UCL Université catholique de Louvain

Faculté de philosophie, arts et lettres (FIAL)

FIAL
Un éventail de formations, une multitude d'horizons

Traitement automatique du moyen français

Analyse de données, étiquetage morphosyntaxique et désambiguïsation contextuelle

Mémoire réalisé par
Hugo LEEMANS

Promoteurs
Mattia CAVAGNA, Cédrick FAIRON

Année académique 2017-2018

Master en linguistique
Finalité: Traitement automatique du langage

Mémoire de master

Traitement automatique du moyen français

Analyse de données, étiquetage morphosyntaxique et désambiguïsation contextuelle

Hugo Leemans

Promoteurs :
Prof. Mattia Cavagna
Prof. Cédrick Fairon

Août 2018

Remerciements

Je souhaite remercier toutes les personnes qui, de près ou de loin, ont contribué à ce que je parvienne au bout de ce travail.

A tout seigneur, tout honneur. Je dois remercier mes promoteurs de m'avoir proposé ce projet et de m'avoir laissé beaucoup de libertés dans la façon de le traiter.

Je voudrais remercier mon papa qui m'a été d'une grande aide, notamment dans les questions relatives à l'analyse de données.

Je ne pourrais pas clore cette section sans évoquer Gilles Souvay. Il a consacré beaucoup de son temps à m'expliquer entres autres ce qu'était LGeRM.

A tous, merci.

Chapitre 1

Introduction

1.1 Situation du problème

Le lemmatiseur LGeRM est un outil qui propose une séries de lemmes pour chacune des formes des textes qui lui sont soumis. Actuellement, il fonctionne pour tous les stades du français mais est développé particulièrement pour le moyen français. L'outil a été conçu pour que le lemme correct se trouve dans la liste, aussi restreinte que possible. Malgré tout, LGeRM produit du « bruit ». La motivation pratique du travail est de faire en sorte de réduire ce bruit, pour les textes en moyen français.

Les lemmes proposés par LGeRM, pour une forme donnée, n'entrent pas tous dans la même catégorie morphosyntaxique. L'idée de base pour résoudre le problème est que si l'on parvenait à déterminer la catégorie morphosyntaxique d'un forme donnée, on pourrait exclure les lemmes proposés par LGeRM dont la catégorie n'est pas identique.

1.2 Étapes envisagées

Le lemmasiteur LGeRM doit sa raison d'être au fait que, en moyen français, de nombreuses variantes graphiques existent pour une forme lexicale. La partie initiale de ce travail est naturellement dédiée à l'explication de la présence massive de cette variation en moyen français.

Il existe déjà deux jeux de paramètres statistiques d'étiquetage morphosyntaxique pour l'ancienne langue. Ces paramètres sont basés sur l'ancien français et le français sans distinction entre 1100 et 1600. En tant que paramètres statistiques, ils sont exploitent la fréquence des mots dans la langue. Une étude détaillée de l'évolution comparée de la fréquence de quelques mots bien choisis n'a encore jamais été faite. Effectuer une telle étude permettra de mieux comprendre les résultats obtenus à l'aide de ces paramètres.

Si la fréquence des mots ne change pas, alors il n'est en effet pas nécessaire de procéder à un nouvel entrainement d'un étiqueteur morphosyntaxique dédié préférentiellement au moyen français. Cependant, nous le verrons, le moyen français se distingue particulièrement, en termes de fréquences d'apparition de mots, de l'ancien français et du français classique. Il s'ensuit que de nouveaux paramètres donneront probablement de meilleurs résultats.

Après avoir rappelé les enjeux de la lemmatisation, un chapitre est consacré à une présentation des bases théoriques nécessaires à la compréhension de la nature des paramètres statistiques. Ensuite seulement on procède à une comparaison des nouveaux paramètres créés avec les paramètres existant.

Chapitre 2

Histoire de la langue

2.1 Introduction

Prince à mort sont tous destinez
Et tous autres qui sont vivans ;
S'ilz en sont courcez n'atinez,
Autant en emporte ly vens.

Dans sa ballade « en vieux langage françois », François Villon use de la stratégie rhétorique qui consiste à donner à son texte une coloration archaïque afin d'accentuer la sensation de passé, d'effacement et de vanité de la lutte contre le temps. Ce faisant, il permet de voir que l'homme du XV^{e} siècle a conscience de ne pas parler le même français qu'auparavant. Donc, vers l'an 1450, on pouvait croire que l'on ne parlait plus, dans sa langue vernaculaire, comme Chrétien de Troyes le faisait, en tout cas selon les écrits que l'on conservait. Cette prise de conscience serait très intéressante à étudier, mais ce n'est pas exactement ce qui nous occupera ici. On tentera d'abord de comprendre ce qu'était le moyen français, la langue parlée par François Villon.

Les textes du passé, par les idées transmises et certaines caractéristiques formelles, font penser que la langue que nous parlons a évolué. Devant ce constat, il est habituel, dans le monde occidental, de procéder à un découpage du temps en plusieurs périodes afin de créer des classes, censément homogènes, permettant de mieux définir le phénomène étudié. Or, « la périodisation en histoire de la langue se fonde sur une convention qui peut subir

des modifications en fonction de l'état de nos connaissances sur la langue et selon la prise en compte de divers phénomènes socio-historiques. » [1] Donc, si la périodisation nous aide à créer des classes qui permettent d'étudier et de connaître la langue parlée à telle ou telle époque, cette connaissance acquise, en retour, permet de créer et de distinguer ces intervalles temporels. Le manuel duquel est tiré cette citation n'applique pas le programme qu'il énonce. En effet, avant de la « périodiser » il faudrait d'abord, selon les auteurs, théoriser la langue.

Qu'est-ce que la langue ? comment on peut connaitre une langue parlée dans le passé ? comment des phénomènes socio-historiques interagissent avec la façon dont on parle ? comment et pourquoi tel ou tel phénomène a eu telle ou telle influence sur la façon dont on a parlé ?... Il faut admettre que ce sont là de vastes questions. L'ambition de ce chapitre est de rappeler très rapidement quelques faits qui ont été découverts sur la langue en tant que pratique sociale et sur l'histoire de cette pratique dans nos régions. Dans un chapitre suivant nous continuerons la discussion à propos de la périodisation et apporterons quelques éléments nouveaux qui permettent de justifier une distinction du moyen français.

2.2 La langue, le français

La langue est un *produit social* qui s'impose aux individus. C'est-à-dire que la langue *dépend des individus* et, conjointement, les dépasse, *les façonne*. Lorsqu'ils parlent, les individus produisent des actes de parole. Il se trouve que l'on observe des variations dans les actes de paroles. Ces variations sont fonctions du temps (variation diachronique), du lieu (variation diatopique), des classes sociales (variation diastratique) et de la situation de la production de la parole dans l'espace social (variation diaphasique). La variation – et subséquemment le changement – est inhérente aux pratiques des locuteurs parce qu'ils ne sont pas des machines, parce qu'ils ne parlent pas en appliquant strictement des règles.

Si la langue n'est plus la même c'est qu'elle a changé. Si elle a changé c'est qu'elle a varié. C'est une des grandes leçons de la sociolinguistique. Le phénomène de variation est à la base du changement linguistique, mais est également la cause de défis et difficultés pour les linguistes qui font du traitement automatique du langage, en particulier lorsque les règles de transcription graphique ne font pas l'objet de consensus, comme c'est le cas en moyen français. Nous verrons et explorerons ces difficultés beaucoup plus en détail dans les chapitres 4 et 6.

1. [Cazal et Parussa, 2015], p 53.

Maintenant que nous avons clarifié le concept de langue, il faut déterminer où et quand est pratiquée la langue française. Pour Bernard Cerquiglini, l'affaire est simple et c'est « après les *Serments de Strasbourg*, et seulement après, [que] le français existe » [2]. Soit, donc, pour le « quand ». Après tout les histoires commencent par « Il était une fois ». En réalité, Cerquiglini, dans l'ouvrage cité, s'oppose à une vision du français théorisée par la notion de *koinè* orale. Pour lui, au contraire, le français résulte de l'homogénéisation des usages écrits, née de la volonté *politique* d'une « *scripta* commune, [...] modèle de la langue juridique carolingienne » [3]. Cette hypothèse, par ailleurs complètement téléologique, ne permet en rien de proposer une histoire de la façon dont les gens ont parlé et qui a, c'est une évidence, influencé la façon dont ils ont écrit. L'inverse n'est pas le seul moteur de l'évolution de la langue, contrairement à ce que semble penser Cerquiglini.

Définir l'espace francophone est une gageüre. Aux alentours du XIV^{e} siècle, il y avait dans chaque région de l'Europe de l'Ouest une ou deux langues vernaculaires. A cela s'ajoutait deux langues de prestiges qui connaissaient une diffusion européenne : le latin – langue de l'Église – et le français, en tant que sociolecte de la classe dominante [4]. En effet, « le » français était la langue des cours princières, des tribunaux – sous l'influence colossale de Saint Louis –, la langue de la culture – séculière et religieuse – et de l'aspiration bourgeoise, y compris dans les îles britanniques. Des manuscrits anglais témoignent de ce que le français était compris au moins passivement autant que le latin, voire plus – à tout le moins par un public urbain –, car dans les grandes villes le français pouvait être entendu, jusqu'environ 1420, lors des proclamations publiques des ordonnances [5]. Néanmoins, selon Serge Lusignan, « le cloisonnement du champ discursif entre le latin et la langue vernaculaire constituait un principe structural de la société et de la culture au Moyen Age » [6].

2.3 Histoire et définition

Durant le haut Moyen-Age, les tendances à la diversification du latin sont plus fortes que les tendances centrifuges : délitement de l'état central, scission des lettrés et des *illiterati* y

2. [Cerquiglini, 1991], p. 4.
3. [Cerquiglini, 1991], p. 121.
4. [Rey *et al.*, 2011], p. 114-115
5. Consulter [Busby et Kleinhenz, 2010]
6. [Chaurand, 1999], p. 97.

compris dans les classes dirigeantes et bourgeoises, ... Ces tendances commencent à s'inverser à partir du moment où les locuteurs romans sont conscients de ne plus parler le latin, au moment des *Serments*, comme l'a justement démontré Cerquiglini. Lentement, la situation politique change et, au moins à l'intérieur ce qui est considéré comme le domaine de la langue d'oïl, les tendances unificatrices se font de plus en plus fortes. La langue de l'administration, du pouvoir, va faire l'objet de codifications diverses afin de pouvoir être véhiculaire sans toutefois devenir parfaitement unifiée. Cette langue est en partie écrite et s'oppose, par là, aux parlers uniquement oraux : les patois. Il est toutefois incorrect de parler de langue française pour cette époque. Ce qui était parlé était alors « le produit de la diversification d'un stade très ancien de la langue » [7] : c'est-à-dire des dialectes et non une langue unifiée. Le français de Lille n'est alors pas celui de Paris. La variation spatiale était très présente, très importante.

Il faut considérer le monde médiéval comme un monde plurilingue. Parfois, les pratiques langagières se distribuent et occupent des fonctions bien distinctes. Parfois, elles entrent en concurrence, comme en témoignent de nombreux manuscrits dans lesquels il est fait usage, au sein d'un même texte, de ces différentes langues. Mais en tant que langue de prestige, le français a été une langue parlée par des catégories dominantes de locuteurs, ce qui explique certaines représentations et conflits que cette situation a pu générer. Cette langue de prestige, pratiquée entre autres à Paris, a subi de nombreuses influences et, en réalité, son creuset est probablement le marché. Le français, vu comme langue de rencontre entre certains groupes sociaux provenant d'aires linguistiques différentes, mais également comme langue forgée dans la région parisienne, lieu géographique de rassemblement de populations attirées par le rayonnement naissant dudit endroit est conceptualisé par la notion de *koinè*. Cette *koinè* populaire a existé en même temps que la cour pratiquait une langue influencée par et influençant les pratiques scripturales. Naturellement ce processus de mélange et d'influences multiples s'est produit partout, et chaque centre urbain disposant d'un certain prestige a été peu ou prou un lieu d'homogénéisation. En somme le français est une *koinè* sociale, géographique et médiale. Ce processus n'a pas de fin et si actuellement les pratiques paraissent partagées e homogènes, ce n'est qu'une représentation due à une habitude très francophone-française de monolinguisme. Cette histoire n'a donc ni début, ni fin, et se déroule partout où les locuteurs parlent. Elle permet de mieux comprendre pourquoi on observe des phénomènes de variations (lexicale, grammaticale, graphique, ...) très présent dans les

7. [Klinkenberg, 1999], p. 35.

textes que l'on conserve.

2.4 Questions épistémologiques

Notre connaissance des langues de cette époque se fonde exclusivement sur les textes écrits que nous conservons. Il est donc inévitable de s'interroger sur la nature, la fonction, la valeur d'un texte écrit. Les textes sont des témoins à manipuler avec précaution et en connaissance de cause. C'est seulement comme cela qu'il est possible aux historiens de la langue d'extraire des faits et d'en tirer des conclusions.

2.4.1 Valeurs et fonctions des textes

« Le texte procède d'une volonté de *structuration* » [8], ce qui l'oppose beaucoup à la communication orale. Cependant, certains textes sont eux-mêmes destinés à être dits et « la communication orale peut participer de la nature du texte » [9]. Le médiéviste propose alors de distinguer *monument* et *document*. Un monument est produit lorsqu'il y a une volonté de pérennisation.

Naturellement, ce qui est au départ un document peut se voir investi d'une valeur monumentale. C'est le cas lorsque le document est interrogé par l'historien pour répondre à certaines questions. Le document devient alors trace et mémoire de l'événement. Le monument, en tant que mémoire de l'événement, permet de le faire revivre. Le document n'en est que le témoin.

La volonté du scripteur de mettre en écrit influe naturellement sur sa considération de ce qu'il est écrit. Cette volonté, la nature du texte transcrit (littéraire, juridique, ...), le message contenu vont être déterminant pour la transcription. Les usages écrits varient autant et comme les usages oraux : en fonction du temps, en fonction du lieu, en fonction de qui écrit et à qui, et en fonction de ce qui est écrit. Sans compter les considérations esthétiques qui sont beaucoup plus présentes que dans une production orale commune. Mais qu'est-ce au juste que mettre par écrit ?

8. [Zumthor, 1960], p. 5-6.
9. [Zumthor, 1960], p. 6.

2.4.2 Transcodage

« L'écriture constitue et a constitué, dans toutes les civilisations, un puissant facteur de standardisation et d'institutionnalisation d'un des systèmes sémiotiques les plus remarquables : le langage verbal », écrit Jean-Marie Klinkenberg [10]. Cette affirmation se révèle particulièrement valable pour le cas du français dans la période qui nous occupe. L'auteur ajoute immédiatement que l'écriture ne joue pas uniquement un rôle social. L'écriture assume une fonction de transcodage qui permet de rendre pérenne et diffusable la parole dans le temps et dans l'espace.

Écriture et langage, ce couple est particulièrement lié durant le Moyen Age, période pendant laquelle les textes étaient principalement dits et non lus. Il existe deux stratégies contraires et non contradictoires qui permettent de transcoder la parole. Soit sont codés les signifiants, soit le sont les signifiés. Dans le premier cas, on parle d'écriture phonographique. Ce type d'écriture est relativement économique en ce qui concerne le nombre de caractères nécessaires à rendre compte des phonèmes d'une langue. L'inconvénient est qu'il faut connaître la langue pour pouvoir comprendre ce qui est transcrit, quand bien même serait-on capable d'en déchiffrer les caractères écrits. Dans le second cas, puisque ce sont les signifiés qui sont transcrits, il n'est pas forcément nécessaires de savoir parler la langue utilisée pour comprendre ce qui est écrit. Cette relative universalité de l'écriture logographique souffre de ce qu'il faut alors apprendre un grand nombre de signes pour maîtriser ce code. Le plus souvent, les écritures sont mixtes : ni totalement phonographiques ni totalement logographiques.

Les signes du code écrit peuvent être de différents types : phonographiques et logographiques comme on l'a vu, mais aussi morphologiques et thématiques. Les signes morphologiques renvoient à des relations morphosyntaxiques. Par exemple, en français, le /s/ en fin de nom commun signifie que ce nom est accordé au pluriel ce qui est inaudible, sauf si le mot suivant, dans le même segment propositionnel, débute par une voyelle. Les signes thématiques indiquent l'appartenance du signifiant voisin à telle ou telle catégorie (animé/inanimé, etc.).

Ce qui précède renvoie aux fonctions *graphémologiques* de l'écriture, c'est-à-dire aux fonctions liées à la représentations de la langue [11]. Or rien n'a encore été dit à propos des signes graphiques ne renvoyant pas à des signes linguistiques comme l'emploi de couleurs, de souli-

10. [Klinkenberg, 2000], p. 223. Son « précis » est essentiel et on consulte également avec profit, du même auteur, « O comme l'orthographe, un monstre sacré ? » dans [Cerquiglini *et al.*, 2000], p. 219-229.

11. Selon la terminologie de Klinkenberg dans [Klinkenberg, 2000], p. 226 et suivantes

gnés, la disposition du texte sur la page, la contiguïté du texte et d'un objet matériel (ainsi la pancarte indiquant la boulangerie), le choix calligraphique, etc. Ces signes relèvent des fonctions *grammatologiques* de l'écriture.

2.4.3 Orthographe

Considérations générales

Il est adéquat de ne pas réduire la notion d'orthographe à un ensemble de règles qui assureraient le bon fonctionnement du code graphique. Cette conception de l'orthographe n'est pas universelle, ni dans l'espace ni dans le temps. Par exemple, suivant la première définition de ce terme que l'on trouve dans le TLFi [12], l'orthographe est la « manière, considérée comme correcte, d'écrire un mot ». Il n'est pas question de marche à suivre mais de correction. Il n'est pas question d'aborder dans ce travail les représentations ou les enjeux de domination associés à l'orthographe. Cependant, ne pas les évoquer serait faire preuve d'aveuglement.

Plusieurs types de signes destinés à transcrire le langage ont déjà été présentés. Les signes phonographiques transcrivent les phonèmes, on l'a dit. C'est-à-dire qu'à un phonème est associé un phonogramme. En français, seules trois lettres sont toujours prononcées et correspondent chacune à un seul son : les lettres « j », « k » et « v ».

Malheureusement, l'association phonème-phonogramme est loin d'être bijective : certains phonèmes peuvent être notés par différents phonogrammes et inversement, certains phonogrammes peuvent correspondre à plusieurs phonèmes. On appelle « graphèmes positionnels » les variantes possibles des phonogrammes. On peut citer, à titre d'exemple, toutes les variantes graphiques transcrivant le son /o/, à savoir « o », « au », « eau », « a » (comme dans « hall ») et « ô ».

Les morphogrammes sont les graphèmes qui transcrivent les morphèmes. Il s'agit des marqueurs de désinence, de flexion, de dérivation, etc. Ces marqueurs sont prononcés ou non.

L'écriture du français est également logographique lorsqu'elle note les homophones. Dans ce cas, le lecteur doit simplement associer de mémoire telle forme à tel concept, sans passer par une étape phonique.

Enfin, certaines lettres ne se prononcent pas et sont simplement la trace d'une histoire : les

12. `http://www.cnrtl.fr/definition/orthographe`

lettres historiques et étymologiques. Les lettres étymologiques renseignent l'homme averti sur une l'étymologie du mot, les lettres historiques sont une marque d'une ancienne prononciation du mot.

L'orthographe du moyen français

C'est parce que l'écriture du (moyen) français est hybride dans son principe et dans ses fonctions que l'orthographe qui la décrit (et non régit) est si complexe. Voyons à présent (succinctement) comment nous en sommes arrivés là.

Le latin parlé a été transformé en une multitude de dialectes. L'écriture du latin était phonographique. C'est cet alphabet-là, ainsi que les habitudes orthographiques partagées qui ont été repris pour les premières inscriptions des langues vernaculaires. Or cet alphabet n'était pas adapté pour transcrire tous les sons nouveaux de ces langues. Lorsqu'il faut adapter un alphabet aux besoins d'une langue pour laquelle il n'est pas conçu, plusieurs solutions sont possibles : inventer des nouvelles lettres, utiliser un groupe de lettres pour noter certains sons.

Rappelons le contexte de l'écriture durant le Moyen-Age. La culture écrite est essentiellement latine : plus de trois quart des incunables édités avant 1500 sont écrits en latin. Près de la moitié des ouvrages traitent un sujet religieux, 30% sont des ouvrages de l'Antiquité classique et 10% sont relatifs au droit et à la science [13]. Ces chiffres sont à considérer avec une certaine distance car les voies de la conservation des ouvrages nous sont inconnues. Le statut des ouvrages écrits en latin était supérieur à celui des ouvrages en langue vernaculaire, de même l'était celui des ouvrages traitant des sujets religieux par rapports aux autres sujets. Il n'est pas surprenant de penser que les ouvrages conservés sont en majorité ceux d'un statut supérieur. Ainsi, ces proportions et, partant, les conclusions sur la culture médiévale, sont à relativiser. Néanmoins cela permet de mieux comprendre l'influence que le latin a exercée sur la façon d'écrire les langues vernaculaires.

La graphie adaptée conserve un caractère essentiellement phonologique. C'est grâce à ce fait que « tout graphie, si complexe soit-elle, reste plus ou moins *compréhensible* – fonctionnelle – dans un pays donné » [14]. La graphie du moyen français se caractérise spécifiquement

13. Nina Catach, dans [Catach, 2001] p. 74-75, fait référence à une étude menée par H.J. Martin dans Febvre L. et Martin H.J., *L'apparition du livre*

14. [Catach, 2001], p. 79.

par des tendances « *étymologique, historique, morphologique et sémantique*, avec le rôle tout particulier joué par la distinction des homonymes et le problème des monosyllabes » [15]. Les scripteurs des langues vernaculaires ont, durant la période du moyen français, de plus en plus eu l'idée de rapprocher les graphies de leurs étymons latins.

Les *scriptae*

Nous avons déjà évoqué le fait que chaque centre urbain pouvait devenir le creuset d'une *koinè*, orale ou écrite. Ce fait se marque, dans les textes que nous conservons, par l'existence de *scritpae*. Une *scripta* est la forme écrite d'une langue, qui varie d'une région à l'autre. On pourrait comparer cette notion à celle de dialecte sous forme écrite. Pour retracer l'histoire de la langue parlée, l'habitude a été prise de comparer la répartition de ces *scriptae* dans l'espace avec l'extension actuelle des dialectes.

Il faut retenir de cela que la cohabitation des dialectes, des sociolectes, des *koinè* orale et écrite, des *scriptae* et les influences mutuelles de toutes ces formes d'expression font que la situation linguistique durant l'époque du moyen français est infiniment complexe.

2.5 Le moyen français

Pour conclure promptement, il faut dire que le moyen français a souvent été présenté comme cet état de langue « transitoire » entre l'ancien français et le français classique. Gaston Zink propose de voir l'évolution de l'ancien français au moyen français comme une conséquence de la guerre de Cent Ans [16].

Le lexique s'étoffe, énormément par emprunts. La syntaxe évolue. Alors qu'elle restait encore assez flexionnelle en ancien français, le déclin de la déclinaison et le bouleversement de l'ordre des mots font que la syntaxe devient plus analytique.

Dans le même temps, la prononciation se « simplifie ». Parmi les changements phonétiques on peut citer : réduction des diphtongues et des hiatus, amuïssement des consonnes implosives et évolution du [e] central.

15. [Catach, 2001], p. 79, c'est l'auteure qui souligne
16. [Zink, 1990], p. 3.

Chapitre 3

Statistiques

3.1 Introduction

La statistique peut être définie comme la « branche des mathématiques ayant pour objet l'analyse (généralement non exhaustive) et l'interprétation de données quantifiables » [1], et l'auteur de la notice de donner comme exemple la « statistique lexicale ». Cette définition ne donne aucune précision à propos des données. Sont-elles recueillies passivement ? Cueillies activement ? Complètement construites ? Quel est le rapport qu'elles entretiennent alors avec la réalité ? Il est difficile de répondre à toutes ces questions, mais ne pas en tenir compte revient à occulter complètement une partie de ce qu'est la statistique.

Comme tous les scientifiques, les statisticiens *construisent* des objets qui décrivent le monde. Leurs méthodes produisent des nombres, une langage chiffré qu'il faut décoder. Ce sont ces méthodes qu'il faut appréhender avant même de se confronter aux « chiffres », sans quoi ils sont dénués de sens. Car, en effet, l'interprétation consiste justement en conférer du sens à ces données, aux analyses effectuées sur ces données.

Il ne faut toutefois pas s'attendre à trouver ici une présentation détaillée et rigoureuse mathématiquement de toutes les méthodes employées. Par contre, on trouvera une description de ces méthodes et, autant que faire se peut, une explication intuitive. Ces méthodes et outils sont d'abord ceux de la statistique descriptive, puis de l'analyse en composantes principales

1. Voir `http://www.cnrtl.fr/definition/statistique`

et enfin de la classification ascendante hiérarchique [2].

Grâce à ces méthodes, il sera possible d'éclairer l'histoire de la langue très différemment de ce qui a été fait jusqu'à présent. En effet, les tableaux que nous construirons seront la base d'explorations inédites et intéressantes. Nous apporterons de nouveaux éléments qui corroborent l'hypothèse que le moyen français est distinct de l'ancien français et du français « moderne »– les subdivisions plus fines des périodes pendant lesquelles ces états de langues s'observent sont également présentées.

Nous livrons au fil du texte les graphiques et tableaux nécessaires à la compréhension de notre propos. Nous joindrons en annexe numérique uniquement le code R écrit et les sorties que le logiciel a produit. Ce fichier, trop volumineux, ne sera pas imprimé.

3.2 Données

La base de données Frantext [3] est une des plus grandes bases de données de textes en langue française. Elle est développée par l'ATILF, initialement pour servir à la recherche d'exemples pour le Trésor de la Langue Française. L'ambition de cette base n'est pas de constituer un recueil des « grands textes de la langue », mais d'être représentative du français dans ses usages. Si le corpus a d'abord été centré sur les XIX^{e} et XX^{e} siècles, il a ensuite été élargi et couvre maintenant l'histoire de la langue de 1100 à nos jours. Ce corpus permet de faire des recherches et, en particulier, de compter les fréquences des formes demandées, sur une période de temps au choix.

Nous avons collecté la fréquence relative des mots « car », « dans », « de », « dedans », « et », « par », « parce » et « pour » sur chaque intervalle de 50 ans, à partir de l'an 1100. Il faut considérer les fréquences relatives pour normaliser les observations et, afin d'observer des nombres un peu « parlant », ces fréquences sont indiquées en millionièmes, puis placées dans une matrice de 19 lignes, correspondant aux période, et de 8 colonnes, pour les mots (voir le tableau 3.1). Les proportions ne changeront jamais de façon significative étant donnée la taille actuelle du corpus.

2. L'idée d'employer les méthodes d'analyses de données comme nous allons le faire ici provient d'un travail déjà présenté l'an passé au cours de statistiques linguistiques. Néanmoins les analyses ont été refaites et modifiées. Les conclusions sont affinées. La présentation des méthodes a été complètement revue et précisée. De nombreuses figures sont nouvelles.

3. `https://www.frantext.fr`

Le choix de ces mots doit être commenté. Il a fallu choisir des mots qui ont toujours existé et ont gardé un sens comparable tout de long de l'histoire de la langue. Il a fallu s'assurer que la variation graphique qui touchait ces formes n'était pas trop importante, à moins de parvenir à la neutraliser au moment du calcul des fréquences. Par exemple, pour déterminer les fréquences du mot « car », il a fallu rechercher « car », « kar », « quar », « quer », « quare », etc. La forme homonyme, qui désigne une voiture de tramway, n'est apparue qu'à la fin du XIX^{e} siècle et, par conséquent, cela ne joue pas sur les fréquences des périodes qui précèdent. Les mots « car » et « parce » ont été sélectionné car ils ont fait l'objet d'études qui seront convoquées en temps utile.

3.3 Exploration des données

Le tableau suivant montre la matrice constituée par les fréquences des mots durant les périodes déterminées. Puisque rien n'est plus semblable à un nombre qu'un autre nombre, le tableau doit être exploré avec tous les outils dont on dispose. Les premiers – et plus simples – sont ceux de la statistique descriptive.

3.3.1 Premiers graphiques

La fréquence relative d'apparition du mot « et » semble (voir le tableau 3.1) diminuer au fil du temps. De plus de 40.000 occurrences par million entre 1250 et 1550, elle chute et passe sous les 20.000 entre 1600 et 2000. L'histogramme 3.1 permet de visualiser cette évolution.

Si l'on peut éventuellement attribuer le faible taux avant 1250 à une erreur de représentativité due au faible nombre de textes dans le corpus pour ces périodes, il faut s'interroger plus sérieusement sur l'évolution globale et consulter les grammaires. En ancien français « classique »– selon l'expression de Geneviève Hasenohr [4] –, « et » recouvre plusieurs fonctions. Ce mot peut être un coordonnant semblable à celui du français contemporain, mais il confère également une plus forte importance au sujet de la proposition qu'il introduit, il peut souligner l'articulation des propositions dans pour autant les coordonner et, dans les dialogue, « et » en tête joue un rôle d'insistance sur le nouveau locuteur. Enfin, M. Wilmet remarque

4. Ce paragraphe est basé entre autres sur les informations que l'on trouve dans [Raynaud de Lage, 1993].

TABLE 3.1 – Taux (en millionièmes) d'apparition des mots dans la base Frantext, par périodes de 50 ans

mot / **période**	car	dans	de	dedans	et	par	parce	pour
1100-1149	1623	0	18161	0	180	6552	0	0
1150-1199	2550	110	19062	0	25370	5841	0	556
1200-1249	3281	34	20090	1	33341	6713	1	1301
1250-1299	1059	36	44670	186	55950	8196	0	4118
1300-1349	2874	6	34628	46	53030	9076	7	8286
1350-1399	4516	7	27778	39	46585	8318	31	8146
1400-1449	3513	6	35449	60	50111	7780	17	7246
1450-1499	3009	18	36536	283	47878	6643	18	7482
1500-1549	3199	333	30763	578	40225	6086	52	6459
1550-1599	2448	632	33886	490	30962	5700	145	6310
1600-1649	1690	2623	36823	315	24328	4742	398	7425
1650-1699	1207	5958	36467	100	22654	4785	691	6902
1700-1749	660	7192	37015	56	21232	4285	459	7053
1750-1799	565	7490	36847	34	20642	4461	543	6106
1800-1849	811	7665	37155	50	21452	4365	364	5728
1850-1899	643	7094	35031	65	20492	3835	307	4976
1900-1949	610	7117	35373	62	18969	4253	474	5086
1950-1999	449	7307	35324	60	17921	4151	479	5204
2000-2049	395	7568	34536	56	16908	3366	527	5521

que le « Eh bien » n'a remplacé « Et bien » « qu'au dix-septième siècle » [5], ce qui souligne encore le problème de la variation graphique. De nos jours, dans la plupart des cas, « et » coordonne deux propositions. En tête de phrase, il peut assumer, selon Wilmet, un rôle de « transitivité factice ou mémorielle » [6] ou encore, s'il ne coordonne rien, être adverbe. Deux facteurs principaux expliquent la diminution des fréquences relatives d'apparition de ce mot. D'une part on ne trouve pour ainsi dire plus d'expressions latines dans les corpus actuels or le mot « et » a exactement la même forme dans cette langue. D'autre part, en « français », ce mot s'emploie dans un nombre plus faible de situations qu'auparavant.

D'un point de vue logique, rien n'empêche de faire des recherches pour expliquer l'évolution de la fréquence d'apparition de tous les mots. Cependant, quoique cela puisse rendre plus documentée la connaissance de l'histoire de la langue, il manque une vue d'ensemble

5. [Wilmet, 1997], p. 570.
6. [Wilmet, 1997], p. 578.

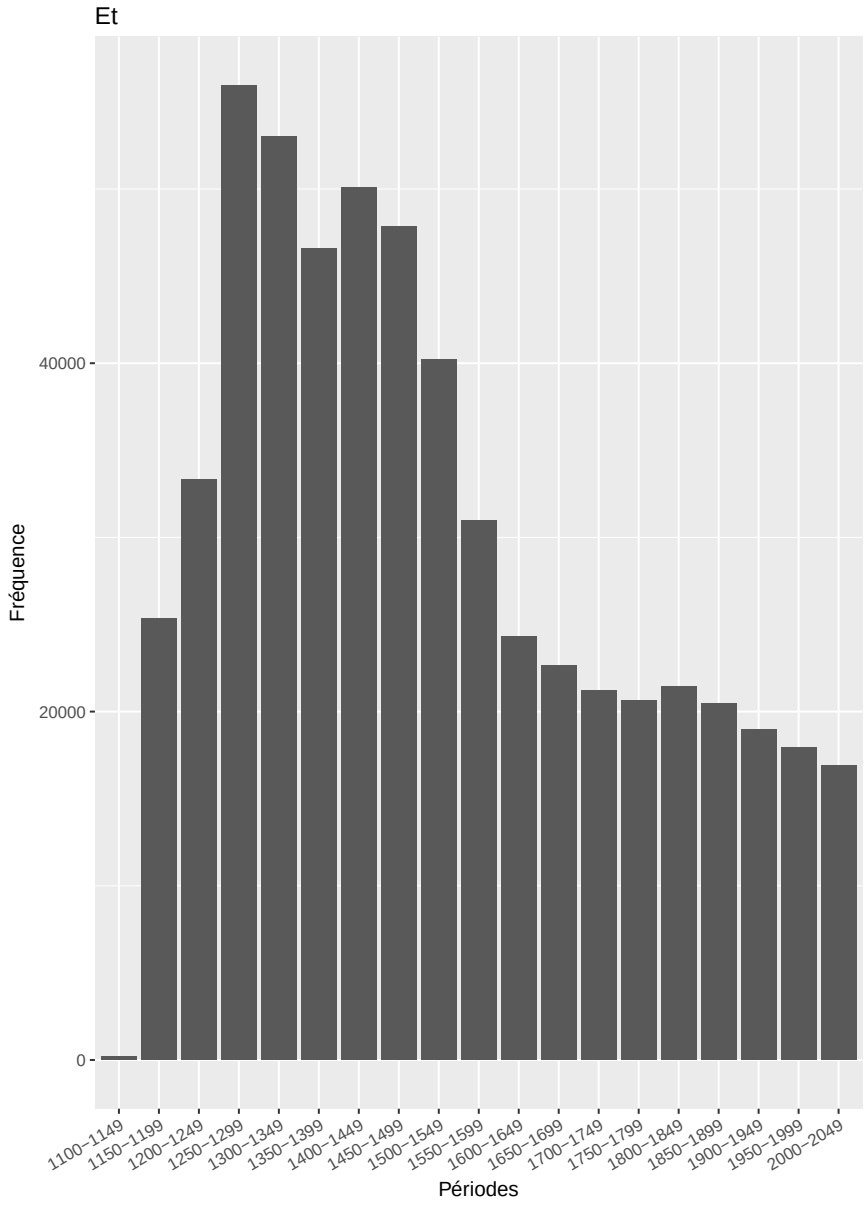


FIGURE 3.1 – Evolution de la fréquence de « et » en fonction du temps

de la dynamique globale de l'évolution des usages. Les outils de la statistique descriptives sont justement conçus pour décrire et résumer un ensemble de résultats. Un de ces outils est l'analyse des corrélations.

3.3.2 Tableau des corrélations

L'historien de la langue dispose donc d'une matière première lorsqu'il envisage l'évolution de la fréquence des mots, mot par mot. En étudiant l'évolution conjointe de ces fréquences, des points d'attention surgissent qu'il faut également analyser. Des liens étroits apparaissent entre certains mots. Comme le rappelle Saussure, « [à] chaque instant [le langage] implique à la fois un système établi et une évolution ; à chaque moment, il est une institution actuelle et un produit du passé. Il semble à première vue très simple de distinguer entre ce système et son histoire, entre ce qu'il est et ce qu'il a été ; en réalité, le rapport qui unit ces deux choses est si étroit qu'on a peine à les séparer » [7]. Le tableau 3.2 fournit un résumé de ces liens « fréquentiels » qui se sont noués. Le graphique *scatterplot* condense de nombreuses informations. Tout cela va être présenté dans les paragraphes subséquents.

TABLE 3.2 – Coefficients de corrélation linéaires

mot / **période**	car	dans	de	dedans	et	par	parce	pour
car	1.00	-0.84	-0.43	0.24	0.64	0.77	-0.76	0.13
dans	-0.84	1.00	0.42	-0.31	-0.61	-0.86	0.91	0.16
de	-0.43	0.42	1.00	0.25	0.30	-0.17	0.44	0.69
dedans	0.24	-0.31	0.25	1.00	0.28	0.04	-0.16	0.34
et	0.64	-0.61	0.30	0.28	1.00	0.77	-0.58	0.45
par	0.77	-0.86	-0.17	0.04	0.77	1.00	-0.82	0.06
parce	-0.76	0.91	0.44	-0.16	-0.58	-0.82	1.00	0.27
pour	0.13	0.16	0.69	0.34	0.45	0.06	0.27	1.00

Les coefficients de corrélations varient entre 1 et -1. Plus leur valeur absolue est proche de l'unité, plus la corrélation est forte entre les deux variables correspondantes. A titre d'exemple, nous observons une corrélation positive entre « parce » et « dans » ainsi qu'une corrélation négative entre « par » et « dans », toutes deux assez fortes. Cela signifie que les fréquences d'apparition de ces mots ont varié respectivement dans le même sens et dans le sens contraire. Donc au moment où la fréquence de « parce » augmentait, celle de « dans » le faisait également, de même lorsque les fréquences diminuaient. Par contre les fréquences de « par » et « dans » ont varié de façon contraires : lorsque l'une augmentait, l'autre diminuait. Ces coefficients ne renseignent toutefois pas sur la direction de l'évolution des fréquences.

7. [de Saussure, 1994], p. 24

Le graphique *scatterplot*[8] (voir la figure 3.2) résume un certain nombre de données unidimensionnelles et multidimensionnelles. On trouve sur la diagonale principale les histogrammes de l'évolution des fréquences, les coefficients de corrélation (en valeur absolue) sur la partie haute et une régression linéaire faites sur les données utilisées pour calculer les coefficients de corrélations sur la partie basse. La taille de police utilisée pour les corrélations est proportionnelle à la valeur absolue du coefficient de corrélation. C'est pour cette raison que certains coefficients, trop faibles, ne se distinguent pas. La graphique de régression linéaire permet d'apprécier la dispersion des données beaucoup mieux que ne le permet la simple indication du coefficient. Par exemple, il semble qu'il y ait une plus grande linéarité dans la dispersion des fréquences de « par » en fonction de celles de « et », le coefficient de corrélation valant 0.77, qu'entre les fréquences de « dans » et « par » ou « parce » alors que leurs coefficients de corrélation sont élevés. Quelques questions se posent à présent. Que signifient ces corrélations par rapport à l'histoire ? Sont-elles sources de connaissance ?

La corrélation la plus forte (entre « parce » et « dans ») est difficilement explicable, tout comme celle entre « dans » et « car », au contraire de la corrélation (négative) qui est observée entre « parce » et « car ». A priori, en effet, il n'y a pas trop de lien entre l'emploi de « parce », ou de « car », et celui de « dans ». Par contre les deux mots « parce » et « car » s'emploient pour exprimer la causalité. Or on sait que l'emploi de « car » diminue au profit de celui de « parce que »[9]. Ceci nous permet d'ailleurs d'expliquer d'autres corrélations. Le mot « parce » était autrefois fréquemment écrit « par ce ». Il s'ensuit directement que plus la graphie attachée est utilisée, plus la fréquence de « par » diminue : ce qui correspond à une corrélation négative entre « par » et « parce ». Par voie de conséquence, on s'attend à trouver une corrélation positive entre « par » et « car », ce que l'on ne pouvait expliquer de prime abord. Bien d'autres liens existent entre les formes dont les fréquences sont ici dénombrées et ce dénombrement aurait pu être fait pour d'autres formes encore. Les explications données jusqu'à présent suffisent toutefois à justifier l'emploi d'autres méthodes d'exploration de données. On éprouve en effet un manque de vision synthétique des données. Pour pallier ce manque, la figure 3.3 représente simultanément l'évolution au cours du temps de la fréquence de tous les mots envisagés. Mais avant de passer à la section suivante, il faut terminer en signalant que les mots « et » et « de » sont beaucoup plus fréquents que les autres (voir la

8. Le nom de ce graphique ainsi que les commandes R pour l'obtenir sont tirés de [Bellanger et Tomassone, 2014], p. 56-57.

9. Voir [Ruppli, 1990] et [Fagard, B. et Degand, L., 2008].

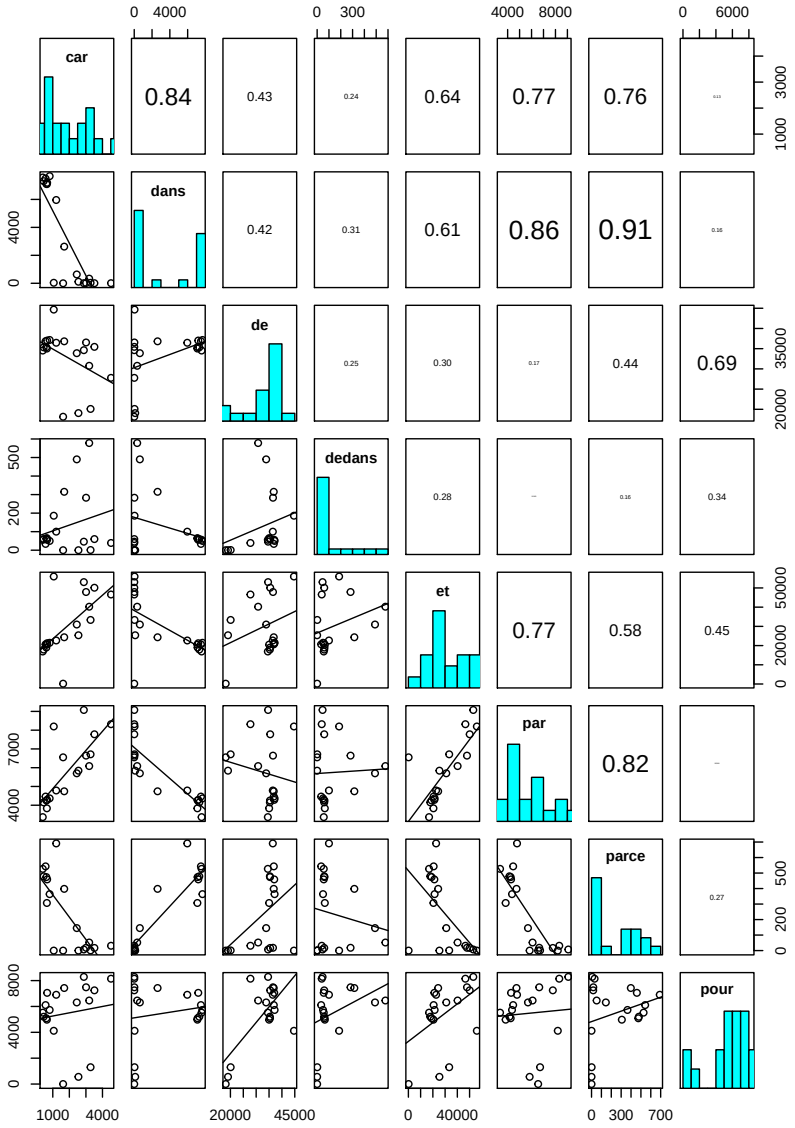


FIGURE 3.2 – Scatterplot : histogrammes, coefficients de corrélation et régressions linéaires

figure 3.3 et le tableau 3.1). Cette caractéristique aura des incidences profondes sur les autres méthodes d'exploration des données que nous emploieront.

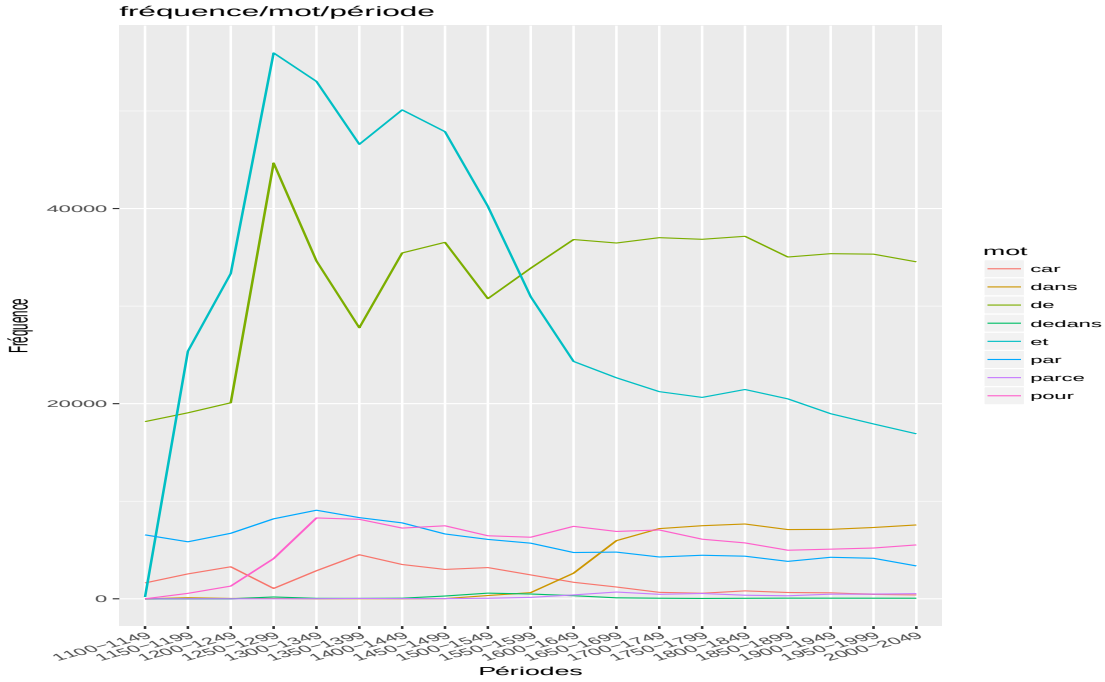


FIGURE 3.3 – Évolution de la fréquence d'emploi des mots en fonction du temps

3.3.3 L'analyse en composantes principales

Notions fondamentales

La méthode de l'analyse en composantes principales (dorénavant abrégée ACP) [10] « a pour objet la description des données contenues dans un tableau individus-caractères numériques : p caractères sont mesurés sur n individus » [11]. Elle permet une représentation graphique de ces individus sur un plan. Cette représentation n'est pas évidente puisque le nombre de dimensions est de p. Dans notre cas, nous pouvons considérer les individus « mots », exprimés en fonction des caractères « périodes », ou à l'inverse considérer les individus « périodes » en fonction des caractères « mots ». Dans le premier cas le nombre de dimensions est de 19 (il y a 19 périodes), il est de 8 (car il y a 8 mots) dans le second cas.

L'ACP est une méthode de réduction du nombre de variables. En réduisant les variables à 2 ou 3 tout en conservant les « différences saillantes » entre les individus, il sera possible

10. Nous renvoyons à [Bellanger et Tomassone, 2014], [Bouroche et Saporta, 1980] et [Escofier et Pagès, 2008] pour tout complément d'information à propos de la méthode et à [Cornillon *et al.*, 2010] et [Husson *et al.*, 2009] pour des informations à propos de R.

11. [Bouroche et Saporta, 1980], p. 17.

d'obtenir un graphique représentatif des individus. Cette opération de réduction consiste à choisir un plan de projection – donc deux axes, deux composantes principales – sur lequel les distances entre les individus seront « en moyenne » le mieux conservées. En effet une opération de projection réduit toujours les distances entre les individus. On se fixe pour critère de trouver les axes de projection qui maximisent la moyenne des carrés des distances entre les projections des individus.

Les axes ainsi déterminés ne correspondent pas avec des dimensions déjà présentes dans l'espace où se trouve les individus. Le système de coordonnées du plan de projection résulte en réalité d'une combinaison linéaire des composantes qui caractérisent les individus. La nature de ce travail empêche de se focaliser plus avant sur des questions mathématiques. C'est une métaphore qui va servir de conclusion préliminaire et justifier le questionnement et la nécessité de bien représenter le nuage de points correspondant aux individus.

Imaginons un dromadaire vu de profil. Si l'on dessine son contour et que l'on noirci la figure obtenue, on distingue facilement que cette ombre chinoise est celle dudit animal. Si l'on avait fait de même, vu de face, on aurait été bien en peine de reconnaître quoi que ce soit. Ainsi, la projection peut changer radicalement l'idée que l'on se fait des données.

Interprétation des différents graphiques

Observons à présent les figures 3.4 et 3.5. Elles représentent la projection des mots, chacun caractérisé par 19 dimensions correspondant aux périodes, sur le plan des composantes principales. Ces deux figures sont très différentes. Dans les deux cas, il est indiqué à côté des axes du plan la « quantité d'inertie conservée », autrement dit la qualité de la projection. L'*inertie*, dont le calcul est précisément celui d'une variance, est une façon de rendre compte de la dispersion des données du nuage initial. Or plus l'inertie du nuage de points projeté dans le plan est proche de l'inertie du nuage initial, meilleure est la qualité de la projection.

La première composante de la figure 3.4 rend compte à elle seule de 88,94% de l'inertie totale du nuage des individus dans l'espace des 19 dimensions-périodes. Cette composante et, partant, la dispersion des individus par rapport à cette composante sont très représentative du nuage de point initial. L'ajout de la deuxième composante permet d'atteindre une inertie totale représentée de 98,68%. Il est tout à fait clair que ce qui induit cela est la fréquence d'apparition bien plus élevée de « et » et de « de » que des autres mots à chaque période. Ces

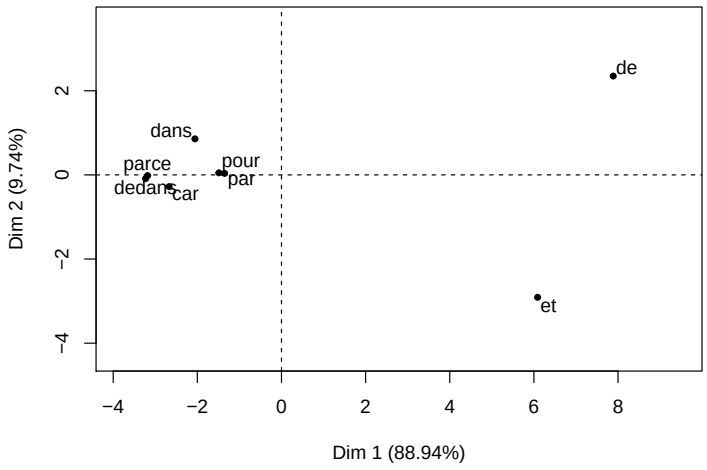


FIGURE 3.4 – Représentation plane des individus avec « et » et « de »

deux individus très éloignés des autres penchent le plan de représentation car cet éloignement participe dans une large proportion à la détermination de la première composante principale, le premier axe du plan de projection.

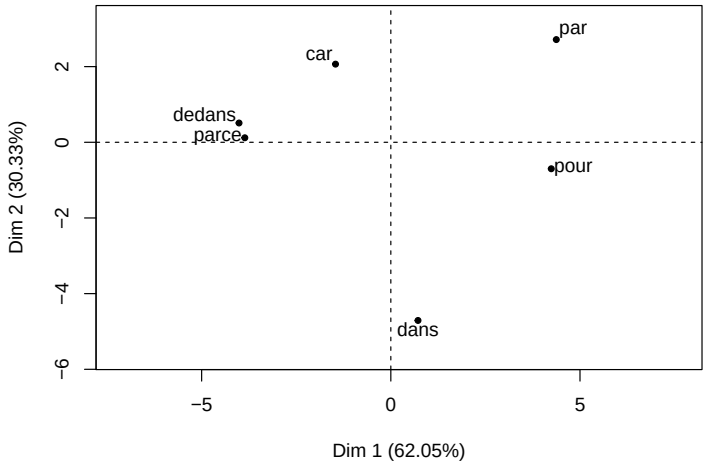


FIGURE 3.5 – Représentation plane des individus sans « et » et « de »

La figure 3.5 représente la projection effectuée sans les deux mots trop fréquents. La

direction des axes permet à présent de conserver respectivement 62.05% et 30.33% de l'inertie totale du nuage des individus (sans « et » et « de »). Ce résultat est très bon. La comparaison entre ce graphique et le précédent permet de visualiser à quel point les mots « et » et « de » sont distincts des autres.

Observons à présent la figure 3.6. Elle représente la projection des périodes, chacune caractérisée par un ensemble de 8 valeurs, les » dimensions-fréquences des mots ». On constate que la qualité de la projection n'est certes pas trop mauvaise, d'après l'inertie conservée, et il y a une « curiosité ». Les points où sont projetées les périodes dans le plan forment visuellement trois groupes distincts.

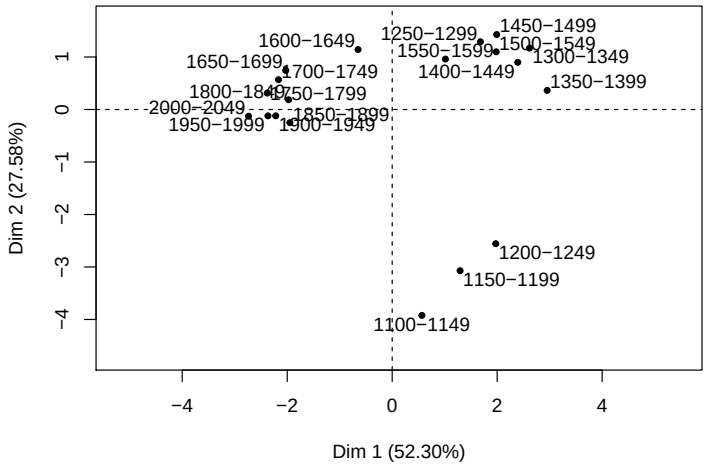


FIGURE 3.6 – Représentation plane des périodes avec « et » et « de »

La projection des individus-périodes, effectuée lorsque ces individus ne sont pas caractérisés par les fréquences de « et » et « de » est représentée par la figure 3.7. D'abord la qualité de la projection est meilleure, d'après l'inertie conservée. Ensuite, cette projection est plus conforme à l'intuition car elle regroupe un peu différemment les périodes. En particulier, la période 1250-1299 se trouve plus proche de la classe 1100-1249 qu'elle ne l'était dans la figure 3.6.

Évidemment, cette comparaison prend tout son sens si l'on peut déterminer la qualité de la projection pour un individu particulier, ici la période 1250-1299.

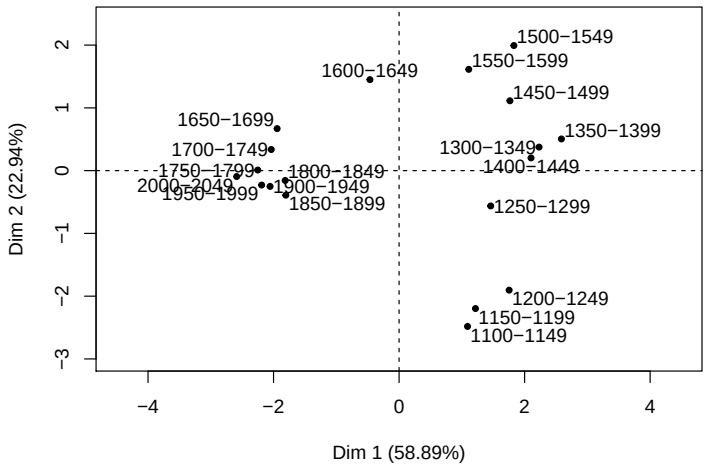


FIGURE 3.7 – Représentation plane des périodes sans « et » et « de »

Contributions des individus et qualité de la projection

Les individus ne contribuent pas tous autant à la direction des composantes principales. Considérons la figure 3.8. Rappelons que toute projection d'un espace dans un sous-espace diminue la distance entre les points projetés. Afin de conserver au mieux l'étendue du nuage de points, il faut que, pour chaque individu, le cosinus de l'angle θ, formé entre OI et le plan des composantes principales, soit le plus proche de 1 possible. Autrement dit, il faut que cet angle soit le plus petit possible. En effet, $d(O', I_p) = cos(\theta) \cdot d(O', I)$.

Ceci explique que des données très éloignées du nuage de point penchent fortement le plan, car l'inertie sur le plan de projection doit être maximale. Autrement dit le cosinus de l'angle formé entre le plan et le vecteur, dont l'origine est le centre de gravité du nuage, du point très éloigné des autres doit être le plus grand possible, donc l'angle le plus resserré. Le plan doit pencher vers ce point éloigné. Il se trouve que le carré du cosinus de cet angle est précisément le rapport entre l'inertie de la projection I_p de l'individu I sur le plan et l'inertie totale de l'individu. Or ce carré est la somme du carré des cosinus entre le vecteur $O'I$ et les vecteurs qui portent les composantes principales. C'est cette relation que suggère les pointillés marquant les coordonnées de la projection I_p de l'individu I dans le plan ACP

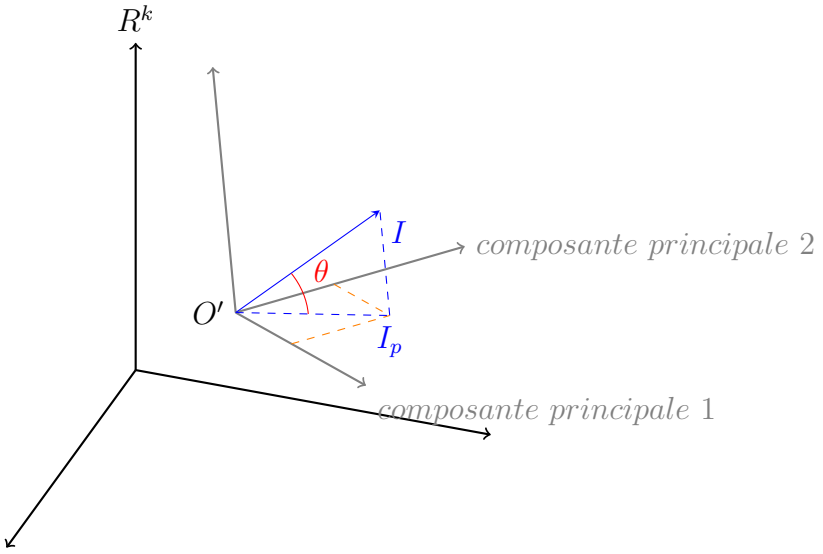


FIGURE 3.8 – Qualité de représentation (I_p) d'un individu (I) par le plan des composantes principales

sur la figure 3.8.

$$cos^2(\theta) = \frac{d^2(O', I_p)}{d(O', I)} = \frac{[\text{inertie de la projection de l'individu } I \text{ sur le plan}]}{\text{inertie totale de } I} \tag{3.1}$$

Pour l'individu-période 1250-1299, le carré du cosinus de l'angle entre le vecteur $O'I$ et le vecteur directeur de la composante principale 1 et le vecteur directeur de la composante principale 2 valent respectivement 0.254 et 0.150, lorsque les mots « et » et « de » entrent en compte. Ces cosinus valent respectivement 0.434 et 0.064 lorsque ces deux mots ne sont pas pris en compte. La qualité totale de la représentation de l'individu-période 1250-1299 vaut donc 0.404 dans le premier cas et 0.498 dans le second cas. Ainsi, non seulement il est projeté plus près de la classe qui lui est chronologiquement proche mais en plus la qualité de sa projection est meilleure.

Conclusion

La méthode de l'ACP nous permet de déterminer trois groupes de données lorsque l'on projette les individus-périodes : de 1100 à 1299, de 1300 à 1599 et de 1650 à 2049. Les périodes 1600-1649 et 1250-1299 se situent entre deux groupes les bordant chronologiquement. La

figure 3.7, construite à l'aide de méthodes statistiques sur base uniquement de la fréquence de 6 mots, montre une distinction entre des groupes de périodes qui recouvrent assez bien les périodes que les linguistes attribuent à l'ancien français, au moyen français et au français moderne. Nous avons choisi de prendre en compte davantage la figure 3.7 plutôt que la figure 3.6 car la projection des individus, dont nous avons détaillé la mesure pour un individu en particulier qui posait question, était meilleure.

3.3.4 Classification ascendante hiérarchique

Nous trouvons dans le TLFi que « [l]a classification proprement dite est une opération de l'esprit qui, pour la commodité des recherches ou de la nomenclature, pour le secours de la mémoire, pour les besoins de l'enseignement, ou dans tout autre but relatif à l'homme, groupe artificiellement des objets auxquels il trouve quelque caractère commun, et donne au groupe artificiel ainsi formé une étiquette ou un nom générique » [12]. Une méthode de classification des individus-périodes permettrait d'affiner l'image des « groupes », « classes » que nous avons cru pouvoir déceler dans la représentation de la projection sur le plan des composantes principales.

Une méthode

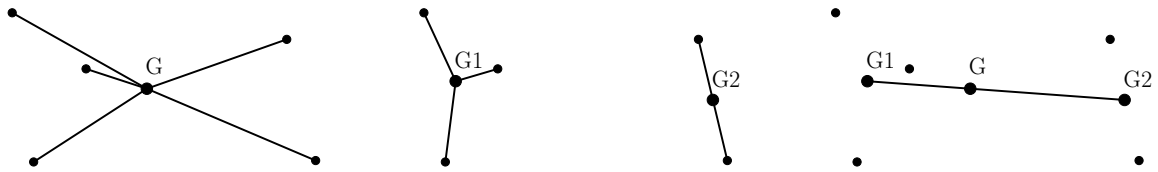
Une méthode de classification est un algorithme qui permet de classer n individus en groupes homogènes en fonction des p caractères qui déterminent ces individus. Dans notre cas, la fréquence des mots détermine les périodes. Il existe deux grandes familles de tels algorithmes. Les méthodes de classification non hiérarchiques partitionnent les données en classes. Les méthodes de classification hiérarchiques font plus. Ces dernières méthodes partitionnent les données en classes qui s'agrègent les unes aux autres pour former des groupes toujours moins nombreux mais plus hétérogènes, jusqu'à retourner à l'ensemble des données initial. L'ordre d'agrégation détermine la proximité relative entre les classes. On appelle dendrogramme l'arbre qui représente de telles successions de partitions, à l'image des arbres phylogénétiques.

Une classification hiérarchique pourrait donc scinder l'ensemble des individus-périodes en classes et indiquer un degré de proximité entre ces classes. Il faut avant cela choisir quel sera

12. Consulter `http://www.cnrtl.fr/definition/classification`

le critère de détermination des classes. Plusieurs solutions sont envisageables et le choix s'est porté sur la méthode de Ward [13].

La méthode de Ward se base sur le critère de l'inertie. On a déjà vu que l'inertie permet de mesurer la qualité de la projection des données, mais ce n'est pas de l'inertie du nuage projeté dont il sera question ici. Dans un espace euclidien – type d'espace dans lequel on a toujours supposé travailler –, l'inertie totale d'un nuage de point ne varie pas si l'on considère la somme des inerties des partitions possibles de ce nuage et des inerties entre les centres de gravité de ces partitions, d'après le théorème de Huygens. L'inertie totale d'un système est égale à la somme des inerties intra-classes et de l'inertie inter-classes, ce qu'illustre la figure 3.9 [14]. Cette propriété permet de comprendre le critère de classification de Ward qui est une classification ascendante.



(a) Inertie totale (b) Inertie intra-classes (c) Inertie inter-classes

FIGURE 3.9 – Illustration du théorème de Huygens

De nombreux algorithmes de classification sont disponibles sur le marché. En effet, définir une méthode de classification hiérarchique constitue un problème identique à établir le(s) critère(s) de regroupement de deux classes, et plusieurs solutions sont envisageables. Nous avons choisi d'appliquer l'algorithme de Ward. Le critère qui caractérise cet algorithme est la minimisation, à chaque étape, de l'augmentation de l'inertie intraclasse. Pour rappel, l'inertie est une manière de mesurer la dispersion des données autour de leur centre de gravité. Plus l'inertie est petite, plus les données sont concentrées. En vertu du théorème de Huygens, l'inertie totale d'un système est égale à la somme des inerties intraclasses et de l'inertie interclasse, où l'inertie interclasse est l'inertie calculée à partir des centres de gravité des classes

13. [Bouroche et Saporta, 1980], p. 48-62
14. Cette figure est inspirée de [Escofier et Pagès, 2008], p. 299.

et du centre de gravité global du système. L'inertie totale du système reste constante. Expliquons à présent pourquoi l'inertie intraclasse augmente au fur et à mesure du regroupement des individus en classes.

Au départ le nombre de partition est de n, c'est le nombre d'individus. Dans ce cas de figure, l'inertie inter-classe est maximale et égale à l'inertie du nuage de points, l'inertie intra-classe est minimale et vaut 0 car chaque individu est seul dans sa classe. A chaque étape, la méthode de Ward fait passer le nombre de classes de k à $k-1$ en fusionnant deux classes. Ces deux classes sont choisies entre toutes celles définies de sorte que l'inertie entre elles soit minimale. C'est une façon de regrouper les deux classes les plus proches. A chaque étape, l'inertie inter-classe totale du système diminue, mais le moins possible, et au contraire, l'inertie intra-classe augmente. C'est encore une métaphore qui va permettre de visualiser le processus. Dans un champ de vigne, les raisins vont d'abord être groupés en grappes, ensuite par pied de vigne, puis par zone et la classe totale sera tout le champ.

Classification de nos données

La figure 3.10 représente le dendrogramme obtenu en appliquant la méthode de Ward aux individus-périodes. Cette méthode permet de regrouper presque sans incohérence historique les périodes chronologiquement proches par groupes. La seule incohérence est de d'abord classer les groupes qui correspondent aux XX^{e} et XXI^{e} siècles avec celui qui correspond au XVIII^{e} siècle et de regrouper seulement ensuite le groupe qui correspond au XIX^{e} siècle. D'après l'algorithme, cela signifie que le XX^{e} est plus proche du XVIII^{e} que du XIX^{e}, lorsque les périodes sont caractérisées par la fréquence d'apparition de certains mots. Ceci étant, ce résultat n'est pas une erreur ; parfois les données contredisent l'intuition.

Le résultat de l'application de cette méthode permet plusieurs affirmations.

— L'évolution de la fréquence d'apparition de certains mots (au moins) ne se fait pas au hasard. En effet, on ne pourrait dans le cas contraire expliquer un regroupement en périodes proches chronologiquement.

— La fréquence des mots étudiés caractérise les périodes de sorte que la fréquence de ces mots étudiés dans un corpus permet d'attribuer une période à ce corpus.

— En considérant trois classes distinctes, on retrouve une partition qui correspond aux périodes de l'ancien français, du moyen français et du français moderne. Les classes observées correspondent sensiblement à ces états de langue définis par les linguistes

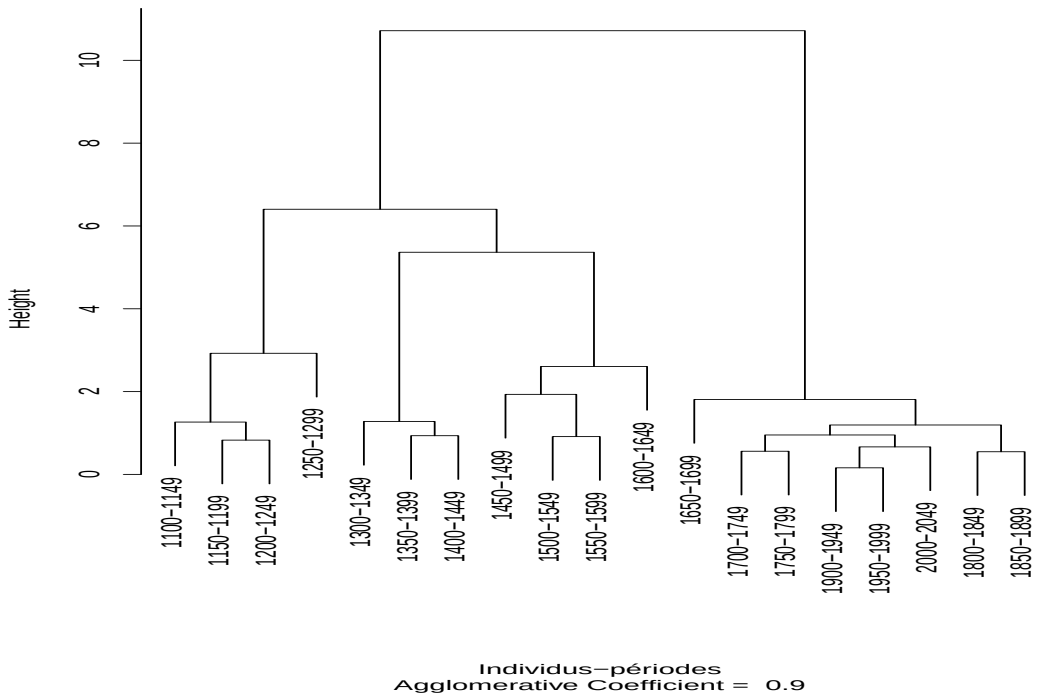


FIGURE 3.10 – Classification ascendante hiérarchique des périodes

sur la base d'autres critères.

— Cette partition en trois classes regroupe les mêmes individus qu'une partition en trois classes effectuées sur base de la projection de ces individus dans le plan factoriel.

3.4 Test

Le tableau 3.3 contient les fréquences relatives d'apparition, en millionièmes, des mots « car », « dans », « dedans », « par », « parce », et « pour » dans « Le mireoir hystorial » de Jean de Vignay. Ce texte a probablement été écrit durant la troisième décennie du XIV^{e} siècle, il était achevé au plus tard en 1332. Le tableau représente donc un vecteur de 6 dimensions.

TABLE 3.3 – Taux (en millionièmes) d'apparition des mots dans « Le mireoir hystorial »

car	dans	dedans	par	parce	pour
4308	15	585	8829	5	6281

Ce vecteur peut être projeté dans le plan ACP déterminé plus haut. La figure 3.11 montre ce plan avec les périodes qui y sont projetées. On observe que les périodes sont cette fois en couleur. Ces couleurs définissent des classes de périodes qui sont celles trouvées lorsque l'on a retreint à 3 le nombre de classes construites par le dendrogramme. Le point qui correspond à la projection du vecteur « Vignay » est indiqué en noir.

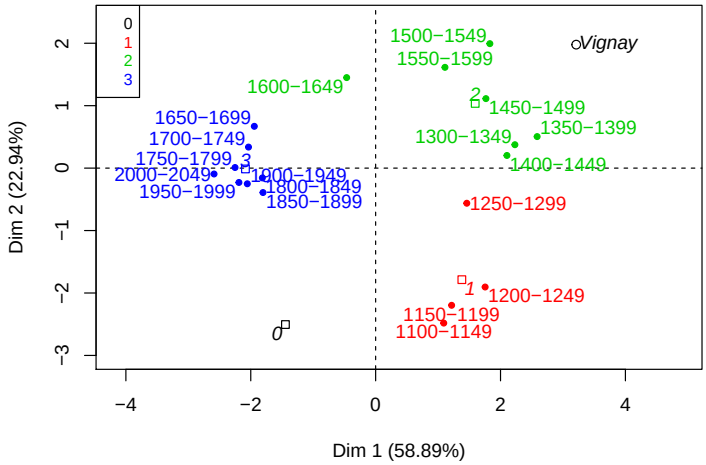


FIGURE 3.11 – Projection du vecteur-Vignay sur le plan ACP

Chaque période étant caractérisée par les mêmes variables lorsque les mots « et » et de » sont laissés de côté, la façon la plus simple d'attribuer une période au texte est de trouver la période qui minimise la distance avec le texte. Nous choisissons la distance euclidienne, la plus intuitive. La figure 3.12 montre un histogramme de toutes les distances calculées. La période la plus proche selon cette mesure est la période 1400-1449. Les deux autres les plus proches sont les périodes, par ordre de distance croissant, les périodes 1350-1399 et 1300-1349. La période exacte se trouve donc parmi les trois premières.

Le texte de Vignay, caractérisé par la fréquence de certains mots, n'est pas placé au plus près de la période où il a été écrit. Cependant, il est projeté dans le plan ACP près des périodes qui lui sont chronologiquement proches et la distance mesurée confirme cette proximité. Il faut tempérer le manque relatif de précision par la qualité des textes qui ont servi à constituer la base de données Frantext, surtout pour les périodes médiévales. En effet, les principes de respect des manuscrits n'ont pas toujours été comparables à ceux que nous

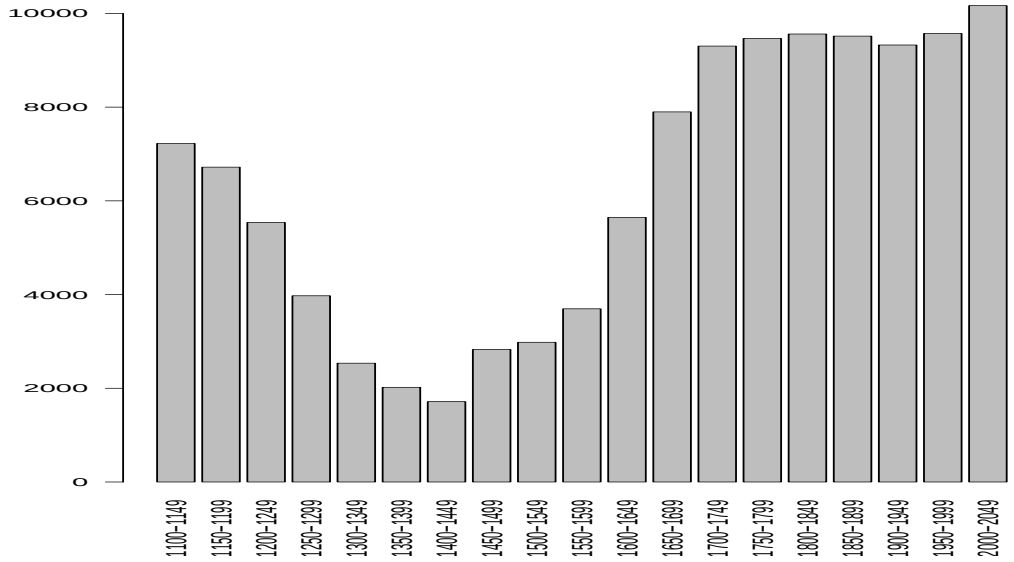


FIGURE 3.12 – Histogramme des distances entre le vecteur-Vignay et les vecteurs-périodes

connaissons aujourd'hui, et encore observe-t-on différentes écoles. Certains éditeurs procèdent à beaucoup plus d'interventions que d'autres : ils interprètent, amendent, délient certaines graphies attachées, ... Dans ce contexte, calculer la fréquence des mots est délicat, quoique nous ayons procédé à une recherche du plus grand nombre de variantes, pour compenser les problèmes dus à la variation graphique et à l'intervention des éditeurs.

3.5 Conclusions

Les linguistes et les historiens définissent les périodes de la langue avec des critères parfois flous. Par exemple, l'ancien français serait une langue à déclinaisons, le français moderne pas et le moyen français serait un entre-deux. Nous avons démontré que ces périodes pouvaient être déterminées par les fréquences relatives d'apparition des mots.

Puisque les fréquences des mots dans un texte permet de caractériser des corpus dont les textes sont écrits pendant un certain intervalle temporel, étudier ces fréquences sur un texte permet de le dater. Pour cela, nous avons utilisé plusieurs méthodes de classification basées sur l'inertie.

De toute évidence, d'après la classification hiérarchique, il y a un avant et un après 1650.

L'évènement politique qui pourrait être responsable d'un changement dans les pratiques observées est la fondation de l'Académie française en 1634, par Richelieu. Après cette fondation, les pratiques linguistiques varient moins, c'est ce que montre le dendrogramme (figure 3.10 p. 34).

Chapitre 4

Variation graphique, lemmatisation et désambiguïsation

4.1 Lemmatisation en moyen français : principe et outil

4.1.1 La lemmatisation, les dictionnaires

Le processus de correspondance entre les formes présentes dans un texte et celles présentes dans un lexique – les lemmes – s'appelle la lemmatisation. Cette application de correspondance entre ensembles n'est ni injective ni surjective : à une forme d'un texte peut correspondre plusieurs lemmes, différentes formes peuvent renvoyer à un unique lemme et il se peut qu'à une certaine forme du texte ne corresponde aucun lemme du lexique.

Ces problèmes d'adéquations se posent particulièrement lorsqu'on lemmatise un texte en moyen français. En effet, d'une part « il existe un grand nombre de possibilités pour une même forme morphosyntaxique et [d'autre part] une liste complète des formes est impossible à établir » [**?**, p. 152].

Constituer un dictionnaire reprenant toutes les formes possibles n'est pas réalisable. D'abord parce que nous n'avons pas un accès numérique à tous les textes écrits en moyen français, ensuite parce que, même alors, il faudrait tous les avoir lemmatisés à la main mais, dans ce cas, le problème d'attribution ne se poserait plus. C'est pourquoi il faut construire une façon de chercher dans un dictionnaire qui ne nécessiterait pas de connaître le lemme de

la forme dont on dispose : il arrive en effet de tomber sur une forme qui nous est totalement inconnue. Cela n'est rendu possible que par l'apparition des dictionnaires électroniques et de différents outils aidant à leur consultation, tel LGeRM.

4.1.2 LGeRM

LGeRM [1] est un lemmatiseur. Ce programme s'occupe de proposer des lemmes aux formes qui lui sont fournies, indépendamment du contexte dans lequel elles apparaissent. Autrement dit, lorsque nous donnons à LGeRM un texte à lemmatiser, il le fait en analysant chaque mot séparément, sans tenir compte des autres. LGeRM renvoie un certain nombre de lemmes possibles pour chaque forme. Il y a donc un certain bruit – *i.e.* un certain nombre de lemmes qui ne sont pas pertinents – mais le lemme correct s'y trouve la plupart du temps. Pour détailler un peu, « le principe général de l'algorithme n'est pas, comme dans la lemmatisation classique, d'essayer de trouver la forme normalisée du lemme, mais de trouver une forme connue dans la base de graphies la plus proche possible de la forme à lemmatiser. » [**?**, p. 160] LGeRM renvoie ensuite le lemme associé à la forme connue dans la base de graphies. LGeRM a été initialement développé pour le moyen français et est actuellement capable de traiter les textes écrits entre 1300 et nos jours. Pour les siècles qui précèdent, LGeRM donne des résultats satisfaisants dans un premier temps mais perfectibles. On l'a vu, la langue n'est pas la même.

4.2 Etiquetage morphosyntaxique

On nomme « *tagging* » ou « étiquetage » l'opération qui consiste à associer une étiquette à un mot. Dans le cas d'une étiquette syntaxique, l'usage est d'employer « part-of-speech tagging » (abrévié POS-tagging). Il existe plusieurs façons de procéder que nous allons décrire sommairement [2].

1. LGeRM, Lemmes Graphies et Règles Morphologiques, voir `http://www.atilf.fr/LGeRM/`
2. Nous renvoyons pour toute cette section à [**?**, chapter 5, pp. 157-206].

4.2.1 Approche par règles

L'étiquetage par règles nécessite deux étapes préalables. La première consiste en l'écriture d'un dictionnaire reprenant toutes les formes, auxquelles est associée une liste de leur(s) étiquettes(s) potentielles(s). La seconde consiste en l'écriture de règles de désambiguïsation qui permettent de déterminer l'étiquette adaptée pour chaque forme. Par exemple, voici la règle qui permet de déterminer si « bien » est un nom commun en moyen français [3] :

Algorithme 1 : « bien » - nom commun

Input : « bien »

if

déterminant
+ *(facultativement) adjectif qualificatif*
+ *bien*
+ ¬*(participe* ∪ *adjectif qualificatif* ∪ *adverbe)*

then « bien » ← « nom commun » ;

Écrire ce type de règle pour chaque forme représente un travail colossal. De plus, il ne faut pas oublier que certaines ambiguïtés ne peuvent être levées sans un recours à la sémantique ; un des exemples les plus célèbres pour illustrer ce cas doit être : « J'ai mangé un avocat ». Sans négliger de réfléchir à certaines règles qui permettraient de déterminer l'étiquette d'une forme dans tous les cas, il faut reconnaître que cette approche n'est pas la plus rapide.

4.2.2 Approche statistique

Soit un corpus de texte dont toutes les formes ont été étiquetées. Déterminons, pour chaque forme, la probabilité qu'elle soit étiquetée de telle ou telle façon, en fonction de la séquence d'étiquettes qui précède. Lorsque nous étiquetons une phrase qui n'est pas présente dans notre corpus d'entraînement, nous choisissons la suite d'étiquettes qui maximise la probabilité d'apparition de cette suite, en fonction des probabilités qui ont été dégagées du corpus d'entraînement.

3. Adapté de [**?**, p. 132]

4.2.3 Approche mixte : *Transformation-Based Tagging*

Nous appelons cette approche « mixte » car elle est inspirée des deux précédentes. Comme la première, elle fait usage de règles ; mais ces règles sont induites automatiquement à partir d'un corpus – ce qui renvoie à la deuxième approche – et non pas générées *ex nihilo* par le linguiste.

4.3 Exposé du problème

Le moyen français n'est pas *une* langue. Ce mot composé (« moyen français ») désigne un ensemble de pratiques langagières ayant eu cours entre 1350 et 1550 (grosso modo) dans les territoires « de la langue d'oïl ». Ces pratiques, orales et écrites, ne sont pas unifiées et sont au contraire très variables dans l'espace, le temps et les situations de production. De cela il nous reste, dans les textes, une énorme variation graphique. L'écriture, à l'époque, n'est pas normée. Il n'existe donc pas un ensemble de règles qui permettraient de prédire sous quelle forme graphique peut se réaliser tel ou tel mot. Dès lors, effectivement, une liste exhaustive est impossible et, lorsque l'on rencontre une forme au fil du texte, il faut en trouver une liste de lemmes de provenance possible. C'est la tâche de LGeRM. C'est donc bien les particularités de variation du moyen français qui ont conduit à faire développer LGeRM comme un producteur de lemmes potentiels. Mais comment déterminer le bon lemme parmi ceux proposés ?

L'opération d'étiquetage morphosyntaxique peut s'avérer utile pour désambiguïser certaines formes d'un texte. Prenons par exemple la phrase suivante : « Et le felon paien s'en courut droit a luy. »[4]. Un lecteur comprendra sans peine que le « a » est une forme de la préposition « à » et non une forme du verbe « avoir » ; ce que ne comprendrait pas un lemmatiseur comme LGeRM. L'assignation de l'étiquette « préposition » à ce mot permettrait au lemmatiseur de déterminer, dans ce cas-ci, de façon univoque le lemme correspondant à la forme étudiée. Cette opération d'étiquetage morphosyntaxique permettrait en tout cas de réduire ce bruit inhérent à LGeRM.

4. Galien D.B., c.1400-1500, 88, exemple trouvé à l'adresse : `http://www.atilf.fr/dmf/definition/à`

Chapitre 5

TreeTagger

5.1 Introduction

Saussure soutient que « dans un discours, les mots contractent entre eux, en vertu de leur enchaînement, des rapports fondés sur le caractère linéaire de la langue, qui exclut la possibilité de prononcer deux éléments à la fois » [1]. Ou, pour l'exprimer autrement, ce serait la place relative des mots entre-eux qui déterminerait en partie leurs relations, et non uniquement les caractéristiques morphologiques dont ils sont porteurs. Il est possible de définir par ce fait les langues analytiques, groupe dont font partie le français et, dans une large mesure, le moyen français.

L'ordre des mots d'un discours dans l'une de ces deux langues sera donc nécessairement dépendant des relations que, dans les discours qui ont précédés, les locuteurs ont attribués aux mots. Tous les mots n'ont pas la même probabilité d'apparition à la i$^{\text{ème}}$ place d'une séquence donnée. Il existe naturellement plusieurs moyens pour utiliser ces relations et en rendre compte : c'est d'ailleurs une des définitions de la grammaire. La méthode que nous allons présenter dans les paragraphes qui suivent est due au mathématicien russe Andreï Markov (1856–1922).

Grâce aux chaînes dites ,« de Markov », il est d'une certaine façon possible de « prédire » le futur. Ainsi, dans notre cas, connaissant une suite de mots, nous pourrions prédire le mot suivant comme dans, par exemple, la phrase suivante : « Quelle heure est-. . . ? ». Si nous

1. [de Saussure, 1994], p. 170

remplaçons cette suite de mots par une suite d'étiquettes, l'exercice serait équivalent pour une machine car le problème abstrait reste identique. Il s'agit de déterminer, dans un contexte donné – ici une suite d'états –, quel sera l'état suivant.

Ce chapitre présente les éléments nécessaires à une compréhension globale de ces chaînes ainsi que différents algorithmes qui permettent de prédire l'étiquette qui apparait après une suite d'étiquettes donnée. D'abord nous présenterons les chaînes de Markov, ensuite la méthode du maximum de vraisemblance pour calculer la probabilité de transition, l'algorithme de Viterbi, puis l'algorithme ID3, utilisé par le logiciel TreeTagger. Le contenu de ce chapitre doit massivement à [Jurafsky et Martin, 2009] et [Quinlan, 1986].

5.2 Les N-grammes

Une *expérience aléatoire* est une expérience dont on ne peut prédire le résultat. Une *variable aléatoire* X est une variable associée à une expérience aléatoire. Dans notre cas où l'expérience aléatoire consiste à rencontrer des mots d'une séquence, les uns après les autres, l'ensemble des valeurs que peut prendre la variable aléatoire est le lexique. Dit autrement, observer une séquence de n mots est le résultat de n expériences aléatoires X et le mot m_{i} est le résultat de l'expérience X_{i}, pour i variant de 1 à n.

Représentons une séquence de n mots par $m_1 m_2 \dots m_{\mathrm{n}}$ ou m_1^n. On note la probabilité d'observer une telle séquence de mots par $P(X_1 = m_1, X_2 = m_2, \dots, X_{\mathrm{n}} = m_{\mathrm{n}})$, ou, plus succinctement, $P(m_1 m_2 \dots m_{\mathrm{n}})$.

$$P(m_1 m_2 \dots m_{\mathrm{n}}) = P(m_1) \cdot P(m_2|m_1) \cdot P(m_3|m_1 m_2) \cdot \dots \cdot P(m_{\mathrm{n}}|m_1 \dots m_{\text{n-1}}) \tag{5.1}$$

$$= \prod_{k=1}^{n} P(m_{\mathrm{k}}|m_1^{k-1}) \tag{5.2}$$

où $P(x|y)$ est la probabilité de réalisation de l'événement x sachant que l'événement y est réalisé.

L'idée de laquelle procèdent les modèles de N-grammes est que l'on peut approximer l'histoire d'un mot (*i.e.* tous les mots qui le précèdent) par une partie d'entre eux. Cette

idée, due à A. Markov, permet la modélisation de très nombreux problèmes et des solutions numériques très fines. On retreint donc l'historique à une séquence de taille N dont le N$^{\text{ème}}$ mot est celui qui nous intéresse. La probabilité d'apparition du n$^{\text{ème}}$ mot s'approxime donc par

$$P(m_n|m_1^{n-1}) \approx P(m_n|m_{n-N+1}^{n-1}) \tag{5.3}$$

et l'équation 5.2, qui rend compte de la probabilité d'apparition de la séquence, devient

$$P(m_1 m_2 \dots m_n) \approx \prod_{k=1}^{n} P(m_k|m_{k-N+1}^{k-1}). \tag{5.4}$$

Une des questions qui va nous occuper tout au long de ce chapitre sera celle du calcul (ou de l'estimation) de la probabilité de ces N-grammes.

La définition classique de probabilité, à savoir le rapport du nombre de cas favorables au nombre de cas possibles, est la plus simple. Dans notre cas, la probabilité d'apparition d'un mot, étant donnés les $N-1$ mots le précédant, est le rapport du nombre d'observations de la séquence des $N-1$ mots suivie du mot dont on cherche la probabilité au nombre de cas total d'apparition de la séquence des $N-1$ mots :

$$P(m_n|m_{n-N+1}^{n-1}) = \frac{C(m_{n-N+1}^{n-1} m_n)}{C(m_{n-N+1}^{n-1})} \tag{5.5}$$

où $C(a)$ est le nombre de cas d'apparition de la séquence « a ». Et ainsi,

$$P(m_n|m_1^{n-1}) \approx \frac{C(m_{n-N+1}^{n-1} m_n)}{C(m_{n-N+1}^{n-1})}. \tag{5.6}$$

Les comptes sont effectués sur le corpus d'entrainement.

Il pourrait arriver d'observer, dans le corpus de test, une séquence qui n'aurait pas été rencontrée dans le corpus d'entrainement. Dans ce cas, nous aurions $C(m_{n-N+1}^{n-1}) = 0$, ce qui rendrait la définition caduque (on ne peut diviser par 0). Pour éviter cet achoppement, on définit la probabilité d'apparition d'un mot après une séquence comme nulle si la séquence n'est pas rencontrée dans le corpus d'entrainement. Notons $C(h)$ le nombre de fois que l'on observe la séquence « h » (l'historique) et $C(h, m)$ le nombre de fois que cet historique est

suivi du mot m ». Nous définissons

$$P(m_n|m_1^{n-1}) \coloneqq \begin{cases} \frac{C(h,m)}{C(h)} & \text{si } C(h) > 0 \\ \\ 0 & \text{sinon.} \end{cases} \tag{5.7}$$

5.3 HMM

Nous avons présenté jusqu'à présent une méthode pour calculer la probabilité d'apparition d'une forme étant donné une séquence qui la précède. Or, le problème qui nous occupe n'est pas exactement celui-là. La question est de déterminer la meilleure séquence d'étiquettes qui corresponde à une séquence de formes donnée.

Notons t_1^n une séquence d'étiquettes (*tag*, en anglais) et m_1^n la séquence de formes. Une solution au problème serait de choisir la séquence d'étiquettes $\hat{t}_1^n$ qui maximise la probabilité d'apparition de cette séquence étant donné la séquence de formes, soit $P(t_n^n|m_1^n)$. Par définition :

$$\hat{t}_1^n \coloneqq \underset{t_1^n}{\operatorname{argmax}} P(t_t^n|m_1^n). \tag{5.8}$$

Le problème revient à définir $P(t_t^n|m_1^n)$ et à la calculer.

Le théorème des probabilités conditionnelles

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{P(y)} \tag{5.9}$$

permet d'exprimer autrement $\hat{t}_1^n$. En remplaçant la probabilité dans l'équation 5.8, nous avons

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \frac{P(m_1^n|t_1^n) \cdot P(t_1^n)}{P(m_1^n)} \tag{5.10}$$

or, puisque $P(m_1^n)$ ne dépend pas de t_1^n, cette probabilité peut être enlevée de l'équation 5.10 sans en modifier le résultat, ce qui donne

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(m_1^n|t_1^n) \cdot P(t_1^n). \tag{5.11}$$

Cette dernière équation va être simplifiée par deux hypothèses.

D'abord, on admet que la probabilité d'apparition d'un mot dépende uniquement de son étiquette, et pas des mots et des étiquettes qui le précèdent, ce qui permet d'approximer

$$P(m_1^n|t_1^n) \approx \prod_{i=1}^{n} P(m_i|t_i). \tag{5.12}$$

Ensuite, on se rappelle de l'idée de Markov et de considérer des N-grammes plutôt que la totalité des séquences. Nous avons donc, dans le cas de bigrammes, $P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})$ et, dans le cas de trigrammes, $P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1}t_{i-2})$.

En remplaçant dans l'équation 5.11, nous obtenons que

$$\hat{t}_1^n \coloneqq \underset{t_1^n}{\operatorname{argmax}}\, P(t_t^n|m_1^n) \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(m_i|t_i) \cdot P(t_i|t_{i-1}) \tag{5.13}$$

La probabilité $P(t_i|t_{i-1})$ est appelée *probabilité de transition* et $P(m_i|t_i)$, la probabilité d'apparition d'un mot étant donné une étiquette, est appelée la *probabilité d'émission*. On peut calculer ces probabilités de façon classique et on a alors

$$P(t_i|t_{i-1}) = \frac{C(t_{i-1}, t_i)}{C(t_{i-1})}\ , \qquad\qquad P(m_i|t_i) = \frac{C(t_i, m_i)}{C(t_i)}. \tag{5.14}$$

Nous avons donc une méthode pour calculer la séquence d'étiquettes la plus probable pour une séquence de mots.

Pour clore cette section, une formalisation de ce qui vient d'être présenté.

Une chaîne de Markov est un automate à états finis probabiliste dont l'état initial, c'est-à-dire l'argument d'entrée de la séquence, détermine les différents états que l'automate va occuper. Un modèle de Markov caché (*hidden Markov model*, abrégé dorénavant HMM) nous permet de rendre compte à la fois des événements observés (les mots dans une séquence) et des événements cachés (les étiquettes qui leur sont associées). Plus formellement, déterminer un HMM nécessite :

$Q = q_1 q_2 \dots q_n$	un ensembles de n états,
$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}$	une matrice A des probabilités de transition, chaque a_{ij} étant la probabilité de passer de l'état i à l'état j, avec $\sum_{j=1}^{n} a_{ij} = 1, \forall i$
$O = o_1 o_2 \dots o_T$	un séquence de T observations, chacune provenant du lexique $V = v_1, v_2, \dots, v_k$
$B = b_i(o_t)$	une séquence de probabilité d'émission, chacune exprimant la probabilité d'observer o_t à l'état i
q_0, q_F	un état initial et un état final qui ne sont pas associés aux observations, avec des probabilité de transition $a_{01} a_{02} \dots a_{0n}$ et $a_{1F} a_{2F} \dots a_{nF}$

5.4 Algorithme de Viterbi

On appelle « décodage » la tâche qui consiste à trouver une séquence d'états la plus probable étant donné une séquence d'observations. Cette tâche est précisément la tâche d'étiquetage que nous cherchons à résoudre. Dans la section précédente, l'emploi des chaines de Markov a permis de trouver que la séquence la plus probable était la séquence qui maximisait un produit de probabilités :

$$\hat{t}_1^n \approx \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(m_i|t_i) \cdot P(t_i|t_{i-1}) \tag{5.15}$$

Toute la question, à présent, est de calculer ce produit et de trouver la séquence d'étiquettes qui le maximise.

Supposons que nous ayons un HMM déjà paramétré. Considérons la séquence de T observations $O = o_1, o_2, \dots, o_T$ et la séquence $Q = q_1, q_2, \dots, q_N$ des N états possibles. L'algorithme de Viterbi consiste à déterminer deux matrices, l'une contenant les suites d'états et l'autre la probabilité associée à ces suites. Il s'agit d'un algorithme de programmation dynamique : les mauvaises solutions sont éliminées au fur et à mesure, il ne faut pas tout calculer car le nombre de suites d'états pour T observations s'élève à N^T.

Soient la matrice, dite de Viterbi, $V = (v_{n,t})_{1\leq n\leq N,\ 1\leq t\leq T+1}$ des probabilités et $M = (m_{n,t})_{1\leq n\leq N,\ 1\leq t\leq T+1}$ la matrice qui tient les séquences en mémoire. Chaque ligne de ces matrices correspond à un état et chaque colonne à une observation. Les trois étapes pour les remplir sont :

1. Initialisation :

$$v_{n1} = a_{0n} \cdot b_n(o_1), \qquad 1 \leq n \leq N \tag{5.16}$$

$$m_{n1} = 0, \qquad 1 \leq n \leq N \tag{5.17}$$

2. Récursion :

$$v_{nt} = \max_{i=1}^{N} v_{i(t-1)} \cdot a_{ij} \cdot b_n(o_t), \qquad 1 \leq j \leq N,\ 2 \leq t \leq T \tag{5.18}$$

$$m_{nt} = \operatorname*{argmax}_{i=1}^{N} v_{i(t-1)} \cdot a_{ij} \cdot b_n(o_t), \qquad 1 \leq j \leq N,\ 2 \leq t \leq T \tag{5.19}$$

3. Fin :

$$v_{n(T+1)} = \max_{i=1}^{N} v_{i(T)} \cdot a_{iF} \tag{5.20}$$

$$m_{n(T+1)} = \operatorname*{argmax}_{i=1}^{N} v_{i(T)} \cdot a_{iF} \tag{5.21}$$

et ainsi chaque v_{nt} contient la probabilité maximale de faire l'observation o_t à l'état n. Cette probabilité est le produit de

— la probabilité d'émission de o_t à l'état n, $b_n(o_t)$,

— la probabilité de transition d'un des états précédents à l'actuel,

— la probabilité précédente calculée de la même façon.

En effet, la probabilité maximale de se trouver à l'état i ($1 \leq i \leq N$) au moment de l'observation o_1 est la probabilité de commencer par cet état, multipliée par la probabilité de faire l'observation o_1 à cet état ($b_n(o_1)$), ce que l'on place dans la première ligne de la matrice. Pour trouver la probabilité maximale de se trouver à l'état n au moment de l'observation o_2, il faut trouver la probabilité de faire cette observation à cet état ($b_n(o_2)$) multipliée par la probabilité précédente($a_{0n} \cdot b_n(o_1)$), et par la probabilité de transition entre l'état précédent et l'actuel (a_{kn}) pour calculer la probabilité de la séquence. Il faut faire ce calcul pour tous les états précédents possibles ($1 \leq k \leq N$) et ne conserver que la probabilité maximale afin

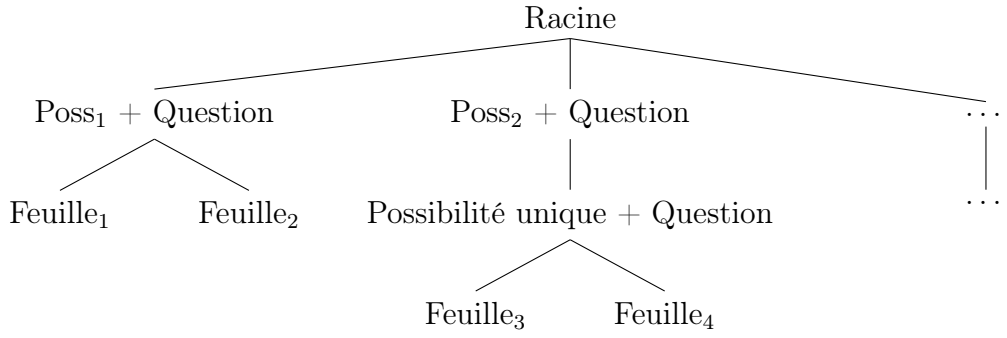
de trouver la séquence la plus probable. Et ainsi de suite. La probabilité finale s'obtient en trouvant le maximum du produit de la probabilité calculée pour chaque état possible en dernière observation (v_nT) et la probabilité de terminer sur cet état (a_{iF}).

5.5 Algorithme ID3

Il existe de nombreux algorithmes qui calculent les probabilités de transition. Cette section est consacrée à la présentation détaillée de l'algorithme qui est utilisé par TreeTagger [2]. Elle est très inspirée d'un article séminal de Quinlan ([Quinlan, 1986]) où l'auteur propose un algorithme qui construit un arbre de décision pour calculer les probabilités de transition.

Un arbre de décision est la représentation graphique, en forme d'arbre, d'un ensemble de choix qui se posent les uns après les autres. En partant de la racine de l'arbre, il faut suivre la branche qui répond à la question posée sur le nœud où l'on se trouve. Chaque embranchement – chaque nœud – mène à différentes possibilités. Les nœuds finaux, les feuilles, contiennent l'information que l'on recherche.

FIGURE 5.1 – Arbre de décision type



5.5.1 La tâche d'induction

Maintenant que l'on comprend le principe d'un arbre de décision, il faut comprendre ce qu'est la tâche d'induction. Postulons que le monde est une collection d'*objets* qui sont décrits et caractérisés par des *attributs*. Un attribut mesure une caractéristique d'un objet.

2. Voir [Schmid, 1994] et [Schmid, 1995]

Les valeurs que peuvent prendre les attributs sont mutuellement exclusives et, surtout, se limitent à un ensemble fini ou, alors, peuvent se résumer à un ensemble fini.

L'exemple qui suit est tiré de l'article de Quinlan déjà cité. Si les objets étaient des samedis matins et que l'objet était la météo, les attributs pourraient être

— nébulosité, avec des valeurs {ensoleillé, couvert, pluvieux}
— température, avec des valeurs {froid, doux, chaud}
— humidité, avec des valeurs {élevé, normal}
— venteux, avec des valeurs {vrai, faux}.

Admettons à présent que chaque objet de l'univers appartienne à une classe d'un ensemble de classes mutuellement exclusives. Pour simplifier, nous continuerons avec les deux classes notées P et N, parfois appelées respectivement les classes d'*instances positives* et d'*instances négatives*.

Considérons que nous sommes en possession d'un *ensemble d'apprentissage*, c'est-à-dire un ensemble d'objets dont la classe est connue. La tâche d'induction consiste à développer une *règle de classification* qui permet de déterminer la classe de tout objet à partir de la valeur de ses attributs. Il faut avant tout vérifier que les attributs permettent de trouver une telle règle. Par exemple, si deux objets sont caractérisés par les mêmes valeurs d'attributs, il n'est pas possible d'établir une règle sur la base de ces seuls attributs. Dans ce cas, on dit que les attributs sont inadéquats.

Le tableau 5.1 contient un exemple d'ensemble d'entraînement dont les attributs des objets le constituant permettent de trouver une règle de classification. La figure 5.2 représente un arbre de décision qui classe correctement chaque objet de l'ensemble d'apprentissage. De façon générale, les feuilles d'un arbre de décision qui a pour tâche de classifier des objets contiennent le nom d'une classe. Les autres noeuds, on l'a dit, représentent des tests basés sur des attributs avec une branche pour chaque résultat possible. Donc, pour classer un objet, il faut commencer à la racine de l'arbre, évaluer le test et suivre la branche appropriée au résultat. Le processus d'évaluation et de suivi de branche continue jusqu'à ce qu'une feuille soit rencontrée. A ce moment, l'objet est censé appartenir à la classe nommée par la feuille. Il est possible que seul sous-ensemble des attributs soit testé sur un chemin particulier en partant de la racine de l'arbre de décision jusqu'à une feuille.

Il est important de souligner que, si les attributs sont adéquats, il est toujours possible de construire un arbre de décision qui classifie correctement chaque élément de l'ensemble

TABLE 5.1 – Un petit ensemble d'entraînement

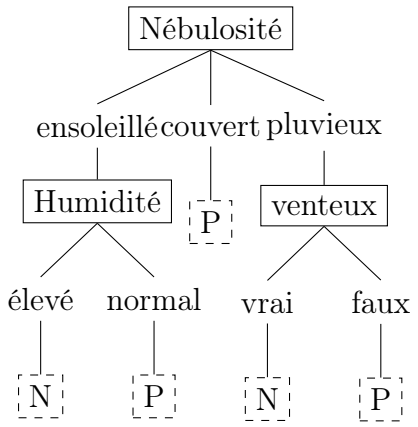
No.	Attributs				Classe
	Perspectives	Températures	Humidité	Venteux	
1	ensoleillé	chaud	élevé	faux	N
2	ensoleillé	chaud	élevé	vrai	N
3	couvert	chaud	élevé	faux	P
4	pluvieux	doux	élevé	faux	P
5	pluvieux	froid	normal	faux	P
6	pluvieux	froid	normal	vrai	N
7	couvert	froid	normal	vrai	P
8	ensoleillé	doux	élevé	faux	N
9	ensoleillé	froid	normal	faux	P
10	pluvieux	doux	normal	faux	P
11	ensoleillé	doux	normal	vrai	P
12	couvert	doux	élevé	vrai	P
13	couvert	chaud	normal	faux	P
14	pluvieux	doux	élevé	vrai	N

d'apprentissage. C'est la pire solution, un chemin correspond alors à un unique objet. Généralement il existe de nombreux arbres de décisions qui classent les éléments d'un ensemble dont les attributs des objets sont adéquats. L'essence de l'induction est de créer une règle qui permette de classer des objets n'appartenant pas à l'ensemble d'entrainement. Pour cela, l'arbre de décision doit capturer une relation significative entre la classe d'un objet et les valeurs de ses attributs. On applique le principe du rasoir d'Occam lorsque deux arbres existent pour un même ensemble d'entrainement. C'est l'arbre le plus simple qui est préféré. Par exemple, l'arbre de décision représentée par la figure 5.3 est également correct pour l'ensemble d'entrainement. Cependant, sa plus grande complexité le rend suspect et il a moins de pouvoir explicatif et prédictif par rapport à un arbre plus simple qui, parce qu'il contient moins de chemins différents, permet de classer plus de données.

5.5.2 ID3

L'algorithme le plus simple pour trouver un arbre de décision correct serait de tous les générer et de choisir le plus simple. Le nombre de ces arbres étant bien trop grand, cette

FIGURE 5.2 – Un arbre de décision simple



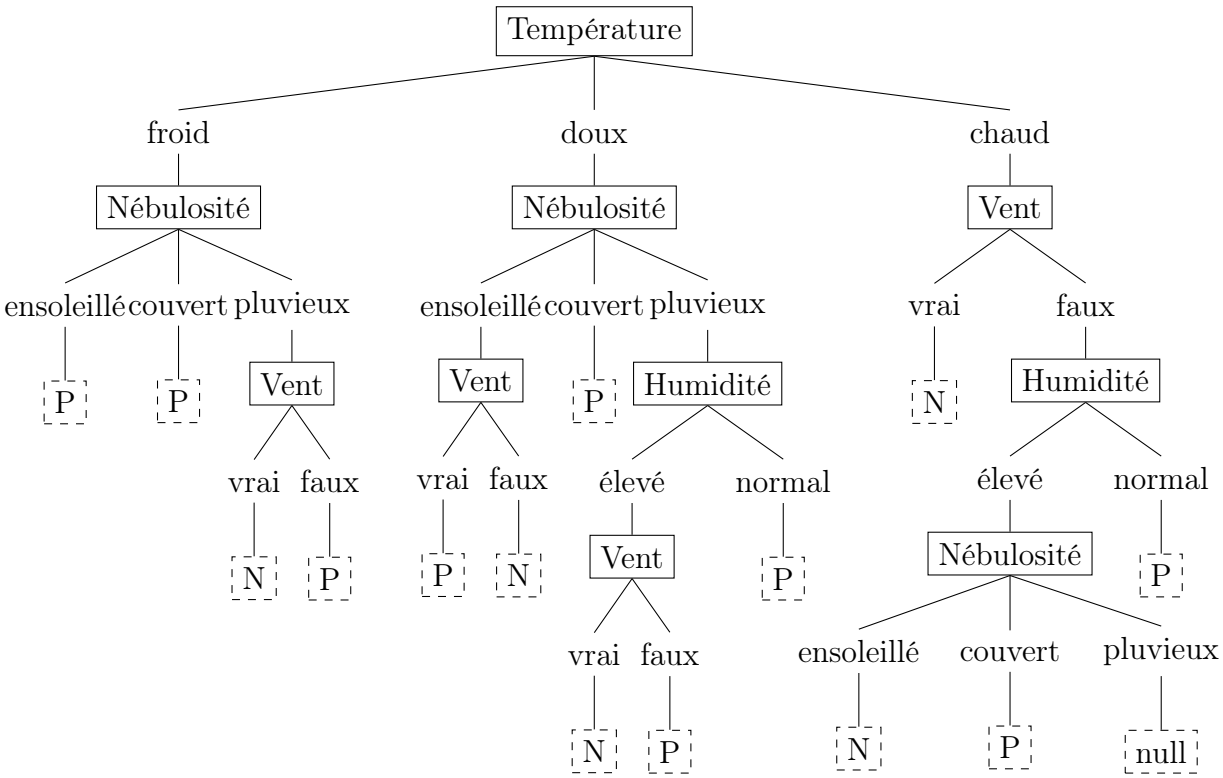
approche ne serait réalisable que pour les petits ensembles d'entrainement. De plus, cette solution manque singulièrement d'intelligence.

L'algorithme ID3 a été pensé pour pouvoir donner une réponse lorsque les ensembles contiennent un grand nombre d'objets et que chaque objet est décrit par de nombreux attributs. Cet algorithme permet de trouver un arbre raisonnablement bon sans trop de calculs, cependant, il ne peut pas garantir que la solution qu'il fournit est optimale. En réalité, la solution est celle d'un minimum local. Parmi les solutions proches, c'est la meilleure mais il n'est pas possible d'affirmer qu'il n'existe pas dans des solutions beaucoup plus différentes une meilleure solution.

ID3 est un algorithme itératif. Un sous-ensemble de l'ensemble d'apprentissage, appelé la *fenêtre*, est choisi au hasard et un arbre de décision est formé à partir de celui-ci. On a que cet arbre, par construction, classe correctement tous les objets de la fenêtre. On utilise ensuite cet arbre pour classer tous les autres objets de l'ensemble d'entrainement. Si la bonne réponse est donnée à chaque fois, alors l'arbre est correct pour tout l'ensemble d'entrainement et le processus prend fin. Si l'arbre ne classe pas correctement tous les objets de l'ensemble d'entrainement, une sélection aléatoire des objets classés de manière incorrecte est ajoutée à la fenêtre et le processus recommence.

L'expérience montre que des arbres de décisions corrects ont été trouvés après seulement quelques intérations pour des ensemble comptant jusqu'à 30.000 objets décrits en termes de 50 attributs. Si, comme il a été noté, le cadre itératif ne peut garantir de converger sur un arbre final avant que la fenêtre ne reprenne tous les éléments de l'ensemble d'entrainement,

FIGURE 5.3 – Un arbre de décision complexe

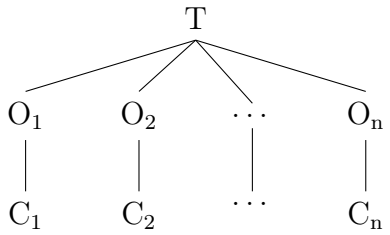


ce cas n'as pas été rencontré dans la pratique. Les constats empiriques montrent qu'un arbre est généralement trouvé très rapidement.

Toute la difficulté revient maintenant que les idées se précisent à former un arbre de décision pour une collection arbitraire C d'objets. Les cas triviaux doivent être mentionnés. Si C est vide ou contient des objets d'une seule classe, l'arbre de décision le plus simple est juste une feuille étiquetée avec cette classe.

Sinon, soit T un test sur un objet avec des résultats possibles $O_1, O_2, \ldots O_n$. Chaque objet dans C donnera un de ces résultats pour T, donc T partitionne C en $C_1, C_2, \ldots, C_n$ avec C_i contenant les objets donnant comme résultat O_i au test T. Cela est représenté graphiquement par la figure 5.4.

FIGURE 5.4 – Un arbre structurant les objets de C



Si pour chaque sous-ensemble C_i on pouvait trouver un arbre de décision qui en classait les objets, le résultat serait un arbre de décision pour C. Il faut remarquer que tant que au moins deux des C_i ne sont pas vides, chaque C_i compte moins d'éléments que C. Cela correspond à la stratégie de « diviser pour régner ». Dans le pire des cas cette stratégie mène à des ensemble ne comptant qu'un élément, ils satisfont alors à l'exigence d'une classe pour une feuille. Ce « pire des cas » est la preuve que l'algorithme mène à une solution et produira toujours un arbre de décision qui classifie correctement chaque objet de C.

Finalement la difficulté repose sur le choix d'un test. On a vu qu'un test doit pouvoir ramifier sur les valeurs d'un attribut. Donc, choisir un test revient d'abord à sélectionner un attribut pour la racine de l'arbre, premier nœud à partir duquel on ramifie. L'algorithme ID3 fait appel à la théorie de l'information et au calcul d'entropie. Soient p objets de la classe P et n objets de la classe N. La méthode est basée sur deux hypothèses :

1. Tout arbre de décision correct pour C classera les objets dans la même proportion que leur représentation dans C. On déterminera qu'un objet arbitraire appartient à la

classe P avec probabilité $p/(p+n)$ et à la classe N avec probabilité $n/(p+n)$.

2. Lorsqu'un arbre de décision est utilisé pour classer un objet, il renvoie une classe. Un arbre de décision peut donc être considéré comme la source d'un message "P" ou "N", avec les informations attendues nécessaires à la génération de ce message donné par

$$I(p,n) = -\frac{p}{p+n} \cdot \log_2(\frac{p}{p+n}) - \frac{n}{p+n} \cdot \log_2(\frac{n}{p+n}) \tag{5.22}$$

Considérons l'attribut A, ayant comme valeurs possibles l'ensemble $\{A_1, A_2, \ldots A_{\mathrm{v}}\}$. Si cet attribut est testé à la racine de l'arbre de décision, il va produire une partition de C en $\{C_1, C_2, \ldots C_{\mathrm{v}}\}$ où C_{i} contient les objets de C qui ont la valeur $\mathrm{A_i}$ de A.

Soit C_{i} les objets p_{i} de classe P et n_{i} de classe N. L'information attendue requise pour les sous-arbres C_{i} est $I(p_{\mathrm{i}}, n_{\mathrm{i}})$. L'information attendue requise pour l'arbre avec A comme racine est alors obtenue comme la moyenne pondérée

$$E(A) = \sum_{i=0}^{v} \frac{p_{\mathrm{i}} + n_{\mathrm{i}}}{p+n} \cdot I(p_{\mathrm{i}}, n_{\mathrm{i}}) \tag{5.23}$$

où le poids pour la $\mathrm{i}^{\mathrm{ème}}$ branche est la proportion des objets dans C qui appartiennent à C_{i}. L'information obtenue en se ramifiant sur A est donc

$$gain(A) = I(p_{\mathrm{i}}, n_{\mathrm{i}}) - E(A) \tag{5.24}$$

L'idée la plus simple à partir de là est de choisir comme attribut de test servant à créer des branches celui qui maximise le gain d'information. Or, puisque $I(p,n)$ est constant pour tous les attributs, maximiser le gain équivaut à minimiser $E(A)$, qui est l'information mutuelle de l'attribut A et de la classe. L'algorithme ID3 examine tous les attributs candidats et choisit A pour maximiser $gain(A)$, forme l'arbre comme ci-dessus, puis utilise le même processus récursivement pour former des arbres de décision pour les sous-ensembles résiduels $C_1, C_2, \ldots C_{\mathrm{v}}$.

Pour illustrer l'idée, soit C l'ensemble des objets du tableau 1. Parmi les 14 objets, 9 sont de classe P et 5 sont de classe N, donc l'information requise pour la classification est

$$I(p,n) = -\frac{9}{14} \cdot \log_2(\frac{9}{14}) - \frac{5}{14} \cdot \log_2(\frac{5}{14}) = 0.940\ bits \tag{5.25}$$

Considérons maintenant l'attribut *nébulosité* de valeurs {ensoleillé, couvert, pluvieux}. Cinq des 14 objets en C ont la première valeur (ensoleillée), deux d'entre eux de la classe P et trois de la classe N, donc

$$p_1 = 2 \qquad n_1 = 3 \qquad I(p_1, n_1) = 0.971 \tag{5.26}$$

et similairement

$$p_2 = 4 \qquad n_2 = 0 \qquad I(p_2, n_2) = 0 \tag{5.27}$$

$$p_3 = 3 \qquad n_3 = 2 \qquad I(p_3, n_3) = 0.971 \tag{5.28}$$

L'information requise attendue après avoir testé cet attribut est donc

$$E(\textit{nébulosité}) = \frac{5}{14} \cdot I(p_1, n_1) + \frac{4}{14} \cdot I(p_2, n_2) + \frac{5}{14} \cdot I(p_3, n_3) = 0.694 \; \textit{bits} \tag{5.29}$$

Le gain de cet attribut est alors

$$\textit{gain}(\textit{nébulosité}) = 0.940 - E(\textit{nébulosité}) = 0.246 \; \textit{bits} \tag{5.30}$$

Des analyses semblables donnent

$$\textit{gain}(\textit{température}) = 0.029 \; \textit{bits} \tag{5.31}$$

$$\textit{gain}(\textit{humidité}) = 0.151 \; \textit{bits} \tag{5.32}$$

$$\textit{gain}(\textit{vent}) = 0.048 \; \textit{bits} \tag{5.33}$$

Ainsi, la méthode de formation d'arbres utilisée dans ID3 choisirait *nébulosité* comme attribut de la racine de l'arbre de décision. Les objets seraient alors divisés en sous-ensembles en fonction de la valeur de leur attribut *nébulosité* et un arbre de décision pour chaque sous-ensemble serait construit de la même manière. En fait, la figure 5.2 montre l'arbre de décision réel généré par ID3 à partir de cet ensemble d'apprentissage.

Un cas particulier serait celui où la collection C ne contient aucun objet ayant A_i comme valeur de l'attribut A, ce qui donnerait un sous-ensemble C_i vide. L'algorithme attribue la valeur « null » à une feuille de cette sorte. Les objets qui arriveraient à cette feuille ne seraient

alors pas classés. Une autre manière de faire serait de conférer à ces objets la classe la plus fréquente que l'on trouve dans C, et donc attribuer cette classe à la feuille. Ou encore, il serait pensable de modifier l'algorithme pour qu'il attribue proportionnellement à la représentation des classes dans C les feuilles de ce genre. Cela ne favoriserait pas outre mesure la classe déjà majoritaire.

L'évaluation des paramètres du modèle – c'est-à-dire des attributs choisis comme nœuds – peut se faire de plusieurs façons. La simplicité de l'arbre est un critère. Quinlan propose d'évaluer la façon dont l'arbre exprime les relations réelles entre classes et attributs. C'est probablement un non-sens. Un modèle statistique est un ensemble de paramètres : de constantes, de variables et de relations entre elles. Entrainer un modèle revient à déterminer les constantes et les relation entre les variables, les relations entre le variables et les constantes. La prédiction se fait en remplaçant dans les équations constituées les variables par la valeur des attributs des objets à classer. Il n'y a rien au monde qui garantit qu'un modèle est l'expression de la réalité. C'est une faute logique et philosophique. Traditionnellement, en traitement automatique du langage, l'évaluation se fait à l'aide des notions de *précision*, de *rappel* et de *F-mesure*. Ces notions seront présentées en détail au moment de l'évaluation des paramètres TreeTagger constitués.

L'algorithme ID3 est correct car, quelle que soit l'instance du problème, une solution existe et sera trouvée : c'est le « pire des cas », une feuille par objet de l'ensemble d'entraînement. Approximons à présent la complexité calculatoire de l'algorithme. A chaque nœud qui n'est pas une feuille un attribut est testé. Pour déterminer l'attribut qui sert de test il faut calculer le gain pour tous les attributs, soit $|A|$ calculs. Le gain (voir les équations 5.23 et 5.24) dépend également des objets qui sont partitionnés au moment du test de l'attribut. Donc, à chaque nœud, la complexité du calcul est directement proportionnelle à $|C|$ et $|A|$. En fait, à chaque nœud, la complexité de calcul est $O(|C| \cdot |A|)$. Ici, l'itération ne change pas la complexité et donc cet algorithme est indiqué pour de grands ensembles de données.

Chapitre 6

Corpus

6.1 Corpus et lexique

Le corpus qui servira pour entraîner les modèles statistiques doit être constitué de phrases étiquetées. Il est nécessaire qu'il s'agisse de phrases et non de mots car l'approche statistique de désambiguïsation se base sur la fréquence des suites de mots, comme cela a été expliqué plus longuement précédemment.

Il n'est pas impératif que le corpus soit lemmatisé afin de servir comme corpus d'entrainement pour un modèle d'étiquetage. Cependant, le logiciel TreeTagger permet d'étiqueter un texte mais également de proposer un lemme pour chaque forme du texte. Afin de profiter de cette fonctionnalité, le corpus qui sert à entraîner un modèle doit être constitué de textes dont les formes sont lemmatisées. Étant donné que LGeRM est conçu pour proposer des lemmes, nous n'avons pas exploité cette possibilité qui, de plus, aurait nécessité énormément de travail.

En effet, dire qu'il faut disposer d'un corpus est une chose, l'obtenir en est une autre. Par exemple, il a fallu consacrer presque la totalité du temps que les autorités de la section « traitement automatique du langage » de l'UCL impose aux étudiants pour constituer un corpus sur base de textes lemmatisés et mis à disposition par l'équipe de l'ATILF. Ce n'est que plus tard, malheureusement, que nous avons trouvé le corpus de la BFM [1]. Présenter

1. La Base de français médiéval est une base de données contenant des textes étiquetés. Elle compte plus de 4 millions de mots. Voir à l'adresse `http://bfm.ens-lyon.fr/`

ici par ordre chronologiques les étapes qui ont été réalisées pour constituer les paramètres statistiques d'étiquetage morphosyntaxique n'a que peu de sens. La solution choisie est de présenter parallèlement les étapes effectuées sur les deux sources de base : l'ensemble de textes de la BFM et celui de l'ATILF.

Le lexique qui servira de base à l'entrainement que nous allons faire est le lexique qui rassemble toutes les attestations de toutes les formes rencontrées à ce jour et encodées dans le dmf. Cela totalise pas moins de 901.335 formes.

6.1.1 Textes

Les textes de la BFM

L'équipe de la BFM nous a fourni l'ensemble de son corpus étiqueté [2]. Seuls quelques textes sont écrits en moyen français, à savoir :

— Les Manieres de langage (1396, 1399 et 1415) ,
— Roman de Jehan de Paris (1494),
— Mémoires de Philippe de Commynes (t. 1, entre 1490 et 1505),

inclus dans le modèle linguistique TreeTagger « fro.par » entrainé par cette équipe et mis à disposition du public [3].

Les textes de l'ATILF

L'équipe de l'ATILF nous a fourni des textes étiquetés et lemmatisés à partir de l'outil glossaire de LGeRM. Malheureusement – et c'est une des causes du temps qu'il a fallu pour extraire de ces textes un corpus exploitable – cet ensemble de texte n'est pas homogène.

Tout d'abord, les linguistes qui ont lemmatisés ces textes n'ont, bien souvent, pas lemmatisés les mots évidents, très nombreux, qui ne représentent aucune difficulté pour un lecteur humain, comme les « a », qui peuvent être préposition ou verbe. Il faut en effet rappeler que certains des textes ont été lemmatisés pour que les étudiants qui préparent le concours de l'agrégation en France puissent plus facilement les lire et les comprendre. C'est un des services que l'ATILF propose depuis maintenant quelques années.

2. Sous licence CC BY-NC-SA 3.0.
3. Voir `http://bfm.ens-lyon.fr/spip.php?article324`

Ensuite, certains textes ont été lemmatisés sur base d'une ancienne version du dmf. Le dictionnaire dmf est un dictionnaire en perpétuelle évolution, c'est une de ses forces mais c'est aussi une source d'ambiguïtés. Cela conduit à une situation où les textes auraient été comme lemmatisés sur base de différents dictionnaires. On peut citer certains exemples typiques de transformations du dictionnaire qui rendent très difficile la tâche d'exploitation des textes. Quelques entrées ont été regroupées. C'est la situation la plus favorable, il suffit alors de faire un programme qui convertit en nouveau lemme l'ensemble des lemmes qui se regroupent les uns avec les autres. Mais ce n'est pas la seule situation et, ainsi, d'autres entrées ont été scindées. Ce qui relevait d'un seul lemme dans l'ancien dictionnaire se retrouve alors sous différentes entrées. Pour certaines formes, l'étiquette associée a été modifiée, ce qui est d'ailleurs souvent le cas lorsqu'un lemme du dmf « ancienne version » recouvre plusieurs lemmes du dmf « nouvelle version ». Dans ces moment-là, il n'est pas possible de procéder automatiquement, puisqu'à une forme lemmatisée peut correspondre plusieurs nouveaux lemmes, si la tâche est d'attribuer automatiquement les lemmes « nouvelle version » à partir des lemmes « ancienne version ».

Mais les lemmes ne sont pas les seuls à poser problème. Les étiquettes associées aux entrées du dmf sont ambiguës. Elles sont conçues pour rendre compte de toutes les différentes formes que peuvent prendre les lemmes auxquels elles sont associées. Ainsi, en consultant au dmf le mot « souverain » [4], on constate que l'étiquette qui lui est associée est « adj. et subst. masc. ». Dans ce genre de cas, il est impossible de désambiguïser automatiquement afin de constituer un corpus étiqueté de façon univoque.

Enfin, il y a encore d'autres difficultés : le premier mot de chaque vers débute par une majuscule et les points ne sont pas employés comme en français contemporains. Les algorithmes pour scinder les phrases ne sont donc pas exactement similaires. Par exemple les points encadrent les chiffres romains et on trouve : .ii. pour signifier 2. Le détail des autres difficultés est indiqué plus loin.

Les textes qui ont servi de base à la constitution d'un corpus sont :

— le *Roman de la Fleur de Lys* de Guillaume de Digulleville,
— *Le Livre du duc des vrais amants* de Christine de Pizan,
— les *Poésies* de Charles d'Orléans,
— le *Miracle de l'abbesse grosse* (anonyme),

4. `http://www.atilf.fr/dmf/definition/souverain`

— le *Miracle de Notre-Dame* de Jean Mielot,
— les *Mémoires* de Philippe de Commynes.

6.1.2 Choix des étiquettes

Les textes qui devraient constituer un corpus homogène n'ont pas été étiquetés de la même façon. Plus précisément, soit parce qu'il y a eu différents lexiques de références, soit parce que la politique d'étiquetage n'était pas partagée par toutes les équipes, certaines formes ont été étiquetées de plusieurs façons alors que, au sein de chaque lexique, elles ne sont pas ambiguës. Il a fallu harmoniser ces étiquettes avant de penser à constituer un corpus qui puisse nous servir.

La discussion d'un choix d'étiquettes pourrait se révéler très intéressante. On sait que plus le jeu compte d'étiquettes, moins bonne est la précision dans l'étiquetage, pour le même nombre de formes d'entrainement. Intuitivement cela se conçoit d'ailleurs fort bien. S'il faut choisir entre une ou deux étiquettes, il y a moins de possibilités d'erreur que lorsque il faut choisir entre 30. Néanmoins, plus le corpus est grand, plus augmenter le nombre d'étiquettes peut se révéler utile. En effet, quelle information peut apporter un jeu de deux étiquettes ?

La question mérite en réalité d'être posée de façon beaucoup plus générale. Qu'est-ce qu'un jeu d'étiquettes ? Lorsque l'on fait de l'étiquetage morphosyntaxique, les étiquettes sont censées rendre compte des classes de mots qui se distinguent par leur nature ou par leur fonction dans la phrase. Mais, et c'est la grande leçon de Marc Wilmet [5], les classes qui existent ne rendent pas compte de grand chose, avec l'abondance de règles et d'exceptions qu'il faut pour les décrire. Comme il le dit : « à l'adage – scandaleux, quand on y réfléchit – "l'exception confirme la règle", j'ai substitué le *credo* intangible que "l'exception *infirme* la règle" » [6]. Les linguistes qui font de l'étiquetage morphosyntaxique ont tout intérêt à utiliser un jeu d'étiquettes qui permettent une identification automatique. C'est en effet le fondement du problème. Il serait tout à fait possible d'atteindre un score d'attribution automatique de 100% si des classes équivoques existaient.

Définir de telles étiquettes n'est pas l'objet de ce travail et, qui plus est, un jeu nous est imposé. En effet, le travail est de constituer un jeu d'étiquette qui servira à étiqueter les

5. Nous renverrons plusieurs fois à son excellent ouvrage [Wilmet, 1997].
6. [Wilmet, 1997], p. 8

textes médiévaux utilisée à l'ATILF. Ce jeu doit être compatible avec le jeu défini pour la nouvelle version de Frantext qui permettra de faire des recherches de formes lemmatisées et étiquetées sur le corpus total, à savoir sur le corpus des textes entre 1100 et maintenant, disponible à l'ATILF. Le nouveau jeu d'étiquettes a été proposé par Sandrine Ollinger, et nous a été fourni le 26 avril dernier. Il a donc fallu refaire correspondre les étiquettes utilisées dans les textes fournis, dans le lexique et dans ce nouveau jeu à partir de cette date. Il a fallu ensuite entraîner à nouveau tous les modèles et procéder encore une fois à tous les comptages, à toutes les analyses.

Le jeu d'étiquettes prévu pour Frantext2 est pensé de façon intéressante. Il permet une précision honnête linguistiquement dans l'étiquetage et permet conjointement d'être relativement équivoque. Cela a déjà été dit : au plus le nombre de classes est petit et au plus les classes sont bien pensées, au plus il est possible d'éviter les ambiguïtés dans la tâche d'étiquetage. Le tableau 6.1 contient les étiquettes que nous avons retenues parmi le jeu proposé pour Frantext2. Nous avons supprimé les étiquettes CLS (clitique sujet), CLO (clitique objet) et VINF (verbe à l'infinitif). Il n'était pas possible de les conserver car aucune équivalence ne pouvait être effectuée, à partir des étiquettes des textes de base, vers ces étiquettes.

TABLE 6.1 – Extrait du jeu d'étiquettes Frantext utilisé pour notre corpus

ADJ	adjectif	P+D	préposition + déterminant
ADV	adverbe	PONCT	ponctuation
CC	conjonction de coordination	PRO	pronom
CS	conjonction de subordination	PROREL	pronom relatif
DET	déterminant	PROWH	pronom interrogatif
ET	mot étranger	P	préposition
I	interjection	V	verbe
NC	nom commun	VPP	verbe participe passé
NP	nom propre	VPR	verbe participe présent
X	mot exclu		

Les textes de la BFM

Les textes de la BFM sont étiquetés avec les étiquettes du jeu d'étiquettes Cattex09min [7].

7. Voir `http://bfm.ens-lyon.fr/IMG/xml/cattex2009.xml`, expliqué dans [Guillot *et al.*, 2013b] et [Guillot *et al.*, 2013a]

Nous présentons en annexe dans le tableau 8.9 les étiquettes qui sont relevées dans notre corpus d'entraînement. Un certain nombres d'étiquettes potentielles [8] n'apparaissent pas et c'est pourquoi nous ne les reprenons pas. Ces étiquettes de la BFM ont été converties, pour notre corpus, en étiquettes Frantext2 correspondantes.

Les textes de l'ATILF

Les étiquettes proposées par LGeRM sont trop nombreuses (183) pour être présentées ici ou en annexe dans un tableau. Il faut noter que ces étiquettes n'ont pas été pensées pour faire de l'étiquetage morphosyntaxique, elles confèrent simplement une information grammaticale à un lemme, lemme qui correspond à une entrée du dmf. Les tables de conversions entre le jeu Frantext2 et le jeu LGeRM est établi sur base du principe de rester le plus exact possible, tout en gardant une grande latitude de traitement automatique. C'est pourquoi l'étiquette « X » a été attribuée aux étiquettes LGeRM qui étaient ambiguës, comme l'étiquette « adj. et subst. ». L'étiquette « X » catégorie correspond à la catégorie d'exclusion. C'est ainsi qu'une grande partie des formes des textes de l'ATILF s'est retrouvée affublée de l'étiquette d'exclusion.

6.1.3 Premiers traitements

Les textes de la BFM

Les textes de la Base de français médiéval se sont révélés une ressource fondamentale. Ils étaient étiquetés et les étiquettes permettaient une conversion automatique vers le jeu Frantext2. Le hasard des corpus a fait que deux d'entre ces textes ont été lemmatisés à l'ATILF. Ceci dit, il n'était pas envisageable dans le cadre de ce travail de lemmatiser les autres textes de cette base. De plus l'intérêt de cette démarche est quasi nul, puisque LGeRM est conçu pour proposer des lemmes parmi lesquels se trouve le lemme correct.

8. Nous renvoyons au manuel d'utilisation Cattex [?] et [?]

Les textes de l'ATILF

Les textes, lorsqu'ils passent à la moulinette de LGeRM, sont commentés. C'est à dire que nous obtenons pour chaque forme une série d'informations qu'il convient d'exploiter judicieusement. Nous donnons ici deux lignes illustratives d'un fichier de sortie de LGeRM.

```
temps;2;[____1:00003:0002][][00001][9][3697];TEMPS@subst.;;;;
passé;2;[____1:00003:0003][][00001][10][2715];PASSÉ@adj.;X;;;
```

Nous avons tout d'abord la forme, un peu plus loin, dans les premiers crochets l'indice du vers et la position du mot dans le vers (à titre d'exemple ces formes sont respectivement les deuxième et troisième formes du troisième vers), puis un peu plus loin le lemme et l'étiquette, autour d'un « @ ». Nous trouvons également d'autres informations qu'il n'est pas nécessaire d'exploiter pour l'attribution automatique des étiquettes. On le voit, il y a une série de traitements à opérer avant de passer à un corpus propre.

Il a d'abord fallu supprimer les µ devant les noms propres (ex. : µPERROT@nom propre) et les symboles € devant les mots étrangers (ex. : €ET@mot étranger), cela afin de pouvoir récupérer les lemmes. Ensuite nous avons remplacé les majuscules en début de vers par des minuscules.

Parfois le fichier initial contenait des ambiguïtés au niveau de la lemmatisation, par exemple on trouvait : OST@subst./OSER@verbe/OS1@subst à la place de l'information du lemme et de l'étiquette. Dans ce genre de cas, les lemmes ainsi que les étiquettes ont été remplacés par « _MOT_EXCLU » et « ERROR ». A titre d'illustration, il existait 22 lignes contenant ce type d'ambiguïté dans le *Miracle de l'abbesse grosse*, 91 dans les *Poésies* de Charles d'Orléans, 1605 dans les *Mémoires* de Philippe de Commynes, 98 dans Evrart de Conty, 4017 dans Jehan de Paris, 758 dans le *Miracle de Notre-Dame*, 13 dans *Le Livre du duc des vrais amants* et 6 dans le *Roman de la Fleur de Lys*.

Enfin, il restait comme prétraitement à regrouper en une ligne les lignes qui contenaient une indication de nombres encadrés par des points, lemmatisés auparavant en trois formes distinctes, et à supprimer les lignes contenant les indications de folio.

Tout cela nous a permis d'obtenir un fichier texte où seuls apparaissent les formes, les étiquettes et les lemmes.

6.2 Exploration du corpus

Le corpus provient d'un ensemble de textes étiquetés de façon hétérogène. On l'a souligné suffisamment. Il résulte de cela que le taux d'ambiguïté – c'est-à-dire de formes portant l'étiquette d'exclusion (X) – est important. Sur 277.370 des formes que totalisent les textes de l'ATILF, 59.669 sont étiquetées par X, ce qui donne un taux de 21,51%.

Une solution pour constituer, à partir de cet ensemble, un corpus qui puisse servir d'entrainement à un modèle statistique d'étiquetage est de sélectionner uniquement les phrases ne contenant pas l'étiquette « X ». Cette collecte ne rapporte que 2051 formes. C'est évidemment beaucoup trop peu. Pendant le stage, un temps a été pris pour sélectionner les phrases ne contenant qu'une seule étiquette « X » afin de lever l'ambiguïté manuellement et ainsi de pouvoir augmenter sensiblement le nombres de phrases, le nombre de formes du corpus. Cela a fonctionné assez bien puisque le total des formes, après cette étape, s'élevait à près de 8000. Cette étape n'est pas envisageable actuellement, car les nouvelles étiquettes n'ont pas été reçues assez longtemps à l'avance. La présence du corpus massif de la BFM joue aussi à amoindrir l'utilité de ce travail.

Il n'est pas nécessaire d'explorer plus avant ce corpus de l'ATILF. Celui qui est utilisé à partir de maintenant est celui de la BFM car il est 10 fois plus conséquent : il compte 83.067 formes, dont seules 28 sont étiquetées par « X ». Le tableau 6.2 contient une exploration plus fine de ce corpus. La première colonne indique un nombre d'étiquettes. La deuxième colonne contient le nombre de formes différentes qui sont étiquetées du nombre de façon qui est indiqué dans la première colonne. Il y a donc 8082 formes différentes qui sont étiquetées d'une seule étiquette. La troisième colonne contient le nombre total d'occurrences comptées. La quatrième colonne représente la fréquence relative que ces formes représentent ensemble et la dernière colonne contient la fréquence cumulée correspondante. On voit que le taux de

TABLE 6.2 – Données relatives au corpus extrait de la BFM

nbr d'étiquettes	formes	occurrences	fréquence	fréq. cumulée
1	8082	57382	69,1%	69,1%
2	416	14567	17,5%	86,6%
3	80	8534	10,3%	96,9%
4	6	1111	1,3%	98,2%
7	1	1476	1,8%	100,0%

formes qui ne sont étiquetées que par une seule étiquette est de 69,1% dans ce corpus. A titre de comparaison, ce taux est de près de 88,9% pour l'anglais, lorsque le taux est mesuré sur le Brown Corpus.

La forme qui représente à elle seule de 1,8% du total des formes et qui est responsable de 5,8% de l'ambiguïté totale $(1,8/(100-69,1)$, soit $1,8/30,9)$ est « que ». Cette forme est étiquetée comme un déterminant, une conjonction de coordination, un mot exclu, un pronom relatif, un pronom interrogatif, un adverbe et une conjonction de subordination. Dans l'idéal il faudrait commencer par ce genre de formes si l'on devait formuler des règles afin de constituer un analyseur statistique. Mais avant d'envisager l'existence-même de ce genre de règles, il faudrait que la communauté des linguistes s'accorde sur un ensemble d'étiquettes.

6.3 Résultats

A partir du tableau 6.2, on comprend que tout étiqueteur doit donner au minimum un score de 69.1% de formes correctes. Pour ces formes qui ne se rencontrent que sous une seule étiquette, en effet, il n'y a aucun doute.

6.3.1 MLE

Pour le sport, nous avons implémenté de façon très simple, avec un tableur, un étiqueteur statistique utilisant le maximum de vraisemblance. Pour cela il s'agissait uniquement de compter les séquences d'étiquettes (disons de 2 ou 3 étiquettes, afin d'obtenir un certaine précision) avant les formes. Par exemple, pour la forme « a », il a fallu dénombrer toutes les séquences de 3 étiquettes la précédent et dénombrer dans chaque cas le nombre de fois que telle ou telle étiquette était attribuée à la forme. Pour cette forme, d'ailleurs, les résultats sont corrects. Seules 99 étiquettes sont incorrectes sur 1167, soit 8,5%. Pour d'autres formes, les résultats sont moins probants. Au total, les paramètres TreeTagger donnent de meilleurs résultats. Ce sont ceux-là que nous commenterons plus en détail.

6.3.2 TreeTagger

Nous disposons de paramètres TreeTagger entrainés par Achim Stein pour l'ancien français et par l'équipe de la BFM pour le français « ancien », de 1100 à 1500. Cependant, le moyen français est une langue qui se différencie très distinctement, au niveau de la fréquence d'emploi des mots « grammaticaux », des autres états de langue – la démonstration de cette affirmation fait l'objet du chapitre 3. Dès lors nous avons trouvé nécessaire de procéder à un entraînement de TreeTagger spécifiquement pour le moyen français. Naturellement nous discuterons tout de même les résultats fournis par les paramètres existants. Le premier entrainement de TreeTagger donne un arbre de profondeur 20 et d'un nombre de nœuds égal à 101. Le tableau 6.3 rapporte les résultats obtenus avec ce jeu de paramètres. On constate

TABLE 6.3 – Evaluation de TreeTagger

Eti.	Et. tot.	Et. corr.	Fréquence
ADJ	272	79	29%
ADV	630	332	53%
CC	407	348	86%
CS	87	65	75%
DET	606	281	46%
NC	1504	1295	86%
NP	183	24	13%
P	613	506	83%
P+D	40	1	3%
PONCT	1642	1642	100%
PRO	843	744	88%
PROREL	52	0	0%
PROWH	34	8	24%
V	1136	1003	88%
Total général	8049	6328	79%

combien les résultats semblent mitigés. Le tableau 6.2 montrait que dans le corpus d'entrainement, le nombre de mots ne contenant pas d'ambiguïté au niveau du tag était de 69,1%. Donc n'importe quel étiqueteur doit au moins réussir à étiqueter ces mots-là correctement. Notre jeu de paramètres permet de dépasser ce score mais pas de beaucoup.

Une exploration plus fine des résultats est nécessaire. On s'aperçoit que, dans le corpus de test, les formes étiquetées « VPR » et qui donc doivent être des verbes au participe présent, qui ne sont jamais reconnues par nos paramètres, ne sont pas des verbes au participe présent.

Le fait que le corpus soit à revoir n'est jamais une simple possibilité.

Le faible score des pronoms relatifs (« PROREL »)et des pronoms interrogatifs (« PROWH ») est dû au seul « que ». On ne conservera donc pas la distinction « PRO » et « PROREL », « PROWH ». Les formes étiquetées « P+D » (forme contractée d'une préposition et d'un déterminant) apparaissent rarement. L'entrainement statistique n'est pas optimal et c'est probablement ce qui cause des problèmes et, en réalité, les étiquettes attribuées sont très censées. Dans la plupart des cas, ces formes reçoivent l'étiquettes « DET » (déterminant) ou « P » (préposition). Quelques autres formes (« du », « as ») se voient parfois attribuer l'étiquette « V » (verbe) ou même « NC ». On conservera par contre l'étiquette « P+D » car elle correspond à l'étiquette « P+D » du corpus de la BFM et que son faible taux d'apparition ne joue pas tant dans le score total. On trouve 40 formes étiquetées « P+D » dans le corpus d'évaluation qui compte 8049 formes, ce qui représente moins de 0,5% du total des formes.

Les adjectifs non reconnus proviennent en majorité du lemme « bon ». Ce lemme pouvant être employé comme adverbe et comme nom commun, l'étiquette du dmf est « adj. et subst. ». Certes cette étiquette ne mentionne pas l'emploi adverbial qui est renseigné dans l'article. Cependant cela nous a fait prendre conscience que le programme que nous avions effectué pour convertir les étiquettes induisait des erreurs. Il ne fallait pas remplacer ce type d'étiquettes du dmf par l'étiquette « X » au moment de la construction du lexique qui sert de référence pour l'apprentissage, car alors le logiciel TreeTagger ne va évidemment pas attribuer une autre étiquette que « X » au moment de sa tâche d'étiquetage.

Le cas des déterminants est probablement un peu différent. Après avoir isolé les déterminants non reconnus, on s'aperçoit que systématiquement les nombres sont étiquetés comme « ADJ », soit adjectifs. Marc Wilmet a fait du déterminant un cheval de bataille. Il nous instruit de l'histoire des classes « adjectif » et « déterminant ». Il fait même plus, il démontre comment ces deux classes ne sont pas naturelles et qu'en réalité les mots qui les constituent peuvent occuper « trois *fonctions* de quantification, de caractérisation et de quantification-caractérisation » [9]. Les deux classes ne forment qu'une seule, dont le nom est celui d'« adjectif », « à laquelle s'agrège, par transfert, des noms, des verbes, des pronoms, des adverbes » [10].

Il ne faut pas, dans ce travail, commencer une épistémologie de la grammaire. Il faut être

9. [Wilmet, 1997], p. 245
10. [Wilmet, 1997], p. 245

plus humble et plus ambitieux à la fois. Plus humble d'abord : il n'est pas possible de faire en quelque ligne une étude critique des grammaires alors que des chercheurs y consacrent leur vie depuis des années. Plus ambitieux ensuite : il faut des classes que l'on puisse implémenter maintenant et il faut que ces classes soient compatibles avec le jeu d'étiquettes Frantext2, dont nous avons montré que le bien-fondé pouvait être remis en question, et avec le jeu utilisé par l'équipe de Lyon.

Nouveaux paramètres

La première solution pour améliorer les performances des paramètres est de corriger le vice de construction du lexique d'apprentissage. Les étiquettes du type « adj. et subst. » ne seront plus remplacée par « X » mais une fois par « ADJ » et « NC », comme s'il s'agissait de deux lemmes différents. Il n'a pas été possible d'éliminer l'étiquette d'exclusion à chaque fois. Par exemple pour l'étiquette « adv. d'intensité et conj. », les étiquettes de remplacement sont « ADV X », car on ne sait pas de quelle conjonction il s'agit. Des nouveaux paramètres ont été entrainés, à partir d'ici ils seront désignés comme « paraMem2 ». Le nombre de nœuds de ces paramètres est identique à celui des premiers constitués, l'arbre a également la même profondeur.

Pour envisager la deuxième solution qui améliorerait les paramètres, un certain pragmatisme est de mise : élaguons les classe par groupements. Nous avons déjà évoqué la proposition de rassembler les classes « PRO ». D'autres groupements sont envisageables. Cependant, réduire autant les classes que Wilmet [11] ne permettrait pas de conserver des paramètres qui puissent *in fine* servir à la désambiguïsation des formes, puisque, étant donné le faible nombre de classes que Wilmet conserve (il propose 3 classes et accepte de monter à 4 si le lecteur insiste), toutes les formes seraient rassemblées dans les mêmes classes. Le tableau 6.4 expose les groupements de classes effectués. Des nouveaux paramètres ont été entrainés, à partir d'ici ils seront désignés comme « paraMem3 ». Le nombre de noeuds (51) de cet arbre est plus restreint que celui des précédents, de même que la profondeur de l'arbre (13).

Nous disposons donc de 4 paramètres TreeTagger : celui d'Achim Stein pour l'ancien français, celui de l'équipe de la BFM à Lyon, celui que nous avons constitué avec les classes d'étiquettes Frantext 2 (paraMem2) et celui que nous avons constitué avec un nombre de

11. [Wilmet, 1997]

TABLE 6.4 – Conversion de l'ancien jeu d'étiquettes au nouveau jeu d'étiquettes

ADJ	ADJ	P+D	P+D
ADV	ADV	PONCT	PONCT
CC	C	PRO	PRO
CS	C	PROREL	PRO
DET	ADJ	PROWH	PRO
ET	ET	P	P
I	I	V	V
NC	N	VPP	V
NP	N	VPR	V
X	X		

classes plus faible (paraMem3). Il n'est pas possible de comparer les résultats donnés par ces paramètres sans discussion. En effet, les étiquettes qu'ils fournissent ne sont pas identiques et pas toujours convertibles sans ambiguïté. On rappelle que le jeu d'étiquettes des paramètres de la BFM est le jeu Cattex09min (voir tableau 8.9).

Nous pouvons comparer les paramètres de l'équipe de Lyon et les nôtres (paraMem2) mais sur un corpus d'évaluation qui n'est pas tiré du corpus de la BFM, auquel cas les paramètres de la BFM seraient évalués sur une partie de leur corpus d'entrainement. C'est ici que l'on va utiliser le corpus que nous avons constitué sur base des textes de l'ATILF. Pour la certitude nous avons tout de même évalué les paramètres avec le corpus tiré de la BFM également.

Nous ne pouvons pas comparer le jeu d'Achim Stein avec les autres car ce jeu ne compte qu'une étiquette rassemblant les conjonctions de coordination et de subordination. Ces paramètres ne seront donc comparés qu'aux paramètres que nous avons constitué sur base du jeu d'étiquettes ne comptant qu'un nombre réduit d'étiquettes (paraMem3).

Pour résumer, nous avons donc 4 paramètres TreeTagger et deux corpus de test. Tous les paramètres sont testés sur les deux corpus. Les étiquettes que les paramètres de Stein et de la BFM utilisent sont converties avant d'être comparées à celles de nos paramètres et aux étiquettes de correction.

Nouveaux résultats

Avant de continuer la présentation des performances des paramètres TreeTagger, nous devons expliquer la façon dont on mesure de telles performances. En TAL, l'usage a consacré

la *précision*, le *rappel* et la *F-mesure*. Nous définissons ces notions en tant qu'elles sont utilisées pour mesurer les performances d'une tâche d'étiquetage.

La précision mesure le pourcentage d'étiquettes attribuées qui le sont correctement :

$$\text{Précision} := \frac{\text{Nombre d'étiquettes correctement attribuées}}{\text{Nombre total d'étiquettes attribuées}}. \tag{6.1}$$

Le rappel mesure le pourcentage d'étiquettes du corpus qui ont été correctement attribuées :

$$\text{Rappel} := \frac{\text{Nombre d'étiquettes correctement attribuées}}{\text{Nombre total d'étiquettes dans le corpus}}. \tag{6.2}$$

La F-mesure est une façon de combiner les deux mesures précédentes en une seule :

$$F := \frac{2PR}{P+R}. \tag{6.3}$$

Il n'est pas utile de s'attarder sur la façon de définir la F-mesure, mais il est important de comprendre pourquoi on combine les deux premières mesures. Étant donné une classe contenant plusieurs formes de la même étiquette. Si les paramètres n'attribuent cette étiquette qu'à une seule forme et que cette forme fait partie de la classe, la précision sera maximale (1). Le rappel, par contre, sera faible car il y aura des formes non reconnues. Si par contre les paramètres attribuent cette étiquette à toutes les formes du corpus. Alors, toutes les formes portant cette étiquette seront correctement attribuées et le rappel sera maximal (1). Par contre, ces paramètres seront d'une piètre précision puisque toutes les autres formes ne seront pas correctement étiquetées.

Les tableaux 8.1, 8.3, 8.2 et 8.4 (que l'on trouvera en annexe, comme tous les autres mentionnés dans cette section) contiennent les résultats des tests effectués avec tous les paramètres sur le corpus provenant des textes de l'ATILF.

Dans un premier temps, les paramètres de la BFM ne donnent pas des résultats inférieurs à ceux que l'on a constitué sur base d'un jeu d'étiquettes « maximal » (paraMem2) (voir les tableaux 8.1 et 8.3). Les deux ensembles de paramètres sont médiocres pour ce qui concerne les adjectifs (F-mesure de 0,5 et 0,51 pour respectivement les paramètres paraMem2 et BFM). Ils donnent d'assez bons résultats pour les conjonctions de coordination (F- mesure de respectivement 0,93 et 0,97) et nos paramètres sont bons également pour les verbes (F-mesure

de 0,92).

Le tableau 8.2 contient les résultats des paramètres destinés à l'ancien français constitués par Achim Stein. Malheureusement ces paramètres ne produisent pas en sortie des étiquettes équivalentes à celles présentées ici, elle sont plus nombreuses. Ainsi, la conversion qui permet la comparaison nous trompe sur la précision des paramètres de Stein et en fin de compte, ce qui est évalué résulte d'un regroupement de classes. Ainsi, la F-mesure totale de 0,87 – relativement haute – ne peut être imputée à la seule qualité des paramètres.

Les tableaux 8.5, 8.7, 8.8 et 8.6 contiennent les résultats de l'évaluation des paramètres testés sur le corpus de textes provenant de la BFM.

Sans surprise, les résultats des paramètres de la BFM sont assez élevés pour chaque classe et la F-mesure globale pour ces paramètres est de 0,89. Cette mesure n'est cependant pas beaucoup plus élevée que celles de nos paramètres qui est de 0,86. Donc, si l'on relativise le score des paramètres de la BFM par le fait que ces paramètres ont été entrainés en partie sur le corpus sur lequel ils sont ici testés, on ne peut pas être admiratif de ces résultats.

On constate dans le tableau 8.6 que nos paramètres constitués sur un jeu minimal d'étiquettes (paraMem3) donnent de bons résultats. Ils donnent la plus haute F-mesure atteinte : 0,9. Pour la même raison que précédemment, on ne peut pas affirmer que la haute F-mesure des paramètres de Stein (0,85) soit due à la qualité de ces paramètres.

Chapitre 7

Conclusion

Effectuons à présent un bref retour sur ce qui vient d'être présenté.

Les pratiques sociales varient, les langues changent. Elles ont donc une histoire. Les historiens de la langue ont l'habitude, comme leurs collègues, de procéder à une périodisation de leur objet d'étude.

Dans ce travail, des éléments nouveaux qui corroborent la périodisation traditionnelle de la langue ont été apportés. Il est maintenant rigoureux d'affirmer que le moyen français est un état de langue qui se distingue de l'ancien français et du français classique. De plus, la distinction entre le français contemporain et le français classique est moins forte que celles qui identifient le moyen français.

La démonstration a été faite à partir de l'analyse de matrice contenant la fréquence relative d'apparition de certains mots pendant des intervalles temporel d'une durée d'un demi-siècle, à partir de 1100 jusqu'à nos jours. Nous avons utilisé pour cela des méthodes classiques d'analyse de données : l'analyse par composantes principales et la classification ascendante hiérarchique.

Nous avons ensuite présenté un problème pratique. Le lemmatiseur LGeRM produit du bruit lorsqu'il propose des lemmes. Nous savions que s'il était possible d'attribuer une étiquette morphosyntaxique aux formes du texte, ce bruit pouvait diminuer car tous les lemmes ne pouvaient pas correspondre. Depuis longtemps l'utilisation de chaines de Markov dans la tâche d'attribution d'étiquettes morphosyntaxiques aux formes d'un texte est bien étudiée.

Des paramètres statistiques ont donc été entrainés en vue de la tâche d'étiquetage, alors qu'il en existait déjà. Ces paramètres n'étaient toutefois pas prévus exclusivement pour le moyen français. Or nous avions démontré que le moyen français se distingue des autres états de la langue par la fréquence d'apparition des mots. Des paramètres statistiques conçus à partir de la fréquence relative des mots ne peuvent donc que donner des meilleurs résultats pour le moyen français s'ils sont (bien) conçus pour le moyen français.

Ce résultat théorique peut servir de point de départ à l'explication de ce que des paramètres entrainés exclusivement sur un corpus en moyen français donnent de meilleurs résultats que d'autres paramètres, entrainés sur un corpus mixte ou prévus pour l'ancien français. Au terme de ce travail, le problème initial qui consistait à réduire le bruit produit par LGeRM est en voie d'être résolu. Parmi les paramètres qui existent et qui permettent de réduire le nombre de lemmes proposés par LGeRM, ceux que nous avons constitué donnent les meilleurs résultats. Nous proposons deux jeux de paramètres (paraMem3 et paraMem2), basés respectivement sur un ensemble d'étiquettes restreint et plus fourni. L'expérience dira lesquels sont les plus utiles.

Chapitre 8

Annexes

8.1 Résultats expérimentaux

Tous les tableaux de cette section ont une architecture comparable. La première colonne indique les étiquettes, la deuxième colonne le nombre de fois que l'étiquette est attribuée dans le corpus (tot. corp.). La troisième colonne contient le nombre de fois que les paramètres testés attribuent cette étiquette (<nom param.> attr.) et la quatrième colonne contient le nombre de fois que cette étiquette est attribuée correctement (< nom param.> corr.). Les trois dernières colonnes contiennent respectivement la précision (P), le rappel (R) et la F-mesure (F). La première colonne porte le nom « eti. max. » ou « eti. min. » selon qu'il s'agissait du jeu d'étiquettes complet ou du jeu réduit après le regroupement des classes.

8.1.1 Corpus de L'ATILF

TABLE 8.1 – Résultats des paramètres paraMem2 sur le corpus de l'ATILF

eti. max.	tot. corp.	paraMem2 attr.	paraMem2 corr.	P	R	F
ADJ	165	101	66	0,65	0,40	0,50
ADV	214	158	145	0,92	0,68	0,78
CC	84	92	82	0,89	0,98	0,93
CS	0	4	0	0,00	–	–
DET	34	134	21	0,16	0,62	0,25
I	23	14	14	1,00	0,61	0,76
NC	347	378	312	0,83	0,90	0,86
NP	5	3	0	0,00	0,00	–
P	146	158	131	0,83	0,90	0,86
P+D	5	3	2	0,67	0,40	0,50
PONCT	394	363	363	1,00	0,92	0,96
PRO	212	198	181	0,91	0,85	0,88
PROREL	31	33	28	0,85	0,90	0,88
PROWH	2	1	1	1,00	0,50	0,67
V	389	402	364	0,91	0,94	0,92
VPR	0	1	0	0,00	–	–
X	0	8	0	0,00	–	–
Total	2051	2051	1710	0,83	0,83	0,83

TABLE 8.2 – Résultats des paramètres paraMem3 sur le corpus de l'ATILF

eti. min.	tot. corp.	paraMem3 attr.	paraMem3 corr.	P	R	F
ADJ	199	227	174	0,77	0,87	0,82
ADV	214	150	139	0,93	0,65	0,76
C	84	94	82	0,87	0,98	0,92
I	23	16	16	1,00	0,70	0,82
N	352	375	321	0,86	0,91	0,88
P	146	158	128	0,81	0,88	0,84
P+D	5	2	2	1,00	0,40	0,57
PONCT	394	363	363	1,00	0,92	0,96
PRO	245	228	210	0,92	0,86	0,89
V	389	434	377	0,87	0,97	0,92
X	0	4	0	0,00	–	–
Total	2051	2051	1812	0,88	0,88	0,88

TABLE 8.3 – Résultats des paramètres de la BFM sur le corpus de l'ATILF

eti. max.	tot. corp.	BFM attr.	BFM corr.	P	R	F
ADJ	165	73	61	0,84	0,37	0,51
ADV	214	187	168	0,90	0,79	0,84
CC	84	89	84	0,94	1,00	0,97
CS	0	5	0	0,00	–	–
DET	34	142	24	0,17	0,71	0,27
ET	0	1	0	0,00	–	–
I	23	6	6	1,00	0,26	0,41
NC	347	314	286	0,91	0,82	0,87
NP	5	40	5	0,13	1,00	0,22
P	146	159	131	0,82	0,90	0,86
P+D	5	21	5	0,24	1,00	0,38
PONCT	394	394	394	1,00	1,00	1,00
PRO	212	201	184	0,92	0,87	0,89
PROREL	31	34	27	0,79	0,87	0,83
PROWH	2	3	1	0,33	0,50	0,40
V	389	301	293	0,97	0,75	0,85
VPR	0	81	0	0,00	–	–
Total	2051	2051	1669	0,81	0,81	0,81

TABLE 8.4 – Résultats des paramètres de Stein sur le corpus de l'ATILF

eti. min.	Tot. Corpus	Stein attri.	Stein corr.	Précision	Rappel	F-mesure
ADJ	199	225	170	0,76	0,85	0,80
ADV	214	147	135	0,92	0,63	0,75
C	84	83	83	1,00	0,99	0,99
I	23	17	16	0,94	0,70	0,80
N	352	342	304	0,89	0,86	0,88
P	146	147	124	0,84	0,85	0,85
P+D	5	0	0	–	0,00	–
PONCT	394	366	366	1,00	0,93	0,96
PRO	245	267	214	0,80	0,87	0,84
V	389	457	371	0,81	0,95	0,88
Total	2051	2051	1783	0,87	0,87	0,87

8.1.2 Corpus de la BFM

TABLE 8.5 – Résultats des paramètres paraMem2 sur le corpus de la BFM

eti. max.	tot. corp.	paraMem2 attr.	paraMem2 corr.	P	R	F
ADJ	272	571	229	0,40	0,84	0,54
ADV	630	514	460	0,89	0,73	0,80
CC	407	365	351	0,96	0,86	0,91
CS	87	90	64	0,71	0,74	0,72
DET	606	420	363	0,86	0,60	0,71
NC	1504	1695	1387	0,82	0,92	0,87
NP	183	24	10	0,42	0,05	0,10
P	613	607	579	0,95	0,94	0,95
P+D	40	1	1	1,00	0,03	0,05
PONCT	1642	1642	1642	1,00	1,00	1,00
PRO	843	783	757	0,97	0,90	0,93
PROREL	52	58	31	0,53	0,60	0,56
PROWH	34	8	4	0,50	0,12	0,19
V	1136	1219	1029	0,84	0,91	0,87
X	0	52	0	0,00	–	–
Total	8049	8049	6907	0,86	0,86	0,86

TABLE 8.6 – Résultats des paramètres paraMem3 sur le corpus de la BFM

eti. min.	tot. corp.	paraMem3 attr.	paraMem3 corr.	P	R	F
ADJ	878	964	799	0,83	0,91	0,87
ADV	630	501	445	0,89	0,71	0,79
C	494	467	411	0,88	0,83	0,86
N	1687	1755	1569	0,89	0,93	0,91
P	613	595	564	0,95	0,92	0,93
P+D	40	1	1	1,00	0,03	0,05
PONCT	1642	1640	1640	1,00	1,00	1,00
PRO	929	821	793	0,97	0,85	0,91
V	1136	1259	1058	0,84	0,93	0,88
X	0	46	0	0,00	–	–
Total	8049	8049	7280	0,90	0,90	0,90

TABLE 8.7 – Résultats des paramètres de la BFM sur le corpus de la BFM

eti. max.	tot. corp.	BFM attr.	BFM corr.	P	R	F
ADJ	272	303	214	0,71	0,79	0,74
ADV	630	550	498	0,91	0,79	0,84
CC	407	385	379	0,98	0,93	0,96
CS	87	81	61	0,75	0,70	0,73
DET	606	489	468	0,96	0,77	0,85
ET	0	2	0	0,00	–	–
I	0	1	0	0,00	–	–
NC	1504	1544	1331	0,86	0,88	0,87
NP	183	217	155	0,71	0,85	0,78
P	613	611	575	0,94	0,94	0,94
P+D	40	44	31	0,70	0,78	0,74
PONCT	1642	1640	1640	1,00	1,00	1,00
PRO	843	836	805	0,96	0,95	0,96
PROREL	52	65	38	0,58	0,73	0,65
PROWH	34	11	5	0,45	0,15	0,22
V	1136	1066	931	0,87	0,82	0,85
VPR	0	203	0	0,00	–	–
X	0	1	0	0,00	–	–
Total	8049	8049	7131	0,89	0,89	0,89

TABLE 8.8 – Résultats des paramètres de Stein sur le corpus de la BFM

eti. min.	tot. corp.	Stein attr.	Stein corr.	P	R	F
ADJ	878	896	706	0,79	0,80	0,80
ADV	630	391	329	0,84	0,52	0,64
C	494	369	359	0,97	0,73	0,83
I	0	8	0	0,00	–	–
N	1687	1636	1392	0,85	0,83	0,84
P	613	553	521	0,94	0,85	0,89
P+D	40	0	0	–	0,00	–
PONCT	1642	1642	1642	1,00	1,00	1,00
PRO	929	1120	852	0,76	0,92	0,83
V	1136	1430	1063	0,74	0,94	0,83
X	0	4	0	0,00	–	–
Total	8049	8049	6864	0,85	0,85	0,85

8.2 Etiquettes

TABLE 8.9 – Jeu d'étiquettes Cattex09min

ABR	abréviation	OUT	mot exclu
ADJcar	adjectif cardinal	PONfbl	ponctuation faible
ADJind	adjectif indéfini	PONfrt	ponctuation forte
ADJord	adjectif ordinal	PONpdr	ponctuation parenthétique
ADJpos	adjectif possessif	PONpga	ponctuation parenthétique
ADJqua	adjectif qualificatif	PONpxx	ponctuation parenthétique
ADVgen	adverbe général	PRE	préposition
ADVing	adverbe interrogatif négatif	PRE.DETcom	enclise dét. après prép.
ADVint	adverbe interrogatif	PRE.DETdef	enclise dét. après prép.
ADVneg	adverbe de négation	PRE.PROrel	enclise pro. après prép.
ADVsub	adverbe « subordonnant »	PROadv	pronom adverbial
CONcoo	conjonction de coordination	PROcar	pronom cardinal
CONsub	conjonction de subordination	PROdem	pronom démonstratif
DETcar	déterminant cardinal	PROimp	pronom impersonnel
DETcom	déterminant composé	PROind	pronom indéfini
DETdef	déterminant défini	PROint	pronom interrogatif
DETdem	déterminant démonstratif	PROord	pronom ordinal
DETind	déterminant indéfini	PROper	pronom personnel
DETint	déterminant interrogatif	PROpos	pronom possessif
DETndf	déterminant non défini	PROrel	pronom relatif
DETpos	déterminant possessif	RED	mot redondant
DETrel	déterminant relatif	VERcjg	verbe conjugué
ETR	mot étranger	VERinf	verbe infinitif
INJ	interjection	VERppa	participe présent
NOMcom	nom commun	VERppe	participe passé
NOMpro	nom propre		

Liste des tableaux

Table des figures

Bibliographie

[Bellanger et Tomassone, 2014] Bellanger, L. et Tomassone, R. (2014). *Exploration de données et méthodes statistiques : data analysis & data mining avec le logiciel R*. Références sciences. Ellipses, Paris.

[Bouroche et Saporta, 1980] Bouroche, J.-M. et Saporta, G. (1980). *L'analyse des données*. Que sais-je ? Presses universitaires de France, Paris.

[Busby et Kleinhenz, 2010] Busby, K. et Kleinhenz, C. (2010). *Medieval Multilinguism. The Francophone World and its Neighbours*. Brepols, Turnhout.

[Catach, 2001] Catach, N. (2001). *Histoire de l'orthographe française*. Honoré Champion, Paris.

[Cazal et Parussa, 2015] Cazal, Y. et Parussa, G. (2015). *Introduction à l'histoire de l'orthographe*. Cursus. Armand Colin, Paris.

[Cerquiglini, 1991] Cerquiglini, B. (1991). *La naissance du français*. Que sais-je ? Presses universitaires de France, Paris.

[Cerquiglini *et al.*, 2000] Cerquiglini, B., Corbeil, J.-C., Klinkenberg, J.-M. et Peeters, B. (2000). *Le Français dans tous ses états*. Flammarion, Paris.

[Chaurand, 1999] Chaurand, J. (1999). *Nouvelle Histoire de la langue française*. Seuil, Paris.

[Cornillon *et al.*, 2010] Cornillon, P.-A., Guyader, A. et Husson, F. (2010). *Statistiques avec R*. Pratique de la statistique. Presses universitaires de Rennes, Rennes.

[de Saussure, 1994] de Saussure, F. (1994). *Cours de linguistique générale*. Payot, Paris.

[Escofier et Pagès, 2008] Escofier, B. et Pagès, J. (2008). *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*. Sciences sup. Dunod, Paris.

[Fagard, B. et Degand, L., 2008] Fagard, B. et Degand, L. (2008). La fortune des mots : grandeur et décadence de « car ».

[Guillot *et al.*, 2013a] Guillot, C., Prévost, S. et Lavrentiev, A. (2013a). Manuel de référence du jeu Cattex09.

[Guillot *et al.*, 2013b] Guillot, C., Prévost, S. et Lavrentiev, A. (2013b). Principes d'annotation Cattex09.

[Husson *et al.*, 2009] Husson, F., Lê, S. et Pagès, J. (2009). *Analyse de données avec R.* Pratique de la statistique. Presses universitaires de Rennes, Rennes.

[Jurafsky et Martin, 2009] Jurafsky, D. et Martin, J. (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition.* Pearson Prentice Hall,, Upper Saddle River, N.J. :, 2nd ed. édition.

[Klinkenberg, 1999] Klinkenberg, J.-M. (1999). *Des langues romanes.* Champs linguistiques. De Boeck, Bruxelles.

[Klinkenberg, 2000] Klinkenberg, J.-M. (2000). *Précis de sémiotique générale.* Points. Seuil, Paris.

[Quinlan, 1986] Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning*, (1):81–106.

[Raynaud de Lage, 1993] Raynaud de Lage, G. (1993). *Introduction à l'ancien français.* Moyen âge. SEDES, Paris, 2e éd. revue et corr édition. couv. ill. en coul. 24 cm. Bibliogr. p. 261-264. Index.

[Rey *et al.*, 2011] Rey, A., Duval, F. et Siouffi, G. (2011). *Mille ans de langue française, tome 1 : Des origines au français moderne.* Perrin, Paris.

[Ruppli, 1990] Ruppli, M. (1990). L'opposition Car/Parce que. *L'Information Grammaticale*, 46(1):22–25.

[Schmid, 1994] Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees.

[Schmid, 1995] Schmid, H. (1995). Improvements In Part-of-Speech Tagging With an Application To German.

[Wilmet, 1997] Wilmet, M. (1997). *Grammaire critique du français.* HU. Hachette supérieur Duculot, Paris, Louvain-la-Neuve.

[Zink, 1990] Zink, G. (1990). *Le moyen français*. Que sais-je ? Presses universitaires de France, Paris.

[Zumthor, 1960] Zumthor, P. (1960). Document et monument. a propos des plus anciens textes de langue française. *Revue des sciences humaines*, 57:5–19.

Table des matières

Place Blaise Pascal, 1 bte L3.03.11, 1348 Louvain-la-Neuve, Belgique **www.uclouvain.be/fial**