Traitements pour l'analyse du français préclassique

Sascha Diwersy[1], Achille Falaise[2], Marie-Hélène Lay[3], Gilles Souvay[4]
(1) Université de Cologne, Albertus-Magnus-Platz, D-50923 Köln, Allemagne
(2) ICAR, ENS de Lyon, 15 Parvis René Descartes, 69342 Lyon, France
(3) FoReLL, Université de Poitiers, 5 rue Théodore Lefebvre, 86073 Poitiers, France
(4) ATILF-CNRS, Université Nancy 2, 44 avenue de la Libération, 54063 Nancy, France
sascha.diwersy@uni-koeln.de, achille.falaise@ens-lyon.fr, marie-helene.lay@univ-poitiers.fr, gilles.souvay@atilf.fr

Résumé. La période « préclassique » du français s'étend sur tout le XVIe siècle et la première moitié du XVIIe siècle. Cet état de langue écrite, qui accompagne les débuts de l'imprimerie, est relativement proche du français moderne, mais se caractérise par une grande variabilité graphique. Il s'agit de l'un des moins bien dotés en termes de ressources. Nous présentons ici la construction d'un lexique, d'un corpus d'apprentissage et d'un modèle de langage pour la période préclassique, à partir de ressources du français moderne.

Abstract.

Treatments for Preclassic French parsing

The "Preclassical" French language period extends throughout the sixteenth century and the first half of the seventeenth century. This state of the written French language, which accompanies the beginnings of printing, is relatively close to the modern French, but is characterized by a large graphic variability. It is one of the most underresourced state of the French language. Here we present the construction of a lexicon, a training corpus and a language model for the Preclassic period, built from modern French resources.

Mots-clés : construction de lexique morphologique, annotation et étiquetage de corpus, linguistique diachronique.

Keywords: morphological lexicon construction, corpus annotation and tagging, diachronic linguistics.

1 Introduction

La période « préclassique » de langue française s'étend sur tout le XVIe siècle et la première moitié du XVIIe siècle. Certaines caractéristiques de cette période, aujourd'hui désuètes, perdurent même jusqu'au XVIIIe siècle. Les écrits de cette époque, qui correspond aux début de l'imprimerie, présentent un début de normalisation graphique, mais la graphie est encore loin d'être stabilisée.

Cet état de la langue est encore peu traité, contrairement à la période médiévale, pour laquelle des ressources ont été créées ces dernières années. Ainsi, la *Base de Français Médiéval*[1] et le *Nouveau Corpus d'Amsterdam* (Gleßgen & Vachon, 2010) offrent des corpus étiquetés pour la période IXe-XVe siècles. À l'opposé, les corpus annotés du français moderne sont nombreux, mais la base *Frantext* catégorisée, par exemple, ne remonte pas avant 1850. Entre ces deux périodes, on peut citer principalement le corpus diachronique *Modéliser le changement : les voies du français*[2], dont quelques textes sont annotés. Toutefois, seuls trois textes annotés (environ 180 000 mots) correspondent à notre période. Il n'existe ainsi guère de ressources et de corpus étiquetés pour le français préclassique et classique.

Il est vrai que cet état de la langue ne présente pas les mêmes difficultés que la langue médiévale. À la différence de cette dernière, le français préclassique reste relativement intelligible pour un locuteur moderne (exemple en Figure 1) ; c'est surtout la variabilité graphique qui pose problème. Il semble ainsi possible d'adapter des ressources et des outils conçus pour le français moderne pour le traitement de cet état de la langue. C'est ce travail d'adaptation qui est présenté

[1] *BFM - Base de Français Médiéval* [En ligne]. Lyon : ENS de Lyon, Laboratoire ICAR, 2012, http://bfm.ens-lyon.fr.

[2] Corpus MCVF annoté XML, sous la direction de France Martineau, avec Paul Hirschbühler, Serge Lusignan, Christiane-Marchello-Nizia, Yves Charles Morin et François Rouget.

ici. Nous nous focaliserons sur la période préclassique (environ 1500–1650), qui est la plus problématique, mais montrerons que cette approche est aussi valable pour la période classique (environ 1651–1800).

2 Le français préclassique

2.1 Caractéristiques

La langue du XVIe est une langue de transition entre la langue à forte variation graphique du Moyen-Âge, sans norme, avec des marquages régionaux, et une langue normalisée, le français classique, qui n'est pas encore la langue moderne que nous connaissons. Au cours du siècle, les graphies évoluent, sans doute sous l'influence de l'impression : les imprimeurs proposent de nouvelles règles typographiques, par exemple les diacritiques. Les textes de nos corpus sont transmis à travers des éditeurs scientifiques, qui soit respectent le texte (édition diplomatique), soit modernisent partiellement (*u/v*, *i/j*) ou entièrement le texte. Lorsque c'est possible, nous privilégions les éditions les plus proches du texte original, c'est à dire conservant les variantes graphiques, typographie exceptée : le corpus ne garde ainsi pas trace des caractères désuets comme le «s» long.

Pour se rendre compte de l'ampleur de cette variation graphique, on peut se référer aux attestations du lemme « *fruit* » dans le corpus Frantext pour la période 1501-1650 : *frui* (24 attestations), *fruict* (1150), *fruicts* (767), *fruictz* (73), *fruis* (2), *fruit* (697), *fruits* (431), *fruitz* (43), *fruyct* (10), *fruycts* (1), *fruyctz* (9), *fruyt* (10). Les traitements vont devoir tenir compte de cette variabilité graphique.

C'est assez dict pour ceste foys.
Quand sçavoir en vous s'assocye,
Monsieur Rien, l'on vous remercye
Du bien qu'avons aprins de vous.
Bazochiens, entendez tous :
Je veulx en triumphant arroy
Eslire et faire ung nouveau roy,
Comme il est coustume de faire ;
Pourtant chacun pense a l'affaire,
Autant les grandz que les petitz,
Et faire les preparatifz ;
Car, ainsi comme liberalle,
Je tendz a monstre generalle
Qui, l'esté qui vient, sera faicte.
En honneur du triumphe et feste,
Ne faillez monstrer vos bons cueurs
Qui font de la vertu approche,
Tant que l'on dye par honneurs :
Vive l'excellente Bazoche !

FIGURE 1: Extrait de *Sottie pour le cry de la bazoche*, Anonyme, 1549

2.2 Constitution d'un corpus et d'un jeu d'étiquettes

Le corpus diachronique du français que nous développons couvre actuellement la période XVIe–XXe siècles. Pour la période XVIe–XVIIIe siècles (périodes préclassique et classique), nous disposons de 189 textes issus de Frantext, des Bibliothèques Virtuelles Humanistes (base Epistémon), des Corpus Électroniques de la Première Modernité et de l'*American and French Research on the Treasury of the French Language*. Comme beaucoup de textes littéraires, même anciens, la plupart d'entre eux sont malheureusement sous droits d'éditeur et ne peuvent pas être redistribués librement. Toutefois, 37 textes (soit environ 2 millions de mots qui constituent le *corpus noyau*, couvrant plusieurs genres littéraires sur toute la période) pourront être diffusés sous licence *Creative Commons* à l'issue du projet.

Le jeu d'étiquettes morpho-syntaxiques est adapté au caractère diachronique du corpus, et vise à simplifier l'annotation de ce dernier. En langue moderne il est parfois délicat d'assigner une forme à une catégorie, même en contexte. Les catégories participe, gérondif et adjectif sont ainsi parfois difficiles à distinguer en langue moderne ; en langue ancienne, c'est même souvent impossible pour un locuteur non spécialiste de la période considérée. Nous avons donc décidé d'adopter un jeu de catégories simple. D'une part, l'annotation s'en tient essentiellement aux parties de discours (verbe, substantif, etc.), et ne « déborde » pas, comme c'est souvent le cas, sur des informations flexionnelles (par exemple mode, temps, etc. pour les verbes). D'autre part, l'annotation introduit des catégories regroupant des cas souvent indécidables ; par exemple, nous avons décidé de créer une étiquette regroupant les adjectifs, participes et gérondifs lorsque leur distinction est difficile, ce qui bien sûr a aussi des conséquences sur la définition des paradigmes verbaux et des auxiliaires.

3 Création de ressources pour le français préclassique

La création de ressources pour les langues anciennes peut s'envisager dans le cadre de la création de ressources pour les langues peu dotées et non standardisées. Notre approche s'inscrit à la suite des travaux de (Sánchez-Marco *& al.*, 2011) pour l'espagnol ancien. Elle consiste à annoter automatiquement un corpus d'apprentissage avec un analyseur pour la langue moderne, puis à corriger manuellement cette annotation. Ce corpus d'apprentissage est ensuite utilisé conjointement à un lexique « archaïsé » à l'aide de règles, pour construire un modèle de langage.

Dans notre cas, le corpus d'apprentissage, en graphie ancienne, est « modernisé » à l'aide du lexique archaïsé, *avant* son traitement avec l'analyseur moderne (cf. section 3.2.1), ce qui permet de réduire l'ampleur du travail de correction manuelle.

Ces ressources seront publiées sous licence *Creative Commons* à l'issue du projet.

3.1 Création du lexique

Le lexique associe un lemme (moderne quand il existe) et une partie du discours, avec les formes rencontrées dans le corpus ; on parle de lexique morphologique. La construction de ce lexique s'appuie principalement sur « l'archaïsation » d'un lexique moderne, mais incorpore aussi quelques éléments empruntés à un lexique du français médiéval, surtout utiles en début de période.

3.1.1 Ressources utilisées

Il existe des lexiques morphologiques pour le français moderne. Nous avons choisi le Le*fff* (Sagot, 2010) car il nous semblait mieux adapté à nos besoins initiaux dans sa version adaptée pour Freeling[3] (en particulier en ce qui concerne les étiquettes, de type EAGLES). Nous avons aussi utilisé Morphalou (Romary & al., 2004) qui nous a semblé plus complet dans la nomenclature de lemmes[4] et pour son appui sur le dictionnaire de référence qu'est le Trésor de la langue Français (TLF). Pour les états médiévaux de la langue, nous nous appuyons sur la nomenclature du Dictionnaire du Moyen Français (1330-1500) (DMF) et sur son lemmatiseur LGeRM (Souvay & Pierrel, 2009). Ce dernier a l'avantage de posséder deux lexiques morphologiques, un pour la période médiévale, et un plus adapté à la langue du XVIIe (appelé « mode » pour « moderne étendu »).

La validation du lexique morphologique va s'appuyer sur le corpus Frantext[5] qui couvre tous les états du français.

3.1.2 Processus de construction

La construction du lexique va se dérouler en 4 grandes étapes qui constituent un cycle. Plusieurs cycles vont être nécessaires pour obtenir un lexique ayant un taux de couverture satisfaisant pour traiter les textes.

La première étape consiste à créer la nomenclature de lemmes et leur flexion moderne. Le lexique de départ Le*fff* est tout d'abord adapté à nos étiquettes morphosyntaxiques. Il est ensuite complété avec les lemmes manquants, pris dans Morphalou pour les modernes, et dans LGeRM pour les médiévaux. Une nomenclature complémentaire de lemmes est ajoutée pour couvrir les lemmes absents des deux dictionnaires de référence (TLF et DMF). Ces nouveaux lemmes sont détectés à la fin d'un cycle.

La deuxième étape consiste à archaïser le lexique. Des règles d'archaïsation sont utilisées pour produire la flexion et la variation spécifique à la langue du XVIe siècle. Elles sont prises dans la base de connaissances du lemmatiseur LGeRM, qui couvre bien les états anciens du français. Afin de ne pas produire un lexique trop volumineux, certaines hypothèses de variations graphiques sont validées par attestation dans le corpus de référence (corpus à annoter et corpus Frantext). Les règles sont appliquées les unes à la suite des autres sur tous les mots du lexique. Une seule règle à la fois est appliquée sur un mot. Il faut donc itérer le processus au cas où plusieurs règles pourraient s'appliquer. Le nombre d'itérations est fixé à trois ; en effet, il paraît incertain d'appliquer plus de trois règles sur un mot (le « bruit » devient alors important), alors que le gain de couverture après deux itérations est déjà faible (*cf.* évaluation).

La troisième étape consiste à compléter le lexique en puisant automatiquement dans les ressources existantes. L'idée est de regarder si les mots absents du lexique mais présents dans les corpus textuels peuvent être analysés. Tous d'abord quelques règles de LGeRM trop générales pour être appliquées sans risque à l'étape deux sont testées (par exemple *an→ en*, *ain→ein*). Ensuite les lexiques LGeRM médiéval et LGeRM moderne étendu sont utilisés, on prend sans vérifier les analyses proposées. Au départ du projet ils étaient les plus riches en termes de variantes graphiques.

---

[3] http://nlp.lsi.upc.edu/freeling

[4] Environ 95 000 lemmes pour Morphalou 2, contre par exemple environ 68 000 lemmes dans le Le*fff* 3.2, ou environ 51 000 lemmes dans la version Freeling du Le*fff*, si l'on excepte à chaque fois les noms propres, que nous traitons à part.

[5] 4 515 textes, 270 millions de mots.

La quatrième étape consiste à analyser les résultats pour détecter les règles manquantes et les lemmes absents de la nomenclature. Le lemmatiseur LGeRM est alors configuré pour proposer des hypothèses, qui sont évaluées manuellement. En effet, ce processus ne peut être fait automatiquement car il produirait trop de bruit, en tout cas sur les premiers cycles de la construction du lexique.

3.2 Création du corpus d'apprentissage

Le corpus d'apprentissage comporte environ 62 000 mots, issus de 5 textes, échelonnés entre 1547 et 1776, pour couvrir aussi la période classique. L'annotation de ce corpus s'est effectuée en plusieurs étapes : une initialisation (« *bootstrapping* ») automatique avec des ressources modernes, puis une correction manuelle.

3.2.1 Initialisation du corpus d'apprentissage : projection lexicale et désambiguïsation

Dans un premier temps, le lexique a simplement été « projeté » sur le corpus. Du fait de la grande variété des formes du corpus, cela impliquait une ambiguïté très importante : à chaque forme pouvaient correspondre un grand nombre de parties du discours et de lemmes. Afin de réduire cette ambiguïté, le corpus a ensuite été « modernisé » à l'aide d'une variante du lexique morphologique, c'est à dire qu'à chaque forme ancienne a été associée une forme moderne. Puis il a été annoté automatiquement, sur la base de ces formes modernisées, à l'aide de TreeTagger (Schmid, 1994) et d'un modèle de langage spécifique, développé par Achim Strein sur un corpus de français moderne, en conservant, pour chaque token, toutes les annotations dont la probabilité dépassait 10 % selon TreeTagger. L'intersection entre l'analyse obtenue par projection lexicale et l'analyse obtenue par TreeTagger (en ne conservant que les analyses validées par les deux méthodes) a ainsi permis de réduire fortement l'ambiguïté, ramenant à 20,4 % le nombre de tokens ambigus (10 200 tokens, sur les 62 000 de départ).

3.2.2 Désambiguïsation et validation manuelles

Lors de l'étape suivante, trois annotateurs experts ont vérifié indépendamment l'annotation obtenue pour chacun des tokens désambiguïsés automatiquement, et annoté manuellement les 10 200 tokens qui ne l'avaient pas été. Généralement, il s'agissait alors seulement de sélectionner l'une des analyses parmi lesquelles il avait été impossible de trancher automatiquement.

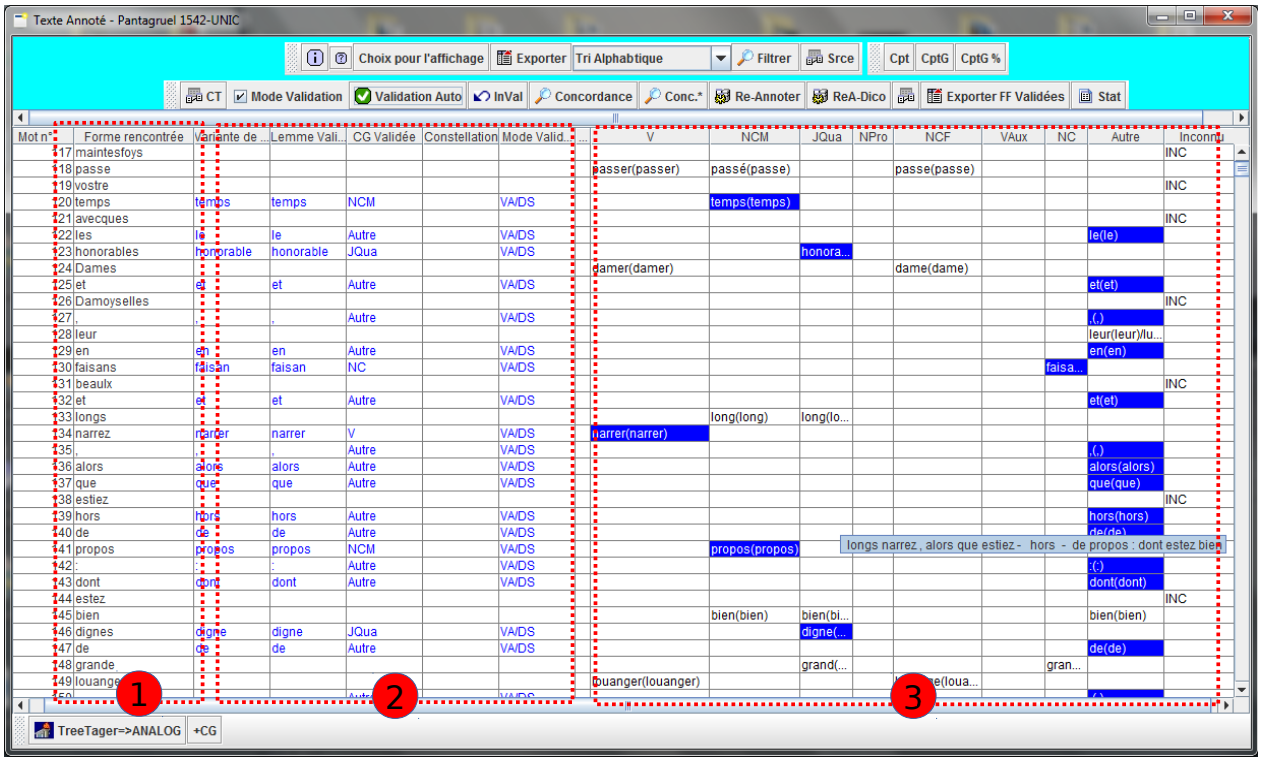

Figure 2: Capture d'écran d'AnaLog, fenêtre de visualisation du croisement du texte et des ressources sur l'analyse de Pantagruel. *À gauche (1), affichage des formes du corpus. Au centre (2), analyse retenue, validée automatiquement (VA/DS) ou manuellement (VM/DS). À droite (3), analyses suggérées pour les cas ambigus.*

Pour ce travail d'amendement, de validation et de correction de corpus, nous avons utilisé AnaLog, un outil dédié à l'exploration humaniste des textes (Lay & Pincemin, 2010). Il propose des fonctionnalités de type « *fonctionnalités documentaires*[6] » disponibles dans les outils d'analyse de données textuelles, en les généralisant et en les adaptant à un environnement de travail du type « linguistique sur corpus ». Ainsi, les index et pourcentages fournis peuvent porter sur toutes les informations disponibles (formes graphiques et variantes, séquences de catégories, ressources dictionnairiques utilisées pour l'annotation, etc), et le concordancier permet d'afficher chacune de ces informations.

Par ailleurs, AnaLog a pour ambition de permettre l'exploration des textes en rendant possible leur annotation manuelle systématique : les concordances permettent de mettre au jour des types d'informations récurrentes (ou non, par contraste), et l'étude des données repose sur l'observation de ces récurrences et l'élaboration de catégories permettant d'en rendre compte : les étiquettes. Ces étiquettes sont dynamiques et peuvent être créées ou supprimées librement, par exemple pour créer des étiquettes de travail temporaires.

Ces fonctionnalités, disponibles à partir d'un corpus brut, le sont aussi en partant d'une pré-annotation, comme c'est le cas ici. Dans le cas où les formes rencontrées ne sont pas ambiguës, on peut les valider en une seule fois. À l'inverse, dans le cas où la ressource propose de multiples annotations, elles sont toutes visibles et l'on procède à une désambiguïsation manuelle, soit au fil du texte (Figure 2), soit en validant un choix pour toute une série de formes répondant à une requête *via* le concordancier.

Le corpus d'apprentissage a ainsi pu être corrigé et désambiguïsé manuellement, à l'aide d'AnaLog, par trois experts[7]. Ces derniers se sont avérés en désaccord dans 9 % des cas. Nous avons alors désambiguïsé automatiquement les cas les plus « évidents », notamment lorsque deux des trois annotateurs étaient d'accord, et lorsque la divergence entre annotateurs ne portait que sur les diacritiques du lemme. Les 5,7 % d'ambiguïtés restantes (3 534 tokens) ont ensuite été résolues manuellement par un expert, qui a « tranché », au cas par cas.

4 Analyse du corpus et évaluation

Le lexique et le corpus d'apprentissage ont ensuite été utilisés pour entraîner un modèle de langage spécifique.

4.1 Couverture lexicale

Nous évaluons la qualité du lexique en regardant son taux de couverture en termes de fréquence sur un corpus de référence (Figure 3). En l'occurrence, nous utilisons Frantext, qui permet d'évaluer la couverture sur toute la chronologie du français. En abscisse du graphique on trouve la tranche temporelle (50 ans) et le nombre de textes concernés. Globalement les taux de couverture sont bons à partir de XVe siècle, jusqu'à la période moderne. L'analyse des lacunes du lexique montre une forte proportion de noms propres et de mots étrangers (essentiellement les mots latins). En terme de graphie on remarque une forte proportion d'hapax. Il s'agit souvent de variantes exotiques ou d'erreurs (numérisation, rupture de mots, impression).

4.2 Évaluation de l'annotation

Chaque texte du corpus d'apprentissage a été découpé en trois parties, selon un ratio 8/1/1. La première partie servait pour l'apprentissage proprement dit, la seconde pour le développement, et la dernière pour l'évaluation. Les résultats de cette évaluation sont synthétisés dans le Tableau 1. Dans un premier temps, nous avons évalué l'exactitude de la projection lexicale seule, sans utiliser de modèle de langage. Dans les nombreux cas d'ambiguïté, une analyse était simplement tirée au sort. La précision obtenue, de 60 %, est évidemment très mauvaise. Nous avons ensuite évalué l'approche par modernisation du lexique, et désambiguïsation en utilisant TreeTagger avec un modèle de langage du français moderne (approche décrite dans la partie 3.2.1). La précision est alors meilleure (81,1 %), et finalement assez proche de la précision du modèle entraîné spécifiquement sur notre corpus d'apprentissage (cf. partie 3.2.2, précision 83,3 %). Il s'agit toutefois de résultats intermédiaires, obtenus à partir de ressources non finalisées. À l'issue d'un important travail sur la normalisation des textes, l'optimisation de la chaîne de traitement et l'adaptation du lexique, nous avons pu obtenir une précision de 94,3 %.

---

6 Pour reprendre la distinction faite par Hyperbase entre fonctions statistiques et documentaires.

7 Deux experts ont annoté tout le corpus d'apprentissage, et le dernier seulement la moitié.

Projection lexicale	Modernisation	Modèle spécifique intermédiaire	Modèle spécifique finalisé
60 %	81,1 %	83,3 %	94,3 %

Tableau 1: Précision obtenue en fonction de la méthode d'annotation.

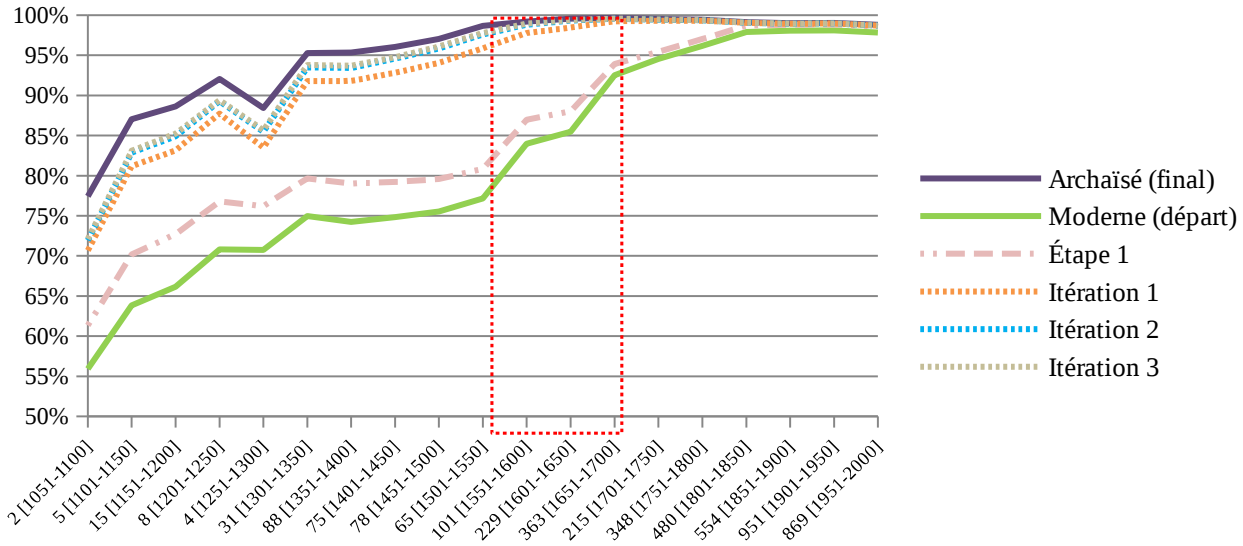

Figure 3: Taux de couverture lexicale du lexique moderne (lexique de départ) et du lexique archaïsé (lexique final), avec étapes intermédiaires, mesurée sur le corpus Frantext (XIème - XXème siècle).

Ces scores sont à mettre en regard de la relative simplicité du jeu d'étiquettes (cf. partie 2.2). De plus, notre évaluation concerne uniquement les étiquettes, et non les lemmes, qui ne sont actuellement pas désambiguïsés. Il reste donc encore une certaine marge de progression, avant de se rapprocher des scores obtenus pour le français moderne, aux environs de 96 % par exemple avec TreeTagger et le jeu d'étiquette GRACE, beaucoup plus complexe que le notre (Allauzen & Bonneau-Maynard, 2008).

Ces résultats sont homogènes sur la période préclassique et classique. Nous avons envisagé la création de modèles de langages distincts en fonction des périodes, mais n'avons pas constaté de gain significatif ; nous préférons donc nous en tenir à la simplicité d'un modèle unique, « panchronique », pour la période préclassique et classique.

5 Conclusion

Ce travail montre qu'il est tout à fait possible d'analyser des textes en français préclassique en adaptant des ressources conçues pour le français moderne. Les outils et les ressources développés dans ce cadre seront librement utilisables à l'issue du projet, ce qui contribuera à combler le manque entre la période médiévale et la période moderne.

Nous envisageons de poursuivre dans cette optique pour le traitement du français moderne, mais surtout du français médiéval. Il est certain que cette approche donnera des résultats de plus en plus dégradés au fur et à mesure que l'on remontera le temps, mais dans quelle mesure ? Enfin, au-delà de la création de ressources, nous envisageons une étude en diachronie longue de la langue française, notamment en adaptant certaines méthodes de *clustering* pour le travail en diachronie. Nous prévoyons aussi une évaluation de cette approche avec d'autres analyseurs : MElt, Morfette, etc.

Remerciements

Ce travail est issu du projet Presto, cofinancé par l'*Agence Nationale de la Recherche* et la *Deutsche Forschungsgemeinschaft*.

Références

ALLAUZEN A., BONNEAU-MAYNARD H. (2008). Training and evaluation of POS taggers on the French MULTITAG corpus. Actes de *International Conference on Language Resources and Evaluation*, Marrakesh, Maroc.

GLEßGEN M.-D., VACHON C. (2010). *Répertoire bibliographique du Nouveau Corpus d'Amsterdam, établi par Anthonij Dees et Piet Van Reenen (Amsterdam 1987), revu et élargi par M.-D.G. et C.V.*, 3. ed., Stuttgart: Institut für Linguistik/Romanistik.

LAY M.-H., PINCEMIN B. (2010). Pour une exploration humaniste des textes : AnaLog. Actes des *Journées internationales d'Analyse statistique des Données Textuelles*, Rome, Italie.

ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004). Standards going concrete : from LMF to Morphalou. *Workshop on Electronic Dictionaries, Coling 2004*, Genève, Suisse.

SAGOT B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th international conference on Language Resources and Evaluation* (LREC 2010), Istanbul, Turquie.

SÁNCHEZ-MARCO CRISTINA, BOLEDA GEMMA, PADRÓ LLUÍS (2011). Extending the tool, or how to annotate historical language varieties. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–9, Portland, Oregon, États-Unis.

SCHMID H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees, actes de *International Conference on New Methods in Language Processing*, Manchester, Royaume-Uni.

SOUVAY G., PIERREL J.-M. (2009). LGeRM : lemmatisation de mots en moyen français. *Traitement Automatique des Langues*, 50-2.