Adaptation d'un lemmatiseur au corpus rabelaisien : naissance d'Humanistica

Marie-Hélène Antoni et Marie-Luce Demonet
MSHS de Poitiers – Faculté des lettres et langues – F-86000 Poitiers- France

Abstract

The "annotated Rabelais" (Humanistica) is a project which is a sequel to the previous CD-ROM of the Electro-chronicques, published in 1995. As for other French texts of the same period, the user has to guess all the spelling combinations to retrieve the form he is looking for. In order to go a step further and to offer increased possibilities of consultation, we needed to solve the problem of heterography, so that we could count not only graphic forms but also lemmatised words. We began to adapt a modern French analyser to an Early modern or Middle French corpus. Thanks to the collaboration of IBM France, we are now pleased to present the first results of this tagging, which, at a rather satisfying level, takes into account the special features of Middle French affecting lexicographic as well as syntactic aspects. The processed sample offers such a large number of technical and linguistic problems that their solutions could be easily transferred to a much larger corpus, and extended to problems of authenticity.

Résumé

L'annotation du corpus rabelaisien s'incrit au départ dans un projet d'édition ayant donné lieu, en 1995, à la parution des Electro-Chronicques. Ce projet répond à un besoin spécifique des états de langue ancien du français : la gestion de ce qu'il est convenu d'appeler l'hétérographie, c'est-à-dire le fait que les états de langue antérieurs au XIX° ne connaissent pas de norme orthographique, et que la graphie est d'autant plus éloignée des formes standard que le texte d'origine est éloigné dans le temps. Cet état de fait prive les chercheurs littéraires d'outils de traitement linguistique automatiques (appliqués à des problématiques telles que l'authenticité p.e.) dont peuvent bénéficier leurs collègues travaillant sur des périodes contemporaines : il leur faut deviner les combinaisons graphiques possibles pour accéder aux formes recherchées. L'idée a donc été de modifier un lemmatiseur existant pour le français contemporain, afin qu'il soit à même de prendre en compte les caractéristiques du moyen français. Nous sommes ici en mesure de présenter les premiers résultats de cette entreprise sur le corpus rabelaisien, et les méthodes que nous avons choisies afin d'être en mesure de transférer cette méthodologie sur d'autres états de la langue.

Mots-clés: Français de la Renaissance - lemmatisation - hétérographie - éditions de textes anciens

1. Introduction

Les états de langue anciens du français, présentent un certain nombre de particularités lexicales, syntaxiques et orthographiques : il faudra attendre le XIX° siècle pour que l'orthographe se stabilise vraiment et que la notion de "faute d'orthographe" prenne son sens. Le mot *côté* pourra se trouver sur une seule et même page orthographié *cousté*, *costé*, *coste* et, dans d'autres pages encore autrement …Cet état de fait ne perturbe pas trop la lecture humaine et ajoute même, au plaisir de la lecture, celui d'un sentiment d'étrangeté qui va du "charmant" au "truculent". Les lemmatiseurs quant à eux ne semblent pas goûter outre mesure la chose, et se contentent d'un imperturbable "unknown word". Nous voilà donc mis en demeure de répondre, et ce pour une communauté relativement restreinte, à une préccupation qui concerne un ensemble de corpus clairement délimité : la gestion de ce qu'il est convenu d'appeler l'hétérographie sur un corpus de littérature en français préclassique ; en effet cette situation prive les chercheurs littéraires de l'exploitation complète des outils de traitement linguistique automatiques dont peuvent bénéficier leurs collègues travaillant sur des périodes

contemporaines. En tout premier lieu, ces outils peuvent aider à la création de ces précieux index et concordances que l'on associe aux œuvres littéraires anciennes. Sur base de recherche d'expressions régulières, de différents tris et analyses statistiques, les outils informatiques permettent de donner une toute autre ampleur au diverses analyses léxicométriques, stylométriques.. Ils ne prennent tout leur sens que si l'on peut travailler sur le lemme, et pas seulement sur les formes graphiques rencontrées. Ceci est d'autant plus vrai pour le français préclassique, que les chaînes de caractères peuvent être extrêmement variables, et qu'il faut pour ainsi dire deviner les combinaisons graphiques possibles pour accéder aux formes recherchéees.

L'idée a donc été de modifier un lemmatiseur existant pour le français contemporain, afin de pouvoir prendre en compte cette particularité. Nous nous sommes basées sur un lemmatiseur robuste, ayant traité des masses considérables de corpus, et dont nous connaissions bien les particularités internes : le tagger TAL, élaboré par la société IBM. Cet outil a participé à l'évaluation Grace (action d'évaluation INaLF-CNRS qui s'est déroulée de 1996 à 1998).

Avant de préciser ce en quoi a consisté cet aménagement, nous présenterons le corpus sur lequel nous avons travaillé,

- du point de vue de son origine, l'édition de ce corpus ayant en soi fait l'objet d'un labeur considérable ;
- du point de vue de ses caratéristiques linguistiques.

2. Etablissement du corpus Rabelaisien.

Le CD-ROM Rabelais des *Electro-chronicques*[1], le premier du genre (1995[2]), a été élaboré sur la base des textes que le laboratoire EQUIL XVI de l'Université Blaise Pascal à Clermont-Ferrand a établis d'après les exemplaires originaux de référence.

Le premier objectif du travail avait été la lecture optique de la première édition du Seuil, à partir de laquelle il avait été possible de construire un index des fréquences (la *Concordance* de Dixon n'était pas encore publiée[3]). Cette édition modifiant naturellement les graphies et la ponctuation selon les principes éditoriaux de l'époque, les recherches de linguistique quantitative que nos collègues entendaient mener pouvaient difficilement être effectuées à partir d'un texte ainsi modifié. La deuxième étape du travail a donc été de revenir aux exemplaires originaux, décision prise à Clermont-Ferrand et adoptée dans les éditions électroniques parues depuis à Poitiers[4]. Le retour aux exemplaires dits de référence (conservés à la BNF, sauf pour la *Briefve Declaration* où nous avons eu aussi recours à l'exemplaire du British Museum), se justifie par la volonté de pouvoir travailler sur les graphies d'époque, qu'elles nous paraissent aberrantes ou non.

[1] Une version en ligne est disponible sur www.lolita.unice.fr

[2] Équipe éditoriale: laboratoire EQUIL XVI, Clermont-Ferrand, dirigé par Guy Demerson puis par M.-L. Demonet (1991-1997). Réalisation technique: Etienne Brunet. Le " corpus Rabelais " des éditions Champion, annoncé depuis juin 1995, n'a pas encore été commercialisé.

[3] Dixon et Dawson, 1992; le texte de référence est celui de l'édition Droz.

[4] Sur le site Epistemon, à la page " Cornucopie ": *Microcosme*, *La Saulsaye*, *La Deplourable fin de Flamette*, de Maurice Scève; *Dialogues de langues*, traduction C. Gruget, de Speroni. En préparation: *Dialogue de Rhetorique*, de Speroni; Boccace, traduction des livres I et II par Laurent de Premierfaict des *Nobles malheureux*; Tory, *Champfleury*; *Delie* (1544) de Scève, etc.

3. Les incidences de l'hétérographie

Une aide considérable est venue de Nice puisque c'est le logiciel Hyperbase écrit par Etienne Brunet (laboratoire du CNRS- INaLF) qui a permis de traiter le corpus ainsi revenu aux sources d'époque. L'avantage d'Hyperbase fut essentiellement son système de traitement statistique des données qui pouvait apporter une aide appréciable dans l'analyse de la richesse du vocabulaire ; il a permis de progresser dans l' examen des problèmes d'authenticité le *Cinquiesme Livre*, soumis au traitement d'Hyperbase, confirme pour le moment l'hypothèse de " brouillons " arrangés par un mystérieux facteur[5]. Toutefois, les résultats obtenus sont à interpréter avec une grande prudence, le logiciel traitant les graphies et non les mots. En l'absence de lemmatisation, les statistiques conservent une assez grande marge d'erreur.

En poursuivant la réflexion amorcée depuis dix ans sur ce problème apparemment insoluble de la gestion des hétérographes, nous est apparue l'urgence d'améliorer les produits consultables actuellement sur deux fronts :

- celui de la consultation par celui qui fait une requête, qu'il soit savant ou non
- celui du traitement statistique pour des recherches avancées (problèmes d'authenticité et analyse automatique des faits de langue et du vocabulaire)

Les deux approches demandent de toute manière une lemmatisation du corpus. L'utilisateur ne peut deviner toutes les variantes possibles d'un mot, l'imagination graphique du français de la Renaissance étant sur ce point sans limite, surtout pour les emprunts, les néologismes ou les calques. Quant au spécialiste, il pourra décider de ne plus recourir à la méthode de l'échantillonnage, pour traiter cette fois des corpus complets.

Nous avons entrepris de mettre au point un prototype sur un petit corpus présentant une variabilité extrême (*Pantagruel* de 1542). Dans un premier temps, nous avons donc décidé de simplement " faire tourner " un analyseur classique sur le corpus de Rabelais, pour ensuite l'adapter. Le taggeur[6] utilisé ici nous permet de « récupérer » le texte annoté qui se présente comme le texte d'origine enrichi d'un certain nombre d'informations, catégorie grammaticale et lemme de rattachement de la forme rencontrée, ou, si la forme rencontrée n'a pas été identifiée, une information " mot inconnu " dans le corps du texte. Un fichier indépendant dans lequel sont recensées toutes les graphies inconnues est généré par ailleurs. Ce fichier est source d'information pour les enrichissements de dictionnaire : il contient ce qui est le plus atypique, c'est à dire le plus caractéristique d'un corpus. On y trouvera les particularités du corpus rabelaisien, formes hétérographes et hapax savoureux, par exemple :

Incornifistubile (inséré) et *desincornifistibulee*, la *gyrognomonique circumbilivagination* (ou pérégrination erratique et circulaire), *filopendole* (qui est pendu à un fil), *matagrabolisme* (élucubration), *emburelucoquees de guilverdons* (emberlificotées de capuchons*), bien grabelez a rouelle* (criblés de rondelles).

4. Typologie de l'hétérographie

En termes statistiques (sur *Pantagruel*) l'exploitation du fichier des "unknowns" nous indique que

[5] Sur les problèmes d'authenticité, voir M. Huchon, 1981 et l' état des lieux présenté par Marie-Luce Demonet au colloque de Rome sur le *Cinquiesme Livre* (octobre 1998), à par. aux éditions Droz.

[6] Il s'agit d'untaggeur développé par la société IBM, TAL. Cet analyseur est un analyseur statistique, présenté dans le cadre de l'action Grace, organisée conjointement par l'Aupelf-Uref et le CNRS.

- sur 47457 mots (tokenisés) identifiés par l'analyseur (le texte ferait donc une vingtaine de pages " standard "),
- 13246 soit près du tiers, ne sont pas reconnus. Cette proportion est tout à fait considérable si l'on pense au fait que bon nombre des tokens identifiés sont des déterminants, des prépositions ….
- 5246 graphies différentes sont identifiées ; statistiquement, une forme inconnue apparaît en moyenne trois fois ; mais cette moyenne globale recouvre une grande variété de situations : 5246 graphies différentes sont identifiées,

  4591 mots inconnus ont une fréquence <= 2 (soit 97%)

  3632 sont des happax (soit 75%)

Dans les faits, **3% des formes** représentent plus de **60% des occurrences** de mots inconnus. C'est donc sur l'observation de ces 3% des formes que nous avons initialement fondé notre stratégie d'élaboration d'Humanistica.

Comme nous l'avons signifié précédemment, la lecture humaine n'est pas perturbée outre mesure par ces variations orthographiques. L'un des faits les plus fréquemment évoqués comme gênant, la non-dissimilation u/v ou i/j, est en fait absolument régulier. Il y a donc fort à parier que la résolution de l'hétérographie par des calculs statistiques aurait été possible[7]. Ce n'est pas le choix qui a été retenu, pour une raison très simple : notre outil, Humanistica, s'adresse à des littéraires qui auront avantage en l'occurrence à connaître les règles qui s'appliquent,

- pour observer les contextes d'apparition de ces règles, sur un corpus d'un auteur, mais aussi en diachronie, afin de documenter par exemple un travail en graphématique diachronique,
- pour avoir éventuellement la maîtrise de la création ou de la suppression de certaines règles, toujours à des fins de description linguistique, ou encore à des fins de tuning de l'analyseur.

L'observation des 3% des formes représentant plus de 60% des occurrences de mots inconnus nous a permis de construire des règles qui agissent aussi sur les 97% restant. Les cas non traités relèvent alors d'une description lexicographique classique : pour l'ensemble du corpus rabelaisien, il nous faudra valider dans le détail environ 2000 lemmes, ce qui n'a rien de particulièrement effrayant. Il est tout à fait naturel, lorsqu'on se trouve confronté à un corpus d'un nouveau genre, de devoir décrire une frange de vocabulaire propre à ce corpus, surtout quand celui ci est l'œuvre d'un certain Rabelais.

On peut distinguer deux grands cas de figure : les expressions de la créativité lexicale d'un auteur en verve (que l'on ne peut résoudre par des règles), les expressions de l'hétérographie ambiante.

Cette hétérographie peut être envisagée de quatre façon différentes :

-1- Il s'agit d'un problème de segmentation : *lage* ou *leage* pour *l'âge*, *daultre* pour *d'autre*

[7] même si le problème n'est pas trivial : les calculs de distance par exemple ne permettent pas de traiter au mieux tous les cas de figure, et les modes de résolution ne sont pas élucidants.

-2- Il ne s'agit pas d'un phénomène régulier : incrémentation du fonds dictionnairique (*incornifistibulé*). Il s'agit alors, soit de la création d'un nouveau lemme, soit de la création d'une variante de lemme, variante qui sera rapportée à un lemme unique qui pourra être pris en compte comme tel dans les calculs ultérieurs (comme *scavoir, sçavoir* sur *savoir*).

-3- Il s'agit de modes de flexion « alternatifs », de régularités morphologiques que l'on peut résoudre en enrichissant les tables de flexion du dictionnaire de base. Une telle décision peut se baser sur des observations du type :

oit -> ait (685 fois) ; *oyt -> ait* (25 fois) ; *ois -> ais* (17 fois) ;

oient -> aient (89 fois) ; *oyent -> aient* (161 fois).

-4- Il s'agit de graphies « qui se prononcent comme », ou qui apparaissent comme des variantes stables pour un état de langue :

- on émet l' hypothèse que si un mot contenant un*y* n' est pas connu du dictionnaire, on cherche ce mot avec un *i* à la place du *y* . En effet *y* apparaît plus de 2000 fois, et seulement 143 fois isolé (comme pronom),

- quant au z, c'est 977 fois qu'il apparaît, dont 887 en finale ! on voit l'intérêt que l'on aura à essayer de le remplacer par *s* ! (il est la marque quasi-régulière du pluriel)

Ces règles s'avèrent donc d'une efficacité redoutable, et sont sans danger car elles correspondent véritablement à un phénomène diachronique stable, et n'affectent que légèrement les mots « réanalysés ». On pourra semblablement formaliser d'autres régularités morphématiques affectant la catégorie des affixes : si *ier* , essayer *er* (*bouchier*) ; si *iers -> ers* ; si *ierz -> ers*

De telles règles ne peuvent cependant être appliquées aveuglément . Il est nécessaire de respecter deux conditions :

- la première est que cette règle soit caractéristique d'une réalité statistique, et pas seulement de notre connaissance de l'évolution des graphies ;

Ainsi, une règle venait rapidement à l'esprit, celle de supposer une conversion possible « sc » -> « s », et ce sur la base du verbe « scavoir ».

Une brève interrogation du corpus a mis en évidence que sur 407 occurrences de « sc », 61 concernaient effectivement *scavoir,* (dont 52 *scav*-), mais que *scavoir* est en fait le seul lemme pour lequel cette règle s'applique dans le corpus, pour tant de *chascun, escumoit, empesche, science, escript, desconfite, preschans, eschaufee*. La diversité des formes amenait trop de risques de collusion avec d'autres règles, et il s'est avéré pour le coup préférable de renseigner le dictionnaire d'un nouveau verbe, qui se fléchit, disons le, de façon assez distrayante. Ce nouveau verbe, *scavoir*, est alors déclaré comme variante de *savoir*, forme sur laquelle il est lemmatisé.

- la seconde condition est que la règle ne provoque pas de catastrophes sur des zones de lexiques qui ne sont pas concernées.

Ce point évoqué, de l'impact d'une abondance de règles entrant en collusion les unes avec les autres ne doit pas être sous-estimé : ainsi, dans un certain état de notre travail, une forme non identifiée : « eschollier » s'est vue lemmatisée sur « écouler » ! ! ! par l'application intempestive de règles en cascade : *es->é; ch->c; o->ou; ll->l; ier->er*

Les solutions que l'on peut apporter résident essentiellement,

- dans le fait qu'on autorise l'application de ces règles dans le seul cas où lea forme n'a pu être idfentifiée par ailleurs. Cette contrainte a quelques inconvénients, ainsi les formes en *ez* qui peuvent être ramenées à un verbe le seront quoiqu'il arrive, même s'il s'agit en fait de noms : *narrez* sur *narrer* et non sur *narré* (au sens de récit); mais on génère sinon pléthore de formes qui affolent complètement l'analyseur.
- par la précision des contextes gauches et droits permettant l'application des règles (par exemple, on ne dit pas que *ul->u* mais que *(o/a)ul(t/x)-> (o/a)u(t/x).* Il faut ici mentionner qu'on ne peut hiérarchiser, ordonnancer l'application de ces règles qui sont compilées sous forme d'automates dans le module de tokenisation. Elles sont toutes appliquées simultanément, sans ralentir notablement les performances de l'analyseur (tout le corpus rabelaisien (1,4mega) est traité en moins de 2 minutes sur un petit portable. La progression en temps n'est par ailleurs pas linéaire puisque sur des temps aussi courts, tout ce qui relève du chargement et de l'accès au fichier prend une part importante du temps.

Dans les cas où les conflits ne peuvent être résolus par les stratégies évoquées ici, les formes rencontrées rejoignent le lot des hapax, et sont renseignées en temps que telles dans le dictionnaire, comme lemme autonome, ou comme variante (c'est le cas de *scavoir*)

Profitons en ici pour insister sur la nécessité de mesurer les conséquences de chacune des règles, et le souci de précision dans lequel les moindres répercussions ont été analysées et répercutées dans notre travail. Pour ce faire, nous avons effectué des contrôles systématiques à l'aide d'un browser faisant apparaître en couleur toutes les différences entre deux fichiers, ici deux états de sortie annotée. La petitesse du corpus initial et l'endurance du lexicographe et du seizièmiste ont rendu ce labeur possible.

Enfin, il nous faut encore souligner qu'adapter un analyseur à un état de langue antérieur ne revient pas simplement à gérer des problèmes d'hétérographie au niveau du lexique, mais qu'il s'agit bien

- de pouvoir s'adapter à des structures syntaxiques assez éloignées du français moderne : « Que ne envoyas tu la mort a moy premier que a elle ? » n'en est qu'un exemple, on en trouve de beaucoup plus éloignés de la norme contemporaine. Il faut souligner ici que le lemmatiseur que nous employons réagit étonnamment bien dans cette situation. Il faut dire aussi qu'il s'agit d'un analyseur reposant sur un modèle de langage statistique tri-classe, et qu'un analyseur syntaxique à proprement parler aurait sans nul doute eu beaucoup plus de difficultés. Il faut dire enfin que nous avons réduit nos ambitions en terme de précision : nous visons dans un premier temps à la seule attribution d'une partie du discours correcte.
- de prendre en compte ayant un l'impact qu'ont sur l'analyse elle-même différents "types" d'hétérographies : ainsi, il va de soi, pour un analyseur du français contemporain que « a » vient de « avoir », et que « à » est une préposition. Or, dans le corpus de Rabelais, « a » peut tout aussi bien être une préposition. Ce qui n'est pas sans incidence sur les analyses de « tout ce qui se trouve autour » sur l'axe syntagmatique. Il s'agissait là d'une épine qui nous a un peu inquiétées.

Les résultats que nous avons obtenus en l'espace d'une quinzaine de jours nous semblent tout à fait satisfaisants : l'intégralité du corpus de Rabelais a été traitée. Des 21395 mots non

identifiés au départ (un mot peut apparaître de 1 à n fois), il n'en reste que 9617. Notons que 7367 sont des happax dont l'origine est généralement étrangère (latin, allemand, limougeot et autres glosses). La situation sera donc encore nettement améliorée par le marquage de ces zones comme « à ne pas analyser ». De plus, la qualité globale de l'annotation n'a pas été « trop » dégradée : nos divers sondages dans le corpus nous laissent au dessus de 97% de réussite. Nos collègues seiziémistes peuvent dès lors se concentrer sur l'amendement des 3% qui restent, en vu de produire une édition annotée de Rabelais qui puisse faire référence dans la communauté des rabelaisants.

Pour donner un aperçu du résultat, nous en donnons ici un extrait :

```
#Pantagruel(Pantagruel,NP) ,(,,SEP) Roy(roi,NM)
des(des,DES) Dipsodes(Dipsodes,NPRO) ,(,,SEP)
restitue(restituer,V)  a(à,PREP) son(son,DPOS)
naturel(naturel,NM),(,,SEP) avec(avec,PREP) ses(son,DPOS)
faictz(fait,NM)  et(et,CCOO) prouesses(prouesse,NF)
espoventables(épouventable,AQ) :(:,SEP)
composez(composer,V) par(PREP) feu(NM) M.(monsieur,NM)
#Alcofribas(Alcofribas,NP) abstracteur(abstracteur,NM)
de(de,PREP) quinte(quint,AQ) essence(essence,NF) .(.,FIN)

Prologue(prologue,NM) de(de,PREP) l'(le,DDEF)
auteur(NM).(.,FIN

Tres(très,ADV)illustres(illustre,AQ) et(et,CCOO)
tres(très,ADV)chevaleureux(AQ) champions(champion,NM)
,(,,SEP) gentilz(gentil,AQ) hommes(homme,NM) ,(,,SEP)
et(CCOO) aultres(autre,AQ) ,(,,SEP) qui(PREL)
voluntiers(volontiers,ADV) vous(vous,PPSUJ)
adonnez(adonner,V) a(à,PREP) toutes(tout,AIND)
gentillesses(gentillesse,NF) et(et,CCOO)
honnestetez(honnêteté,NF)  ,(,,SEP)   vous(vous,PPSUJ)
avez(avoir,V) n'(ne,NEG) a(avoir,AUX) gueres(guère,ADV)
veu(voir,V) ,(,,SEP) leu(lire,V) ,(,,SEP) et(et,CCOO)
sceu(savoir,V) ,(,,SEP) les(le,DDEF) grandes(grand,AQ)
et(CCOO) inestimables(inestimable,AQ)
Chronicques(chronique,NF) de(de,PREP) l'(le,DDEF)
enorme(énorme,AQ) geant(géant,NM)
#Gargantua(Gargantua,NP) :(:,SEP)  et(et,CCOO)
comme(comme,ADV) vrays(vrai,AQ) fideles(fidèle,NM)
les(le,PPOBJ) avez(avoir,AUX) creues(croire,V) ,(,,SEP)
gualantement(galamment,ADV) ,(,,SEP) et(et,CCOO)
y(y,PPOBJ) avez(avoir,AUX) maintesfoys(maintefois,ADV)
passe(passer,V) vostre(votre,DPOS) temps(temps,NM)
avecques(avec,ADV) les(le,DDEF) honorables(honorable,AQ)
Dames(dame,NF) et(et,CCOO) Damoyselles(damoiselle,NF)
```

5. Conclusion

Le premier CD-ROM Rabelais, doté du puissant logiciel d'analyse statistique hyperbase, était toutefois limité par le fait qu'ill ne fonctionne que sur MacIntosh et qu'il ne résoud pas les problèmes liés à l'hétérographie. Le lemmatiseur adapté permettra de publier une mise à jour

disponible en ligne et sous Windows, associée à une procédure de requête assistée qui gèrera automatiquement les hétérographes.

Le prolongement de l'informatisation du corpus rabelaisien vers la lemmatisation d'un état de langue antérieur au français moderne est pour nous une perspective stimulante. Cette étape devrait marquer une réelle avancée dans la gestion documentaire des grands corpus et offrir des applications à d'autres états de la langue non standard. Son intérêt patrimonial est considérable, pour l'archivage et la consultation assistée des textes anciens, quelle que soit leur nature[8]. Associée à des bases déjà existantes mais non lemmatisées, ou à des logiciels de reconnaissance de caractère encore peu performants pour les documents anciens, elle devrait participer aux nouvelles formes de l'encyclopédisme, historique et numérique.

Références

Rabelais F. (XVI°)

- *Les Electro-chronicques de F. Rabelais*, Paris, Les Temps qui courent, 1995.
- *Œuvres*, édition établie et annotée par Mireille Huchon, Paris, Gallimard, « Bibliothèque de la Pléiade », 1994.
- *Œuvres*, édition établie, annotée et préfacée par Guy Demerson, texte original établi par Michel Renaud, avec une translation de G. Demerson, Paris, Seuil, 1995-1996.
- *Œuvres romanesques*, édition établie par le laboratoire EQUIL XVI de l' Université Blaise Pascal, sous la direction de M.-L. Demonet, Poitiers, La Licorne, 1999.

Antoni M.-H. (1997). TAL et le projet Grace. Rapport technique interne IBM.

Demonet M.-L. (1996). L'édition électronique d'une oeuvre littéraire : Les *Electro-chronicques* de Rabelais. *Banques de données et hypertextes pour l'analyse du roman*, éd. N. Ferrand, Presses Universitaires de France, pages119-136.

Demonet M.-L. (1997). Littérature de la Renaissance et informatique. Sur les *Electro-chronicques* de Rabelais . *Éditer et traduire Rabelais à travers les âges*. éd. Paul J. Smith, Amsterdam, Atlanta, pages 233-247.

Demonet M-L, Antoni M-H (1999) Informatisation et lemmatisation du corpus rabelaisien, in Le médiéviste et l'ordinateur, n°38 à paraître (hiver 1999-2000)

Dixon J. E. G. et Dawson J.L. (1992) *Concordance des oeuvres de François Rabelais*, Genève, Droz.

El-Bèze M. (1993) Les modèles de langage probabilistes: quelques domaines d'application. Thèse HDR. LIPN (Paris XIII).

[8] Ce projet ne pourrait pas être mené à bien sans le soutien financier exceptionnel de la MSHS et du CNRS. Les autres collaborateurs sont Pierre Martin et Stéphan Geonget (Université de Poitiers).