entièrement arbitraire et ne marque en rien, par exemple, un ordre chronologique. Il marque simplement l'ordre dans lequel le fichier a été téléchargé dans PALM.

Quelques exemples :

FrP26AL15Fr26 : Une lettre de rémission émise en français et en France, en 1403.
FrP60AL13En1 : La version française de la Magna Carta.
EnV60PP15En27 : Le poème politique en moyen anglais « The World Upside Down ».
LaP60TP13Fr10 : Giles of Rome, De Regimine Principum, livre 1, partie 1.

3. Gérer son espace de travail

PALM permet de préparer des corpus de textes médiévaux pour leur traitement par des logiciels conçus pour des langues modernes, la plateforme offre également un système de gestion de corpus qui permet de construire un corpus prêt à l'exportation.

Pour accéder à ce système, sélectionner l'option « Gérer un corpus » à partir du menu « Espace de travail ». Si l'utilisateur n'a pas encore ajouté de texte à son espace de travail, par exemple à partir de la bibliothèque de PALM, il sera vide.

Une fois qu'un corpus a été construit dans l'espace de travail, il est possible de le lemmatiser en sélectionnant « Etiquetage morphosyntaxique » à partir du menu « Analyses linguistiques » (voire 4) avant de passer à l'exportation (voire 5).

3.1 Ajouter des textes à son espace de travail

Il est possible de créer un corpus soit en ajoutant des textes à son espace de travail à partir de la bibliothèque de PALM (voir paragraphe 2.4, « Ajouter un texte de la bibliothèque à son espace de travail »), soit en téléchargeant ses propres textes.

3.2 Télécharger un nouveau texte

Pour télécharger un texte vers son espace de travail dans PALM, sélectionner l'option « Ajouter un texte » à partir du menu « Espace de travail ». Un formulaire apparaît, demandant de compléter certaines informations sur le texte afin de faciliter sa recherche. Les champs suivants peuvent être complétés (les champs obligatoires sont marqués par un astérisque).

3.2.1 Vers/prose*

Le texte est-il en vers ou en prose ? Si le texte est un mélange des deux, sélectionner la forme majoritaire.

3.2.2 « Champ » et « Type de texte »*

Ces deux entrées servent à identifier la nature générale du texte.

Le « champ » du texte renvoie au contexte socio-littéraire de sa production (voir annexe 1).

Le « type de texte » est un système moins rigoureux de classification que le « champ » ; il vise à aider l'utilisateur à trouver des textes d'un type particulier (acte ou lettre, sermon, poème politique, etc.). Le « type de texte » ne prétend pas offrir un système universel de classification par genre, et les options offertes dérivent du type de corpus de textes politiques pour lesquels PALM a été conçu.

Ces deux entrées peuvent paraître subjectives. On pourrait par exemple discuter sur le fait que la *Cité de Dieu* d'Augustin est un texte religieux ou politique, par exemple. Nous souhaitons que chaque utilisateur agisse selon son jugement, pour aider de futurs utilisateurs à rassembler des textes similaires.

3.2.3 Titre*

Cette entrée correspond au titre court qui permet à l'utilisateur de trouver rapidement un texte. La langue par défaut est le français, sauf si le texte est connu sous un nom dans une langue différente. Si ce nom est seulement une convention éditoriale, une traduction française peut être suggérée entre parenthèses. Pour les auteurs bien connus en France, les noms sont donnés en français.

Remarque : les textes longs seront découpés en plusieurs fichiers plus courts. Quand cela est possible, on suit les subdivisions éditoriales ou auctoriales du texte, mais la division est parfois nécessairement arbitraire.

Quelques exemples :

Tractatus de regimine principum ad regem Henricum Sextum
On the Times [Sur les maux du temps]
Deux poèmes sur la mort de Piers Gaveston
Adam Orleton, Apologia (1/2)
Augustin d'Hippone, De Ciuitate Dei contra Paganos, Liber XV
John Russell, Sermon "In corpore multa quidem sunt membra…", 1484
John Kemp, Discours d'ouverture du Parlement, nov. 1450

3.2.4 Langue principale*

Si le texte comporte plus d'une langue, il est possible d'ajouter des langues supplémentaires dans un champ plus bas dans le formulaire. Dans cette rubrique, c'est la langue principale qui est mentionnée.

3.2.5 Pays d'origine*

Quand cela est possible, identifier le pays ou la région d'origine au moment de sa composition.

3.2.6 Période*

Un repère général du moment où le texte a été composé, en demi-siècles. Si le texte n'est pas daté précisément, choisir soit le demi-siècle le plus probable, ou le plus ancien. Si le texte a été composé sur plusieurs années, choisir la période la plus ancienne.

3.2.7 Niveau d'accès*

Tous les textes dans PALM se voient assignés un niveau d'accès qui s'appliquera si le texte est inclus dans la bibliothèque. Le niveau 3 (« Utilisateur ») correspond à un accès général : tous les utilisateurs peuvent le lire. Les textes de niveau 2 (« Expert ») peuvent seulement être vus par les experts accrédités. Seul l'administrateur de PALM (et l'utilisateur dans son espace de travail) pourra lire les textes de niveau 1 (« Administrateur »).

3.2.8 Type de texte

Une opportunité supplémentaire de spécifier la nature du texte, dans un champ ouvert plutôt que dans un menu formaté.

3.2.9 Auteur

L'auteur connu ou induit du texte. Cliquer sur « + » pour ajouter un auteur supplémentaire, où si l'auteur est connu sous différents noms dans différentes langues.

Pour les textes littéraires ou les textes dont on attendrait un auteur qui reste inconnu, il est possible de spécifier que le texte est « anonyme ». Cela n'est pas nécessaire pour les textes

composés par des institutions, pour lesquels les questions de paternité sont moins utiles. Dans ce cas, la rubrique peut ne pas être remplie.

Il est certain que ce n'est pas la pratique diplomatique habituelle, qui tend à identifier l'auteur comme la personne au nom de laquelle le document est émis, mais pour des raisons historiques, nous avons préférer éviter ce qui, pour nos textes, constitue souvent une identification trompeuse (le roi Jean comme auteur de la Magna Carta, Henri III comme auteur des déclarations de ses opposants baronniaux, etc.).

3.2.10 Édition*

Cette rubrique doit être complétée. Elle constitue, en quelque sorte, une note de bas de page, permettant à l'utilisateur d'identifier et de localiser l'édition utilisée. Les standards français de citation sont suivis, bien que le titre et le nom de l'auteur soient cités dans la même langue que dans l'édition. Normalement, seul le lieu de publication doit être complété, sauf si le nom de l'éditeur est nécessaire pour identifier une édition précise.

Si le titre d'un texte court ou d'un poème inclus dans un recueil est déjà donné dans le titre court, ce n'est pas la peine de le répéter, mais la pagination doit être fournie.
Quelques exemples (comme pour les « Détails » dans la bibliothèque de PALM) :

> Aegidius Romanus, De Regimine Principum, Rome, 1607.
> Rotuli Parliamentorum, éd. J. Strachey et al., Londres, 1767-77, vol. V, p. 16-17.
> The political songs of England : from the reign of John to Edward II, éd. et trad. Th. Wright, Londres, 1839, p. 258-261.
> Ptolomaeus lucensis [Bartholomeo Fiadoni], De Regimine Principum, dans Thomas Aquinas, Opuscula philosophica, éd. R.M. Spiazzi , Turin, 1954, p. 280-358.
> Londres, British Library, Royal MS 8.B.xxiii, ff. 9-10v.
> Lille, Archives du Nord, B 517/11679.
> Corpus Thomisticum <http://www.corpusthomisticum.org>.

Cliquer sur « + » pour ajouter des éditions multiples.

3.2.11 Langue

Une possibilité d'identifier une deuxième ou une troisième langue apparaissant dans le texte. Cliquer sur « + » pour ajouter davantage de langues.

3.2.12 Numérisé par…

Compléter par son nom et les noms de ceux qui ont été impliqués dans la numérisation du texte. Utiliser « + » pour ajouter des noms supplémentaires.

3.2.13 Date

Une possibilité d'identifier plus précisément une date. Les standards suivants doivent être suivis :

1467	Un texte daté clairement de 1467
[1467]	Un texte dont nous pouvons déduire qu'il a été composé en 1467
[?1467]	Un texte qui a peut-être été composé en 1467
[1215-1258]	Un texte qui a été composé durant la période 1215-1258
[avant 1327]	Un texte composé avant 1327
[après 1292]	Un texte composé après 1292
[c. 1340]	Vers (circa) 1340

3.2.14 Notes

Une possibilité d'ajouter des détails supplémentaires : notes techniques, par exemple, ou notes sur la nature particulière d'une édition ou d'un manuscrit complexe.

3.2.15 Texte

Couper-coller le texte dans cette rubrique.

Les textes doivent être copiés sans annotation à l'exception de la pagination. Cette dernière peut être insérée en usant soit le style <p=1>, <p=2> etc. ; ou le style <p=1|25> le 25 renvoyant à la page de l'édition.

Il faut ajouter au moins <p=1> au début du texte pour que le téléchargement fonctionne.

Les éléments insérés entre crochets [ ] ne seront pas pris en compte dans l'analyse automatique. Les crochets peuvent donc être employés pour insérer des commentaires dans le texte. Il faut souligner, néanmoins, que ces commentaires seront supprimés lorsque le texte sera soumis à l'opération d'étiquetage morphosyntaxique (voir ci-dessous, chapitre 4).

Quand le formulaire est complet et que le texte a été copié, cliquer sur le bouton « Télécharger ». Le texte sera alors téléchargé dans l'espace de travail. Un message apparaîtra confirmant que le texte a été correctement téléchargé.

3.3 Télécharger directement un texte

Les textes peuvent être téléchargés directement dans un format brut TXT. Pour sélectionner cette option, cliquer sur « Ajouter un texte par fichier » en haut du menu « Ajouter un texte ».

3.4 Gérer les textes dans son espace de travail

L'espace de travail peut être utilisé pour vérifier les détails d'un texte en cliquant sur « Gérer le corpus » dans le menu « Espace de travail ». L'écran de l'espace de travail permet d'accomplir un certain nombre d'actions sur les textes sélectionnés.

3.4.1 Détails du texte

Pour voir les détails d'un texte dans l'espace de travail, faire un clic droit sur le texte et sélectionner « Afficher les détails ».

3.4.2 Voir le texte

Un aperçu du texte peut être affiché, comme dans la bibliothèque du texte. Faire un clic droit sur le texte et sélectionner « Voir le texte ».

3.4.3 Modifier le texte

Si des erreurs sont détectées, soit dans les « Détails » du texte soit dans le texte lui-même, il est possible de les corriger en cliquant droit sur le texte dans l'espace de travail, puis en sélectionnant « Modifier le texte ».

Cette action affiche un écran similaire à celui utilisé pour télécharger un texte. Il est possible de changer les informations sur un texte, ou le texte lui-même, avant de le télécharger à nouveau dans l'espace de travail.

3.4.4 Supprimer le texte

Il est possible de supprimer le texte en faisant un clic droit sur un texte de l'espace de travail et de sélectionner « Supprimer le texte ». Attention : cette action est irréversible !

3.4.5 Ajouter un texte à la bibliothèque

Seuls les « experts » et l'administrateur peuvent transférer de nouveaux textes d'un espace de travail vers la bibliothèque de PALM en effectuant un clic droit sur le texte dans l'espace de travail et en sélectionnant « Ajouter à la bibliothèque ».

Pour les « Utilisateurs », il est possible de soumettre un texte pour inclusion en utilisant le même bouton. Une fois examiné, il pourra être inclus dans la bibliothèque.

Remarque : les utilisateurs sont invités, dans la mesure du possible, à rendre leurs textes disponibles pour la communauté de PALM-MEDITEXT. Cela permettra d'améliorer la qualité de l'annotation automatique proposée par PALM, et donc de réduire le temps passé sur les corrections manuelles.

4. Annoter un texte

La fonction principale de PALM est d'enrichir les textes par des annotations grammaticales et sémantiques afin de permettre l'exploration, par exemple, via des logiciels d'analyse textométrique.

Pour accomplir cette opération, il faut d'abord transférer les textes souhaités dans son espace de travail, puis, pour une annotation semi-automatique des textes, il faut cliquer sur « Etiquetage morphosyntaxique » dans le menu « Analyses linguistiques ».

Il est prévu d'intégrer à PALM un système de reconnaissance des entités nommées, pour le moment, toutefois, l'option « Entités nommées » dans le menu « Analyses linguistiques » n'est pas encore active.

4.1 La page du balisage morphosyntaxique

L'écran de l'étiquetage morphosyntaxique prend la même forme que celui de la gestion du corpus. Il liste les textes présents dans l'espace de travail par code, titre, langue, pays d'origine et période.

Pour lemmatiser un texte, effectuer un clic droit sur son titre et sélectionner « Etiquetage morphosyntaxique ». Il peut y avoir une courte attente après laquelle s'affichera le message : « L'analyse est en cours… Veuillez patienter… Cette opération peut prendre quelques minutes ». L'écran de l'annotateur de PALM s'affichera ensuite.

4.2 L'annotateur

Lorsqu'un texte est soumis à l'étiquetage morphosyntaxique, PALM applique un certain nombre de ressources linguistiques numériques (dictionnaires de formes-lemmes et « règles » écrites manuellement) afin d'identifier la description morphosyntaxique de chaque mot (occurrence) du texte.

Ces outils ont été développés pour un usage sur des textes de la fin du Moyen Âge (du milieu du XIII^{e} siècle au début du XVI^{e} siècle) en français originaires de la France et d'Angleterre. L'efficacité de ces outils s'accroit quand le texte soumis à analyse présente des similarités avec les corpus qui ont servi à leur développement.

L'annotateur permet d'accéder à la l'annotation automatique proposée par PALM et de la corriger manuellement quand il faut.

Chaque mot du texte est initiallement pourvu d'une couleur parmi trois possibles. Pour le « thème » par défaut de PALM, ces couleurs sont violet, jaune et rouge.

Quand un mot est coloré en violet, PALM estime que le mot a été correctement annoté. Pour s'assurer que cela est correct, il faut pointer le curseur sur le mot. La description morphosyntaxique assignée apparaîtra dans une petite fenêtre.

Quand un mot est coloré en jaune, PALM a identifié plusieurs descriptions morphosyntaxiques possibles, l'utilisateur peut alors effectuer un choix manuellement.

Quand un mot est coloré en rouge, PALM considére qu'il est inconnu et aucune description morphosyntaxique n'est attribuée.

4.3 Corriger un texte dans l'annotateur

C'est souvent au moment de l'annotation que l'utilisateur remarque des erreurs ou des incohérences dans le texte. Pour corriger un mot, faire un clic droit dessus et sélectionner « Modifier ». Pour ajouter un mot omis, faire un clic droit sur le mot suivant et sélectionner « Ajouter un nouveau mot ». Pour supprimer un mot, faire un clic droit et sélectionner « Supprimer ».

Cette option peut également être employée pour modifier la segmentation du texte en mots si l'utilisateur le souhaite (voir ci-dessous, paragraphe 4.7.4).

4.4 Corriger une annotation

Pour corriger une annotation, faire un clic gauche sur le mot. Une fenêtre s'ouvrira, dans laquelle on peut corriger le lemme et la partie du discours ou identifier un mot dans une autre langue.

En cliquant sur « Valider », le mot sélectionné sera corrigé. Les formes corrigés apparaissent en blanc dans l'annotateur.

Dans le cas de formes ambigües (en jaune), un choix de lemmes est proposé. Nous notons que si le lemme existe dans le *Dictionnaire du moyen français (DMF)*, il est possible de consulter la définition en cliquant sur le bouton « DMF ».

4.5 Corriger toutes les occurrences d'une forme

Il est également possible de corriger toutes les occurrences d'une forme dans un texte. En premier lieu, cliquer sur « Lancer les concordances ». Cela produira une concordance de toutes les occurrences de cette forme dans le texte. Si elles correspondent aux mêmes lemmes, cliquer d'abord sur une forme dans la conconrdance, et ensuite sur « Corriger tout ». Chaque occurrence de cette forme sera corrigée de la même manière.

4.6 Corriger à partir de la liste des formes les plus fréquentes

La même opération peut être effectuée à partir de la liste des fréquences à gauche de l'annotateur. Cliquer sur une forme de cette liste, puis lancer les concordances si cela est approprié et annoter les occurrences de la forme de la même manière.

4.7 Définition d'un lemme dans PALM

Un lemme est souvent défini comme la forme canonique d'un mot apparaissant dans un dictionnaire. Pour le moyen français, il n'y a pas de dictionnaire de référence pouvant être utilisé pour définir un lemme.

Le choix des lemmes utilisés dans PALM, et que l'utilisateur devrait également suivre pour de meilleurs résultats, doit donc être expliqué.

4.7.1 Lemmes en moyen français

Dans la mesure du possible, les lemmes utilisés par PALM correspondent à ceux de l'édition électronique du *Dictionnaire du moyen français* développé et mis à jour par le groupe de recherche ATILF du CNRS (UMR 7118) de l'université de Lorraine (Nancy). Le dictionnaire est disponible en ligne à l'adresse suivante : <http://www.atilf.fr/dmf/>.

Du point de vue de PALM, le DMF est une référence adéquate, à la fois parce qu'il couvre la période de notre corpus et parce qu'il distingue clairement les homonymes.

Le DMF utilise la forme en français moderne d'un lemme si elle est toujours en usage. Si le lemme n'est plus employé, le DMF utilise une forme attestée en moyen français. Étant donné que de nombreux textes de notre corpus sont d'origine anglaise, des lemmes n'ayant pas d'équivalent dans le DMF ont parfois été trouvé. Dans ces cas-là, nous avons utilisé l'*Anglo-Norman Dictionary,* qui peut être consulté en ligne : <http://www.anglo-norman.net/>.

Dans les très rares cas où nous n'avons trouvé d'équivalents dans aucun de ces ouvrages, nous avons créé nos propres lemmes sur la base de la forme attestée.

4.7.2 Note sur la segmentation des textes

Pour le moyen français, les scribes et les éditeurs ont choisi de diviser les mots de différentes manières.

Les choix que nous avons fait en annotant la bibliothèque de PALM, et donc ceux qui seront vraissemblablement proposés automatiquement par PALM, reflètent le souhait d'intervenir aussi peu que possible dans la correction des pratiques scribales, ou même des pratiques éditoriales.

Ainsi les mots composés dans les manuscrits sont subdivisés, ce qui entraîne quelques résultats discutables pour l'annotation automatique. Par exemple, en moyen français, l'adjectif « tresredouté » communément traité comme une unité dans les manuscrits et les éditions, est ségmenté en « tres » et « redouté ». Dans la mesure où, il n'existe pas un seul lemme qui peut être attribué à cette forme, il est donc annoté en deux formes disctinctes.

D'un autre côté, certains mots, en particulier les opérateurs logiques, sont parfois regroupés dans une seule unité, parfois séparés par un espace. Ainsi, par exemple, « toutefois » peuvent également s'écrire « toute fois ». Pour l'annotation automatique, il a été nécessaire d'analyser chaque mot dans un tel groupe séparément et ensuite une analyse lexicale est effectuée afin de traiter ces désagglutinations et l'annotée comme une seule forme.

Nous notons également que les formes qui sont liés par un trait d'union sont annotées et considérées comme des mots composés et ne seront pas divisées par notre système de segmentation. Si l'utilisateur souhaite contrôler la segmentation des mots composés, il peut donc intervenir en ajoutant des traits d'unions quand cela est adéquat avant de télécharger un texte. Cela vaut particulièrement la peine si l'utilisateur souhaite travailler sur certains mots composés ou une séquence des formes.

4.8 Définition des parties du discours dans PALM

Le jeu d'étiquettes proposé pour le moyen français a été composée de manière aussi simple que possible. Certains choix (que ce soit pour faire ou non la distinction entre les conjonction de subordination et de coordination, par exemple, ou entre les nombres ordinaux et cardinaux) ont été dictés par les ressources des dictionnaires qui ont servies de référence pour la définition d'un lemme (voir ci-dessus, section 4.7).

4.8.1 Parties du discours en moyen français

PALM propose les parties du discours suivantes pour le moyen français (voire l'annex B) :

Verbe
Adverbe
Nom commun
Nom propre
Pronom
Adjectif
Déterminant
Préposition
Conjonction of subordination
Conjonction of coordination
Nombre cardinal
Nombre ordinal
Interjection
Ponctuation

4.8.2 Note sur les noms propres

Les noms propres simples, en un mot (« John », « Paris ») sont étiquetés dans PALM avec la partie du discours « nom propre ». Toutefois, on trouve nombre d'expressions courtes (« Notre-Dame-de-Paris », « St Albans Abbey », « Stratford-atte-Bowe », « John of Salisbury ») qui renvoient à une entité nommée. Si ces mots appartiennent à notre dictionnaire des mots composés ou reconnues par notre système de reconnaissance des entités (qui sera bientôt intégré à PALM), chaque séquence de formes sera donc considéré comme une seule forme composé. Sinon, chaque séquence de formes sera segementée en plusieurs forme. Dans ce cas, par exemple « Notre Dame de Paris » sera ainsi analysé comme (adjectif) (nom commun) (préposition) (nom propre). Une telle analyse peut donc être modifiée manuellement par l'utilisateur s'il souhaite en ajoutant des traits d'unions afin de créer une unité simple. Ainsi, si le système rencontre « Notre-Dame-de-Paris », cette forme sera annotée comme une unité simple, avec le lemme « Notre-Dame-de-Paris ». L'utilisateur peut ainsi regruoper des séquences de formes.

Nous avons toutefois l'intention de mettre en place un système plus élaboré d'étiquetage pour les entités nommées et pour le regroupement des séquences de formes qui n'exigera pas ce traitement préliminaire.

4.9 Naviguer au sein du texte dans l'annotateur

Il est possible de naviguer dans le texte soit en cliquant sur les boutons « page précédente » et « page suivante », soit en sélectionnant une page spécifique.

4.10 Annoter un corpus

Une fois un texte annoté, sélectionner « Etiquetage morphosyntacique » dans le menu « Analyses linguistiques », ou en cliquant droit sur le texte souhaité, pour continuer avec les autres textes du corpus.

Si pour une raison quelconque on souhaite supprimer l'annotation d'un texte, faire un clic droit sur le texte et sélectionner « Supprimer l'étiquetage ». Attention : cette action est irréversible.

Une fois le corpus annoté de manière satisfaisante, on peut procéder à l'exportation.

5. Exportation

5.1 Exporter un corpus

Pour exporter un corpus, sélectionner l'option du menu « Exporter un corpus ».

L'écran d'exportation permet d'exporter des fichiers à partir de l'espace de travail, qu'ils soient lemmatisés ou non.

En premier lieu, sélectionner les textes en les tirant de la boite de gauche (l'espace de travail) vers la boite de droite (les fichiers à exporter). Une fois les textes sélectionner, cliquer sur le bouton « Options d'exportation » en bas de la page.

Il est possible de choisir le format du logiciel prévu pour l'exportation : NooJ, Lexico 3, Hyperbase, Tramer ou TXM. Il est également possible de créer un fichier texte brut en sélectionnant « Format TXT ».

On peut choisir d'exporter le corpus non-annoté ou le corpus enrichi par des annotations. Aussi, il est possible d'exporter un corpus partiellement annoté en sélectionnant des parties du discours en particulier. Il suffit donc de cliquer sur le bouton « Valider » pour confirmer le choix des options d'exportation et de cliquer ensuite sur le bouton « Exportation » pour exporter ce corpus. Il sera téléchargé en format compressé ZIP.

L'utilisateur ordinaire en aura terminé avec PALM, avec un corpus de textes annotés prêt à être utilisé dans un logiciel textométrique par exemple.

5.2 Note sur les formats d'exportation

5.2.1 TXM

Les fichiers exportés pour une utilisation avec TXM sont au format XML/w+CSV. Cela génère un fichier XML avec « Valeurs séparées par des virgules » que l'on peut importer dans TXM. Il faudra sélectionner ce format lors de l'importation.

Chaque dossier exporté de cette manière contient un court en-tête avec les informations sur le texte en question. Il est suivi par le corps du texte annoté dans un format XML adapté pour TXM. Chaque mot (occurrence) est annoté par lemme et par partie du discours.

Une fois importé dans TXM, les fichiers annotés dans PALM peuvent être explorés, utilisés pour générer les concordances et faire l'objet d'analyse textométrique (cooccurrences, spécificités, analyse factorielle…), soit par forme, soit par lemme – sachant qu'il est toujours possible de retourner à l'original.

Remarque : les « propriétés » annotées sont respectivement « lemme » (le lemme de ce mot), « explana » (un code pour la partie du discours) et « forme » (la forme particulière attestée).

Dans sa forme brute, cela ressemble à :

```
<?xml version="1.0" encoding="UTF-8"?>
<text titre="Common petition of the Good Parliament" langue="Français"
pays="Angleterre" periode="2éme moitié du XIVe siècle">
  <body>
    <pb n="331" />
    <p>
      <s />
      <s>
        <w lemma="de" explana="PREP" form="De">De</w>
        <w lemma="le" explana="DET" form="les">les</w>
        <w lemma="grand" explana="A" form="grant">grant</w>
        <w lemma="charte" explana="Nc" form="chartre">chartre</w>
        <w lemma="et" explana="CONJC" form="et">et</w>
        <w lemma="charte" explana="Nc" form="chartre">chartre</w>
        <w lemma="de" explana="PREP" form="de">de</w>
        <w lemma="le" explana="DET" form="la">la</w>
        <w lemma="foret" explana="Nc" form="forest">forest</w>
        <w lemma="." explana="SENT" form=".">.</w>
      </s>
```

… et ainsi de suite.

5.2.2 Lexico 3 et Hyperbase

Les fichiers exportés pour une utilisation dans Lexico 3 et Hyperbase peuvent être exportés dans leur format lemmatisé ou non. Les textes non lemmatisés sont simplement préparés pour l'importation dans le logiciel choisi, avec les balisages des titres et des numéros de page. Dans les textes lemmatisés, d'un autre côté, chaque mot (occurrence) est remplacé par son lemme dans le fichier exporté. Cela permet d'effectuer une analyse textométrique sur les lemmes des textes, mais les rendent très difficile à lire.

6. Gérer son compte

6.1 Modifier les paramètres de l'utilisateur

Pour changer ou modifier un nom, un email ou un « rôle », sélectionner « Mon compte » dans le menu « Espace de travail ». Cliquer sur le champ souhaité, corriger et cliquer sur « Valider ».

6.2 Modifier son mot de passe

Pour changer un mot de passe, sélectionner « Mon compte » dans le menu « Espace de travail ». En bas de la page, taper son ancien mot de passe, puis le nouveau deux fois pour le confirmer.

7. Administrer PALM et MEDITEXT

Le menu « Administration » est utilisable par les « Experts » et l'administrateur, c'est-à-dire ceux qui sont accrédités pour construire de nouvelles ressources linguistiques pour PALM.

7.1 Ajouter un nouvel utilisateur

Pour ajouter un nouvel utilisateur, sélectionner « Ajouter un nouvel utilisateur » dans le menu « Administration ».

Cette page permet d'entrer un nouvel utilisateur avec son nom, son email, son niveau d'accès, son identifiant et son mot de passe.

7.2 Gestion des comptes d'utilisateur

L'administrateur peut modifier les détails d'un compte en sélectionnant « Gestion des comptes d'utilisateur » dans le menu « Administration ». Cliquer sur un champ pour le modifier puis cliquer sur « Sauvegarder ».

7.3 Gestion de la bibliothèque

Un « Expert » ou l'administrateur peut gérer la bibliothèque de PALM en sélectionnant « Gérer la bibliothèque » dans le menu « Administration ».

Faire un clic droit sur un texte pour consulter les informations, voir le texte ou le supprimer.

Les textes de la bibliothèque ne peuvent être directement modifiés par l'administrateur. Ils doivent être téléchargés dans l'espace de travail, modifiés, puis re-téléchargés dans la bibliothèque.

Bibliographie

(Abney, 1991a) ABNEY, Steven P. Parsing by chunks. In : *Principle-based parsing*. Springer, Dordrecht, 1991. p. 257-278.

(Adda et al., 1997) DOLMAZON, Jean-Marc, BIMBOT, Frédéric, ADDA, Gilles, *et al.* Organisation de la premiere campagne AUPELF pour l'évaluation des systemes de dictée vocale. *Journées Scientifiques et Techniques du Réseau Francophone d'Ingénierie de la Langue de l'AUPELF-UREF*, 1997, p. 13-18.

(Adda et al., 1999) ADDA, Gilles, MARIANI, Joseph, PAROUBEK, Patrick, *et al.* L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Langues*, 1999, vol. 2, no 2, p. 119-129.

(Aouini, 2014) AOUINI, MOURAD. A NOOJ MODULE FOR NAMED ENTITY RECOGNITION IN MIDDLE FRENCH. *Formalising Natural Languages with Nooj 2014*, 2015, p. 99.

(Baccini et al, 2010) BACCINI, Alain. Statistique Descriptive Multidimensionnelle (pour les nuls). *Institut de Mathématiques de Toulouse-UMR CNRS*, 2010, vol. 5219.

(Baker, 1976) Baker J. K., Stochastic modeling as a means of automatic speech recognition, PhD Thesis, Carnegie-Mellon University, 1976.

(Bennett, 2007) BENNETT, Philip E. Anglo-Norman Dictionary. *The Modern Language Review*, 2007, vol. 102, no 2, p. 500-503.

(Bernard et al., 2002) BERNARD, Pascale, DENDIEN, Jacques, LECOMTE, Josette, *et al.* Un ensemble de ressources informatisées et intégrées pour l'étude du français: FRANTEXT, TLFi, Dictionnaires de l'Académie et logiciel Stella, présentation et apprentissage de leurs exploitations. In : *Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues*. 2002.

(Blackburn et al., 2001) BLACKBURN, Patrick, DE RIJKE, Maarten, et VENEMA, Yde. Modal logic, volume 53 of Cambridge tracts in theoretical computer science. 2001.

(Boileau, 1879) BOILEAU, Etienne. *Les métiers et corporations de la ville de Paris: xiiie siècle. Le livre des métiers*. Imprimerie nationale, 1879.

(Brugman et al., 2004) BRUGMAN, Hennie, RUSSEL, Albert, et NIJMEGEN, Xd. Annotating Multi-media/Multi-modal Resources with ELAN. In : *LREC*. 2004.

(Brunot, 1913) BRUNOT, Ferdinand et BRUNEAU, Charles. *Histoire de la langue française: des origines à 1900*. A. Colin, 1913.

(Burnard L., 1995) BURNARD, Lou. Users Reference Guide British National Corpus Version 1.0. 1995.

(Campedel & Hoogstoël, 2011) CAMPEDEL, Marine, HOOGSTOËL, Pierre. Sémantique et multimodalité en analyse de l'information. Traité RTA, série Recherche d'information et web, 2011

(Carolus-Barré, 1964) CAROLUS-BARRÉ, Louis (ed.). *Les Plus anciennes chartes en langue française...: Problèmes généraux et recueil des pièces originales conservées aux Archives de l'Oise, 1241-1286. Publié par Louis Carolus-Barré,.*. C. Klincksieck, 1964.

(Carreras & Màrquez, 2005) CARRERAS, Xavier et MÀRQUEZ, Lluís. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In : *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2005. p. 152-164.

(Chanod & Tapanainen, 1995) CHANOD, Jean-Pierre et TAPANAINEN, Pasi. Tagging French: comparing a statistical and a constraint-based method. In : *Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. Morgan Kaufmann Publishers Inc., 1995. p. 149-156.

(Chomsky & Miller, 1968) CHOMSKY, Noam et MILLER, George A. Introduction to the formal analysis of natural languages. 1968.

(Chomsky, 1957) LEES, Robert B. et CHOMSKY, Noam. Syntactic structures. *Language*, 1957, vol. 33, no 3 Part 1, p. 375-408.

(Chomsky, 1959) CHOMSKY, Noam. On certain formal properties of grammars. *Information and control*, 1959, vol. 2, no 2, p. 137-167.

(Chomsky, 1961) CHOMSKY, Noam. On the notion ‘rule of grammar’. In : *Proceedings of the Twelfth Symposium in Applied Mathematics*. American Mathematical Society, 1961. p. 6-24.

(Chomsky, 1965) CHOMSKY, Noam. Syntactic structures. The Hague: Mouton.. 1965. Aspects of the theory of syntax. *Cambridge, Mass.: MIT Press.(1981) Lectures on Government and Binding, Dordrecht: Foris.(1982) Some Concepts and Consequences of the Theory of Government and Binding. LI Monographs*, 1957, vol. 6, p. 1-52.

(Chomsky, 1975) CHOMSKY, Noam. *The logical structure of linguistic theory*. New York : Plenum press, 1975.

(Cleverdon et al, 1966) CLEVERDON, Cyril W., MILLS, Jack, et KEEN, Michael. Factors determining the performance of indexing systems. 1966.

(Coates-Stephens, 1993) COATES-STEPHENS, Sam. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities*, 1992, vol. 26, no 5, p. 441-456.

(Courtois & Silberztein, 1990) COURTOIS, Blandine et SILBERZTEIN, Max. Dictionnaires électroniques du français. *Langue française*, 1990, vol. 87, no 1, p. 3-4.

(Darmesteter et al., 1895) DARMESTETER, Arsène, SUDRE, Léopold, et MURET, Ernest. *Cours de grammaire historique de la langue française*. Librairie C. Delagrave, 1894-1898, v. 1, 1895.

(De Marneffe et al., 2006) DE MARNEFFE, Marie-Catherine, MACCARTNEY, Bill, MANNING, Christopher D., *et al.* Generating typed dependency parses from phrase structure parses. In : *Proceedings of LREC*. 2006. p. 449-454.

(Dees et al. 1980) DEES, Anthonij. *Atlas des formes et des constructions des chartes françaises du 13e siècle*. Walter de Gruyter, 1980.

(Dees et al. 1987) DEES, Anthonij. *Atlas des formes linguistiques des textes littéraires de l'ancien français*. M. Niemeyer, 1987.

(Dees, 1971) DEES, Anthonij. *Etude sur l'évolution des démonstratifs en ancien et en moyen français*. Wolters-Noordhoff Publ, 1971.

(Delbouille, 1939) DELBOUILLE, Maurice. A propos de l'infinitif historique dans les langues romanes. *Revue belge de philologie et d'histoire*, 1939, vol. 18, no 2, p. 625-639.

(Di Marneffe & Manning, 2008) DE MARNEFFE, Marie-Catherine et MANNING, Christopher D. The Stanford typed dependencies representation. In : *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*. Association for Computational Linguistics, 2008. p. 1-8.

(Doddington et al. 2004) DODDINGTON, George R., MITCHELL, Alexis, PRZYBOCKI, Mark A., *et al.* The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In : *LREC*. 2004. p. 837-840.

(Dupont & Tellier, 2014) DUPONT, Yoann et TELLIER, Isabelle. Un reconnaisseur d'entités nommées du Français. *Jean-Marie Pierrel [P-Demo1. 2] Utilisabilité d'une ressource propriétaire riche dans le cadre de la classification de documents*, 2014, p. 40.

(Eherman, 2008) EHRMANN, Maud. *Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguïsation*. 2008. Thèse de doctorat. Paris 7.

(Eshkol et al., 2010) ESHKOL, Iris, TELLIER, Isabelle, SAMER, Taalab, *et al.* Etiqueter un corpus oral par apprentissage automatique\a l'aide de connaissances linguistiques. *arXiv preprint arXiv:1003.5749*, 2010.

(Farkas et al., 2010) FARKAS, Richárd, VINCZE, Veronika, MÓRA, György, *et al.* The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In : *Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*. Association for Computational Linguistics, 2010. p. 1-12.

(Febvre, 1953) FEBVRE, Lucien Paul Victor. Combats pour l'histoire. 1953.

(Fletcher, 2013) C. Fletcher, 'Les mots et les choses dans l'historiographie du parlement anglais de la fin du Moyen Âge', in Consensus et représentation : Actes du colloque tenu à Dijon, le 14-16 mars 2013, dir. J.-P. Genet et al., forthcoming.

(Fletcher, 2014) C. Fletcher, 'What makes a political language? Key terms, profit and damage in the Common Petition of the English Parliament, 1343-1422' in The Voices of the People in Late Medieval Europe: Communication and Popular Politics, ed. Jan Dumolyn, Jelle Haemers, H.R. Oliva Herrer and Vincent Challet. Studies in European Urban History 33. Turnhout: Brepols, 2014, pp. 91-106.

(Fleury, 2007) FLEURY, Serge. Le Trameur, Manuel d'utilisation. 2007.

(Fleury, 2009) FLEURY, Serge. Programmation et projet encadré. Université Sorbonne Nouvelle-Paris 3-M. FLEURY Serge. 2009.

(Foucault, 1966) MICHEL, FOUCAUlt. Les Mots et les Choses-Une archéologie des sciences humaines (1966). *Paris, Gallimard, coll.«Tel*, 1998.

(Fouché, 1967) FOUCHÉ, Pierre. *Le Verbe français: étude morphologique.*. Klincksieck, 1967.

(Francopoulo et al., 20013) FRANCOPOULO, Gil (ed.). *LMF Lexical Markup Framework*. John Wiley & Sons, 2013.

(Francopoulo et al., 2006) FRANCOPOULO, Gil, GEORGE, Monte, CALZOLARI, Nicoletta, *et al.* Lexical markup framework (LMF). In : *International Conference on Language Resources and Evaluation-LREC 2006*. 2006.

(Friburger et al., 1993) FRIBURGER, Nathalie. Linguistique et reconnaissance automatique des noms propres. *Meta: Journal des traducteurs/Meta: Translators' Journal*, 2006, vol. 51, no 4, p. 637-650.

(Friburger, 2002) FRIBURGER, Nathalie. *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*. 2002. Thèse de doctorat. Tours.

(Fuchs & Victorri, 1993) FUCHS, Catherine, VICTORRI, Bernard. Sémantique. In Fuchs, C. (ed.), Linguistique et traitement automatique des langues, Supérieur, pages 139-169. Hachette, Paris, 1993.

(Fuchs, 1993) FUCHS, Catherine, LACHERET-DUJOUR, Anne, VICTORRI, Bernard, *et al.* Linguistique et Traitement automatiques des Langues. *Hachette université langue, linguistique, communication*, 1993.

(Gahbiche-Braham, 2013) GAHBICHE-BRAHAM, Souhir, BONNEAU-MAYNARD, Hélene, LAVERGNE, Thomas, *et al.* Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier. In : *LREC*. 2012. p. 2107-2113.

(Galliano et al., 2009) GALLIANO, Sylvain, GRAVIER, Guillaume, et CHAUBARD, Laura. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In : *Tenth Annual Conference of the International Speech Communication Association*. 2009.

(Gazdar, 1985) GAZDAR, Gerald. *Generalized phrase structure grammar*. Harvard University Press, 1985.

(Gazdar, 1988) GAZDAR, Gerald. Applicability of indexed grammars to natural languages. In : *Natural language parsing and linguistic theories*. Springer Netherlands, 1988. p. 69-94.

(Genet et al., 2011) GENET, Jean-Philippe, BERTRAND, Jean-Marie, *et al. Langue et histoire: actes du Colloque de l'École Doctorale d'Histoire de Paris 1, INHA, 20 et 21 octobre 2006*. Publications de la Sorbonne, 2011.

(Genet, 2003) GENET, Jean-Philippe. *La genèse de l'État moderne: Culture et société politique en Anglaterre*. Presses Universitaires de France-PUF, 2003.

(Genet, 2010) GENET, Jean-Philippe. Identité, espace, langue. *Cahiers de recherches médiévales et humanistes. Journal of medieval and humanistic studies*, 2010, no 19, p. 1-10.

(Gerdes & Kahane, 2017) GERDES, Kim, KAHANE, Sylvain. *Trois schémas d'annotation syntaxique en dépendance pour un même corpus de français oral : le cas de la macrosyntaxe.* Actes de la 24e conférence sur le traitement automatique des langues (TALN), Atelier sur les corpus annotés du français (ACor4French), Orléans, 2017, p. 9.

(Gossen C, 1968) GOSSEN, Charles Théodore. *L'interprétation des graphèmes et la phonétique historique de la langue française*. Centre de philologie et de littératures romanes de Strasbourg, 1968.

(Gossen, 1968) GOSSEN, Charles Théodore. *L'interprétation des graphèmes et la phonétique historique de la langue française*. Centre de philologie et de littératures romanes de Strasbourg, 1968.

(Green & Manning, 2010) GREEN, Spence et MANNING, Christopher D. Better Arabic parsing: Baselines, evaluations, and analysis. In : *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010. p. 394-402.

(Grice, 1979), GRICE, H. Paul. Logique et conversation. *Communications*, 1979, vol. 30, no 1, p. 57-72.

(Grishman & Sundheim, 1996) GRISHMAN, Ralph et SUNDHEIM, Beth. Message understanding conference-6: A brief history. In : *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. 1996.

(Gross, 1997) GROSS, Maurice. 1 The Construction of Local Grammars. *Finite-state language processing*, 1997, p. 329.

(Guillot et al., 2007) GUILLOT, Céline, LAVRENTIEV, Alexei, et MARCHELLO-NIZIA, Christiane. La Base de Français Médiéval (BFM): états et perspectives. 2007.

(Guiraud, 1963) GUIRAUD, Pierre. *Le moyen français*. Presses universitaires de France, 1963.

(Güngör, 2010) GÜNGÖR, Tunga. Part-of-Speech Tagging. 2010.

(Habert et al., 1997) HABERT, Benoît, NAZARENKO, Adeline, et SALEM, André. Les linquistiques de corpus. Armand Colin. 1997.

(Harris, 1951) HARRIS, Zellig Sabbettai. Methods in structural linguistics. 1951.

(Harris, 1954) HARRIS, Zellig S. Distributional structure. *Word*, 1954, vol. 10, no 2-3, p. 146-162.

(Harrison, 1978) HARRISON, Michael A. *Introduction to formal language theory*. Addison-Wesley Longman Publishing Co., Inc., 1978.

(Heiden, 2004) HEIDEN, Serge et LAVRENTIEV, Alexei. Ressources électroniques pour l'étude des textes médiévaux: approches et outils. *Revue française de linguistique appliquée*, 2004, vol. 9, no 1, p. 99-118.

(Hin et al., 2012) HINTON, Geoffrey E., SRIVASTAVA, Nitish, KRIZHEVSKY, Alex, *et al.* Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

(Hong, 2017) HONG, Yong-Jin. Langue et politique à la cour de Philippe VI et de Jean II. *Traduction et culture en France et en îles Britanniques*. Edité par Jean-Philippe Genet, Paris (Garnier), 2017, p.52-91.

(Hopcroft et al, 2001) HOPCROFT, John E., MOTWANI, Rajeev, et ULLMAN, Jeffrey D. Introduction to automata theory, languages, and computation. *ACM SIGACT News*, 2001, vol. 32, no 1, p. 60-65.

(Hudson, 2000) Hudson R., Dependency Grammar, Essli Summer School, University of Birmingham, UK, 2000

(Huizinga & Le Goff, 1989) HUIZINGA, Johan et LE GOFF, Jacques. *L'automne du Moyen Age*. Payot, 1989.

(Huizinga et al., 1967) HUIZINGA, Johan, BASTIN, Julia, et HANOTAUX, Gabriel. *Le déclin du moyen âge*. 1967.

(Iberkwe-SanJuan, 2007) IBERKWE-SANJUAN, F. Fouille de textes methods, outils et apllications. *Expert System With Applications*, 2007.

(Ide & Romary, 2004) IDE, Nancy et ROMARY, Laurent. A registry of standard data categories for linguistic annotation. In : *4th International Conference on Language Resources and Evaluation-LREC'04*. 2004. p. 135-138.

(Ide & Veronis, 1994) IDE, Nancy et VÉRONIS, Jean. MULTEXT: Multilingual text tools and corpora. In : *Proceedings of the 15th conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 1994. p. 588-592.

(Imbs P., 1971) IMBS, Paul (ed.). *Trésor de la langue française: Lot-Natalité*. Éditions du Centre national de la recherche scientifique, 1971.

(Isozaki et Kazawa, 2002) ISOZAKI, Hideki et KAZAWA, Hideto. Efficient support vector classifiers for named entity recognition. In : *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002. p. 1-7.

(Jean-Marie, 2000) JEAN-MARIE, Pierrel. Ingénierie des langues. *Paris: Hermes*, 2000.

(Jejcic et Fondet, 2011) FONDET, Claire et JEJCIC, Fabrice. OrthoFonic: un projet de didacticiel pour l'apprentissage de l'orthographe française. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 2011, vol. 54, p. 73-92.

(Joshi, 1987) JOSHI, Aravind K. An introduction to tree adjoining grammars. *Mathematics of language*, 1987, vol. 1, p. 87-115.

(Kaplan & Bresnan, 1982) KAPLAN, Ronald M. et BRESNAN, Joan. Lexical-functional grammar: A formal system for grammatical representation. *Formal Issues in Lexical-Functional Grammar*, 1982, p. 29-130.

(Karttunen et al, 1997) KARTTUNEN, Lauri, GAÁL, Tamás, et KEMPE, André. Xerox finite-state tool. *Rapport technique, Centre de recherche Xerox de Grenoble*, 1997.

(Karutten et al, 2003) BEESLEY, Kenneth R. et KARTTUNEN, Lauri. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*, 2003.

(Kilgarriff, 2000) KILGARRIFF, Adam. Wordnet: An electronic lexical database. 2000.

(Klein & Simpson, 1963) KLEIN, Sheldon et SIMMONS, Robert F. A computational approach to grammatical coding of English words. *Journal of the ACM (JACM)*, 1963, vol. 10, no 3, p. 334-347.

(König et al., 2003) KÖNIG, Esther, LEZIUS, Wolfgang, et VOORMANN, Holger. Tigersearch 2.1 user's manual. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 2003.

(Kripke, 1972) KRIPKE, Saul A. Naming and necessity. In : *Semantics of natural language*. Springer Netherlands, 1972. p. 253-355.

(Kucera & Francis, 1967) KUČERA, Henry et FRANCIS, Winthrop Nelson. *Computational analysis of present-day American English*. Dartmouth Publishing Group, 1967.

(Kurdi, 2017) KURDI, Mohamed Zakaria. *Traitement automatique des langues et linguistique informatique 1*. ISTE editions, 2017.

(Lafferty et al.,2001) LAFFERTY, John, MCCALLUM, Andrew, et PEREIRA, Fernando CN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

(Lanly, 1977) LANLY, André. *Morphologie historique des verbes français: notions générales, conjugaisons régulieères, verbes irréguliers*. Bordas, 1977.

(Le Meur et al., 2004) LE MEUR, Céline, GALLIANO, Sylvain, et GEOFFROIS, Edouard. Conventions d'annotations en Entités Nommées-ESTER. *Rapport technique de la campagne Ester*, 2004.

(Lebart & Salem, 1994) LEBART, Ludovic et SALEM, André. Statistique textuelle. *Paris: Dunod,| c1994*, 1994.

(Lezius et al., 2002) LEZIUS, Wolfgang, BIESINGER, Hannes, et GERSTENBERGER, Ciprian. TIGER-XML quick reference guide. *IMS, University of Stuttgart*, 2002.

(Ligozat, 1994) LIGOZAT, Gérard. *Représentation des connaissances et linguistique*. Colin, 1994.

(Luce et al., 1858) LUCE, Siméon, RAYNAUD, Gaston, MIROT, Léon, *et al.* Chroniques de J. Froissart. *Paris: SHF*, 1868, vol. 1975, p. 60-141.

(Lusignan, 2012) LUSIGNAN, Serge. *Essai d'histoire sociolinguistique: le français picard au Moyen Âge*. Classiques Garnier, 2012.

(Macdonald et al., 2013) MCDONALD, Ryan T., NIVRE, Joakim, QUIRMBACH-BRUNDAGE, Yvonne, *et al.* Universal Dependency Annotation for Multilingual Parsing. In : *ACL (2)*. 2013. p. 92-97.

(Makhoul et al. 1999) MAKHOUL, John, KUBALA, Francis, SCHWARTZ, Richard, *et al.* Performance measures for information extraction. In : *Proceedings of DARPA broadcast news workshop*. 1999. p. 249-252.

(Manning & Schutze, 1999) MANNING, Christopher D., SCHÜTZE, Hinrich, *et al. Foundations of statistical natural language processing*. Cambridge : MIT press, 1999.

(Marchello- Nizia, 1979) MARCHELLO-NIZIA, Christiane. *Histoire de la langue française aux XIVe et XVe siècles*. Bordas, 1979.

(Marchello-Nizia, 1999) MARCHELLO-NIZIA, Christiane. *Le français en diachronie: douze siècles d'évolution*. Éditions Ophrys, 1999.

(Marchello-Nizia, 2005) MARCHELLO-NIZIA, Christiane. *La langue française aux XIVe et XVe siècles*. Armand Colin, 2005.

(Marconi, 1997) MARCONI, Diego. *La philosophie du langage au vingtième siècle*. Éditions de l'éclat, 1997.

(Martin & Bazin-Tacchella, 2012) MARTIN, Robert et BAZIN-TACCHELLA, Sylvie. Dictionnaire du moyen français (DMF2012). 2012.

(Martineau et al., 2009) MARTINEAU, France, *et al.* Le corpus MCVF. Modéliser le changement: les voies du français. 2009.

(Martineau et al., 2010) MARTINEAU, F., HIRSCHBÜHLER, P., KROCH, A., *et al.* Corpus MCVF (parsed corpus), Modéliser le changement: les voies du français, Département de français, University of Ottawa. *CD-ROM, first edition (http://www. arts. uottawa. ca/voies/voies_fr. html)*, 2010.

(Martineau, 2009) MARTINEAU, F. Le corpus MCVF. *Modéliser le changement: les voies du français. Université d'Ottawa, Ottawa*, 2009.

(Mazziotta, 2010) MAZZIOTTA, Nicolas. Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples. *Recherches qualitatives. Hors-série*, 2010, vol. 9, p. 83-94.

(Mazziotta, 2012) MAZZIOTTA, Nicolas. Le _Syntactic Reference Corpus of Medieval French_: Structure, outils et exploitation. *11es Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 2012, p. 701-713.

(Mazziotta, 2013) MAZZIOTTA, Nicolas. Traitement de la coordination dans le Syntactic Reference Corpus of Medieval French (SRCMF). *Actes du XXVIe Congrès de Linguistique et Philologie Romanes*, 2013, p. t. 7, 229-238.

(McDonald, 1996) MCDONALD, David D. Internal and external evidence in the identification and semantic categorization of proper names, Corpus processing for lexical acquisition. 1996.

(Merisalo, 1988) MERISALO, Outi. La langue et les scribes. Etude sur les documents en langue vulgaire de la Rochelle, Loudun, Châtellerault et Mirebeau au XIIIe siècle. *Commentationes humanarum litterarum*, 1988, vol. 87, p. 7-336.

(Mesfar, 2008) MESFAR, Slim. *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*. 2008. Thèse de doctorat. Besançon.

(Mikheev et al., 1998) MIKHEEV, Andrei, GROVER, Claire, et MOENS, Marc. Description of the LTG system used for MUC-7. In : *Proceedings of 7th Message Understanding Conference (MUC-7)*. Fairfax, VA, 1998. p. 1-12.

(Mitchell, 1997) MITCHELL, Tom M., *et al.* Machine learning. WCB. 1997.

(Moeschler, 1997) MOESCHLER, Jacques et AUCHLIN, Antoine. Introduction a la linguistique contemporaine. Paris: Armand Colin. 1997.

(Nivre et al., 2016) NIVRE, Joakim, DE MARNEFFE, Marie-Catherine, GINTER, Filip, *et al.* Universal Dependencies v1: A Multilingual Treebank Collection. In : *LREC*. 2016.

(Nouvel et al., 2015) NOUVEL, Damien, EHRMANN, Maud, et ROSSET, Sophie. *Les entités nommées pour le traitement automatique des langues*. ISTE editions, 2015.

(Nouvel, 2012) NOUVEL, Damien. *Reconnaissance des entités nommées par exploration de règles d'annotation-Interpréter les marqueurs d'annotation comme instructions de structuration locale*. 2012. Thèse de doctorat. Université François Rabelais-Tours.

(Paroubek & Rajman, 2000) PAROUBEK, Patrick et RAJMAN, Martin. Etiquetage morpho-syntaxique. *Ingénierie des langues*, 2000, p. 131-150.

(Petersen et al., 2005) PETERSEN, Ulrik. Evaluating corpus query systems on functionality and speed: TIGERSearch and Emdros. In : *International Conference Recent Advances in Natural Language Processing*. Incoma, Ltd., 2005. p. 387-391.

(Pierrel, 2000) PIERREL, Jean-Marie. *Ingénierie des langues*. Hermes, 2000.

(Pioche & Marchello-Nizia, 1989) PICOCHE, Jacqueline et MARCHELLO-NIZIA, Christiane. *Histoire de la langue française*. 1989.

(Poibeau, 2003) POIBEAU, Thierry. The multilingual named entity recognition framework. In : *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 2*. Association for Computational Linguistics, 2003. p. 155-158.

(Polar & Sag, 1994) POLLARD, Carl et SAG, Ivan A. *Head-driven phrase structure grammar*. University of Chicago Press, 1994.

(Pollard, 1984) POLLARD, Carl. Generalized context-free grammars, head grammars and natural language. *Standord: Stanford University dissertation*, 1984.

(Prévost & Heiden, 2002) PRÉVOST, Sophie et HEIDEN, Serge. Etiquetage d'un corpus hétérogène de français médiéval: enjeux et modalités. *Romanistische Korpuslingustik: Korpora und gesprochene Sprache (Romance Corpus Linguistics: Corpora and Spoken Language). Tübingen: Gunter Narr Verlag*, 2002, p. 127-136.

(Prévost & Stein, 2012) PRÉVOST, Sophie et STEIN, Achim. Syntactic Reference Corpus of Medieval French et l'ordre des compléments du verbe en ancien français. 2012.

(Prévost & Stein, 2013) PRÉVOST, Sophie et STEIN, Achim. Syntactic Reference Corpus of Medieval French (SRCMF). *Lyon/Stuttgart: ENS de Lyon*, 2013.

(Pustet, 2003) PUSTET, Regina. *Copulas: Universals in the Categorization of the Lexicon*. OUP Oxford, 2003.

(Reenen & Mulder, 2000) REENEN, Pieter van et MULDER, Maaike. Un corpus linguistique de 3000 chartes en Moyen Néerlandais du 14e siècle. *Corpus: Méthodologie et applications linguistiques*, 2000, p. 209-217.

(Remacle, 1948) REMACLE, Louis. *La structure interne du wallon et l'influence germanique*. Michels, 1948.

(René et al., 1991) RENÉ, Carré, DÉGREMONT, Jean-François, GROSS, Maurice, *et al.* Langage humain et machine. 1991.

(Rosset et al. 2011b) ROSSET, Sophie, GROUIN, Cyril, et ZWEIGENBAUM, Pierre. *Entités nommées structurées: guide d'annotation Quaero*. LIMSI-Centre national de la recherche scientifique, 2011.

(Roth & YIH, 2002) ROTH, Dan et YIH, Wen-tau. Probabilistic reasoning for entity & relation recognition. In : *Proceedings of the 19th international conference on Computational linguistics-Volume 1*. Association for Computational Linguistics, 2002. p. 1-7.

(Salem, 1986) SALEM, André. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, 1986, p. 5-28.

(Samuelsson & Voutilainen, 1997) SAMUELSSON, Christer et VOUTILAINEN, Atro. Comparing a linguistic and a stochastic tagger. In : *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997. p. 246-253.

(Schmid, 2013) SCHMID, Helmut. Probabilistic part-ofispeech tagging using decision trees. In : *New methods in language processing*. 2013. p. 154.

(Schmidt, 2004) SCHMIDT, Thomas. Transcribing and annotating spoken language with EXMARaLDA. In : *Proceedings of the LREC-Workshop on XML based richly annotated corpora, Lisbon 2004*. 2004.

(Seraj et al., 2017) SERAJI, Mojgan, GINTER, Filip, et NIVRE, Joakim. Universal Dependencies for Persian. In : *LREC*. 2016.

(Silberztein, 1993a) SILBERZTEIN, Max. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Masson, 1993.

(Silberztein, 2005) SILBERZTEIN, Max. NooJ's dictionaries. *Proceedings of LTC*, 2005, vol. 5, p. 291-295.

(Silberztein, 2015) SILBERZTEIN, Max. *La formalisation des langues: l'approche NooJ*. ISTE éd., 2015.

(Siouffi & Van Raemdonck, 2009) SIOUFFI, Gilles et VAN RAEMDONCK, Dan. 100 fiches pour comprendre la linguistique. 1999.

(Sleator et Temperley, 1991) SLEATOR, D. et TEMPERLEY, D. *Parsing with a link grammar*. Technical Report CMU-CS-91-196, School of CS, Carnegie Mellon University, Pittsburgh PA, 1991.

(Souvay & Pierrel, 2009) SOUVAY, Gilles et PIERREL, Jean-Marie. LGeRM Lemmatisation des mots en moyen français. *Traitement Automatique des Langues*, 2009, vol. 50, no 2, p. 21.

(Souvay, 2004) SOUVAY, Gilles. Vers un dictionnaire électronique du moyen français. In : *Actes du Colloque Euralex 2004, European Association for Lexicography congress*. 2004. p. 671-678.

(Taji et al., 2017) TAJI, Dima, HABASH, Nizar, et ZEMAN, Daniel. Universal Dependencies for Arabic. *WANLP 2017 (co-located with EACL 2017)*, 2017, p. 166.

(Taylor, 2003) TAYLOR, John R. *Linguistic categorization*. Oxford University Press, 2003.

(Tellier, 2010) TELLIER, Isabelle. Introduction au TALN et à l'ingénierie linguistique. *Polycopié de cours: Univer-sité de Lille*, 2010, vol. 3.

(Tesnière, 1959) LUCIEN, Tesnière. Eléments de syntaxe structurale. *Paris, Klincksieck*, 1959.

(Tian et al., 2015) TIAN, Tian, DINARELLI, Marco, TELLIER, Isabelle, *et al.* Etiquetage morpho-syntaxique de tweets avec des CRF. In : *TALN 2015*. 2015.

(Touratier, 2010) TOURATIER, Christian. *La sémantique*. Armand Colin, 2010.

(Tran, 2006) TRAN, Mickaël. *Prolexbase: un dictionnaire relationnel multilingue de noms propre: conception, implémentation et gestion en ligne*. 2006. Thèse de doctorat. Tours.

(Turing, 1950) Turing A., « Computing Machinery and Intelligence », Mind, vol. 49, p. 433-460, 1950

(Tzourkermann et al., 1996) TZOUKERMANN, Evelyne et RADEV, Dragomir R. Using word class for part-of-speech disambiguation. In : *Proceedings of the Fourth Workshop on Very Large Corpora*. 1996. p. 1-13.

(Van Reenen & Schoesler, 2000) VAN REENEN, Pieter et SCHOESLER, Lene. Corpus et stemma en ancien et en moyen français. *Le moyen français: le traitement du texte*, 2000, p. 25-54.

(Véronis et al., 1995) VÉRONIS, Jean et KHOURI, Liliane. Étiquetage grammatical multilingue: le projet MULTEXT. *TAL. Traitement automatique des langues*, 1995, vol. 36, no 1-2, p. 233-248.

(Véronis, 2000) VÉRONIS, Jean. Annotation automatique de corpus: panorama et état de la technique. *Ingénierie des langues*, 2000, vol. 4.

(Weizenbaum, 1966) WEIZENBAUM, Joseph. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 1966, vol. 9, no 1, p. 36-45.

(Yvon, 2006) YVON, François. Des apprentis pour le traitement automatique des langues. *HDR, Université Pierre et Marie Curie*, 2006.

(Zipf, 1930) Zipf, K. (1930). Arch. exp. Path. Pharmak. 157. 95.

ÉD LETS

Titre : Approche multi-niveaux pour l'analyse des données textuelles non-standardisées : corpus de textes en moyen français

Mots clés : approche multi-niveaux, données textuelles non-standardisées, moyen français, étiquetage morphosyntaxique, reconnaissance des entités nommées.

Résumé : Cette thèse présente une approche d'analyse des textes non-standardisé qui consiste à modéliser une chaine de traitement permettant l'annotation automatique de textes à savoir l'annotation grammaticale en utilisant une méthode d'étiquetage morphosyntaxique et l'annotation sémantique en mettant en œuvre un système de reconnaissance des entités nommées. Dans ce contexte, nous présentons un système d'analyse du moyen français qui est une langue en pleine évolution dont l'orthographe, le système flexionnel et la syntaxe ne sont pas stables. Les textes en moyen français se singularisent principalement par l'absence d'orthographe normalisée et par la variabilité tant géographique que chronologique des lexiques médiévaux.
L'objectif est de mettre en évidence un système dédié à la construction de ressources linguistiques, notamment la construction des dictionnaires électroniques, se basant sur des règles de morphologie. Ensuite, nous présenterons les instructions que nous avons établies pour construire un étiqueteur morphosyntaxique qui vise à produire automatiquement des analyses contextuelles à l'aide de grammaires de désambiguïsation. Finalement, nous retracerons le chemin qui nous a conduits à mettre en place des grammaires locales permettant de retrouver les entités nommées. De ce fait, nous avons été amenés à constituer un corpus MEDITEXT regroupant des textes en moyen français apparus de la fin du XIIIème à la fin XVème siècle.

Title: Multi-level approach for the analysis of non-standardized textual data: corpus of texts in Middle French

Keywords: Multi-level approach, non-standardized textual data, Middle French, morphosyntactic tagging, named-entity recognition.

Abstract: This thesis presents a non-standardized text analysis approach which consists a chain process modeling allowing the automatic annotation of texts: grammar annotation using a morphosyntactic tagging method and semantic annotation by putting in operates a system of named-entity recognition. In this context, we present a system analysis of the Middle French which is a language in the course of evolution including: spelling, the flexional system and the syntax are not stable. The texts in Middle French are mainly distinguished by the absence of normalized orthography and the geographical and chronological variability of medieval lexicons.
The main objective is to highlight a system dedicated to the construction of linguistic resources, in particular the construction of electronic dictionaries, based on rules of morphology. Then, we will present the instructions that we have carried out to construct a morphosyntactic tagging which aims at automatically producing contextual analyzes using the disambiguation grammars. Finally, we will retrace the path that led us to set up local grammars to find the named entities. Hence, we were asked to create a MEDITEXT corpus of texts in Middle French between the end of the thirteenth and fifteenth centuries.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon