Annoter facilement un corpus complexe

Ariane Pinche

Actes des

Rencontres lyonnaises des jeunes chercheurs en linguistique historique

Édités par
Timothée Premat
Ariane Pinche

Diachronies Contemporaines

PREMAT, Timothée & PINCHE, Ariane (dir.) (2019). *Actes des rencontres lyonnaises des jeunes chercheurs en linguistique historique*. Lyon : Diachronies contemporaines, 70 p.

DOI : 10.5281/zenodo.3462309

Annoter facilement un corpus complexe

L'exemple de Pyrrha, interface de post correction, et Pie, lemmatiseur et tagueur morphosyntaxique, pour l'ancien français

Ariane PINCHE
Université Lyon 3 (CIHAM UMR 5648)
École nationale des chartes

Résumé

L'annotation morphosyntaxique est une tâche qui occupe activement la communauté scientifique depuis l'émergence des outils numériques dans les sciences humaines, car ces informations sont précieuses aussi bien aux yeux des linguistes que de chercheurs moins spécialisés en sciences du langage pour, entre autres, des études de stylométrie ou de lexicométrie. Toutefois, certaines langues, comme l'ancien français, représentent un véritable défi parce qu'il est très difficile d'annoter des langues sans normalisation graphique.

Pyrrha est une interface d'aide à l'annotation linguistique développée au sein de l'équipe numérique de l'École nationale des chartes. Cette dernière s'appuie sur le lemmatiseur et annotateur morphosyntaxique Pie qui ne possède pas de dictionnaire prédéfini. Pyrrha propose également une interface de relecture pour assurer à l'utilisateur une complète maîtrise de ses données. Ainsi, annoter un texte est plus rapide et plus simple. Il est alors facile de constituer à partir de son corpus un glossaire qui répertorie sous une entrée unique toutes les variations graphiques d'un même lemme ou d'analyser un corpus étendu pour observer, par exemple, l'état du système casuel ou encore détecter certains traits dialectaux.

Mots-clés : ancien français ; interface ; lemmatiseur ; morphosyntaxe ; tagueur.
DOI : 10.5281/zenodo.3464473

1. Introduction

Faire un commentaire linguistique sur un corpus donné est une tâche extrêmement fastidieuse qui demande de consacrer beaucoup de temps à la préparation de données qui bien souvent resteront dans l'ombre de la recherche. Se pose alors la question de l'exhaustivité du relevé, souvent écartée en regard du manque de rentabilité entre le temps consacré à la préparation du corpus et ses possibilités d'exploitation. Il semble alors souvent préférable de consacrer son travail à un phénomène délimité et aisément quantifiable offrant au chercheur une pleine maîtrise de son objet.

Pourtant, les corpus annotés peuvent se révéler très utiles dans de nombreux champs d'investigation en sciences humaines. Une annotation proposant des lemmes pour chaque mot peut aider à repérer des phénomènes phonétiques et à identifier certains traits linguistiques qui à terme pourront être des indices du lieu de composition. Une annotation morphosyntaxique est utile pour étudier des

phénomènes d'ampleur sur un corpus complet comme l'utilisation du système casuel ou encore l'utilisation ou non du pronom personnel sujet, phénomènes en ancien français qui permettent parfois de dater un texte. Elle permet aussi des études encore plus pointues comme les analyses de stylométrie, notamment pour les études d'attribution d'auteur (Mellet, 2002). L'étendue des possibilités fait de ces données cachées des éléments extrêmement précieux pour la recherche, même si l'exploitation immédiate d'une telle masse de données n'est pas toujours évidente.

L'annotation des corpus a souvent été automatisée par les linguistes notamment dans le champ des TAL (Traitement automatisé de la langue) pour ensuite réexploiter les données de manière quantitative. Des lemmatiseurs et annotateurs automatiques ont été développés depuis les années soixante[1]. Les outils existants sont relativement performants pour les langues normées, mais perdent en fiabilité dès lors qu'ils sont confrontés à une grande variété graphique et à une langue en transition où les marqueurs morphologiques ne sont pas toujours utilisés de manière homogène. En outre, bien souvent ces lemmatiseurs n'ont pas été pensés pour un usage par les novices en numérique et ne possèdent pas d'interface graphique qui permettrait un usage sans connaissance préalable en informatique, donnant ainsi parfois l'impression de priver le chercheur d'un complet contrôle sur son annotation.

En effet, la plupart des lemmatiseurs produisent des données brutes en CSV ou TSV sans proposer d'interface de correction de l'annotation. Pourtant cette étape peut s'avérer cruciale pour la qualité du travail scientifique, ne serait-ce que pour un contrôle de la qualité du corpus. Cette étape permet d'améliorer l'annotation du set de données et d'éliminer le « bruit » lié au traitement automatique pour garantir la fiabilité des résultats de l'exploitation future des données et laisser une pleine maîtrise de ces dernières à l'humain. Ces points sont d'autant plus importants quand la production des données dépasse le simple cadre d'un usage individuel pour constituer un corpus annoté pérenne qui pourra par la suite être réutilisé par la communauté scientifique.

Face à ce constat, l'interface d'annotation et de post-correction Pyrrha a été développée dans le cadre de la cellule humanité numérique de l'École nationale des Chartes (Clérice, Pilla & Camps, 2018) pour proposer une interface simple d'utilisation qui permette de contrôler tout le cycle d'annotation de son corpus. On peut alors lemmatiser un texte en ancien français ou en latin, contrôler la qualité des données produites, voire les corriger si nécessaire grâce à l'interface de post-correction, et enfin les télécharger en CSV ou en XML-TEI pour leur exploitation et leur conservation.

2. Lemmatisation d'un corpus

L'annotation morphosyntaxique est une tâche qui occupe activement la recherche depuis l'émergence des outils numériques dans les sciences humaines, toutefois, certaines langues, comme l'ancien français, représentent un véritable défi parce qu'il est très difficile d'annoter des langues sans normalisation graphique, les lemmatiseurs traditionnels fonctionnant à partir de règles prédéfinies et d'un dictionnaire établi.

[1] On peut citer l'entreprise de lemmatisation, démarrée dans les années soixante et toujours en cours, des textes latins de l'équipe de recherche du LASLA (Laboratoire d'analyse statistique des langues anciennes de l'université de Liège).

2.1. Présentation de l'annotateur Pie

Pyrrha s'appuie sur l'annotateur Pie (Manjavacas, Kestemont & Clérice, 2019) qui associe à chaque mot un lemme, une nature[2] et une analyse morphosyntaxique. Afin de s'affranchir des problématiques des lemmatiseurs traditionnels, Pie a été conçu à partir d'algorithmes d'intelligence artificielle qui ne s'appuient pas sur un dictionnaire (Manjavacas, Kádár & Kestemont, 2019), ce qui le rend indépendant du langage auquel il est appliqué et rend possible le traitement des langues à forte variation graphique ou dont les règles morphosyntaxiques ne sont pas toujours suivies de manière homogène[3]. L'algorithme est capable d'apprendre la langue des textes sur lesquels il est appliqué par comparaison entre les résultats qu'il parvient à obtenir par lui-même et les résultats du corpus d'entraînement qui a été préalablement annoté à la main. Les performances de Pie s'améliorent au fur et à mesure des entraînements et plus le corpus réunit des textes nombreux et variés, plus la fiabilité des résultats de l'annotation automatique augmente. Les données d'entraînement permettent de produire un modèle d'annotation qui pourra être utilisé pour des corpus similaires. Dans le cadre de Pyrrha, deux modèles ont été générés, un pour le latin et un autre pour l'ancien français[4].

2.2. Les modèles d'annotation de Pyrrha

Le premier modèle a été établi à partir d'un corpus annoté de textes latins de plus d'un million de mots produit par le LASLA (Clérice, 2019). Ce modèle propose une série de lemmes constitués à partir du dictionnaire Forcellini (Forcellini & Furlanetto, 1965) et d'un jeu d'étiquettes morphosyntaxiques créées pour les projets du LASLA et accessible via la liste de contrôle « latin LASLA » disponible sur Pyrrha.

Le deuxième modèle est un modèle pour l'ancien français (Clérice, 2019). Les lemmes ont été établis à partir du dictionnaire Tobler-Lommatzsch (Tobler & Lommatzsch, 1952) et adaptés dans les cas où les entrées n'étaient pas homogènes[5]. L'étiquetage morphosyntaxique se subdivise en deux champs *POS* pour la nature du mot et *Morph* pour indiquer : le cas, le genre, le nombre, le degré de l'adverbe ou de l'adjectif si nécessaire ou bien le mode, le temps, la personne et le nombre. Le nommage et la constitution de ces catégories s'appuient sur le référentiel *Cattex 2009* développé par l'équipe de la base de français médiévale à Lyon (Guillot, Prévost & Lavrentiev, 2013).

[2] POS (Part Of Speech).

[3] L'explication présente le fonctionnement de manière extrêmement simplifiée (cf. Manjavacas, Kádár & Kestemont, 2019).

[4] D'autres modèles pourraient être ajoutés si des chercheurs contribuent à constituer des données d'entraînement nécessaires à la constitution d'un modèle.

[5] Voir la documentation : https://github.com/Jean-Baptiste-Camps/Geste/wiki, (consulté le 28 mai 2019).

Form	Lemma	POS	Morph
conmence	comencier	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s

Table 1. Exemple d'annotation verbale

Form	Lemma	POS	Morph
signeur	seignor	NOMcom	NOMB.=s|GENRE=m|CAS=r

Table 2. Exemple d'annotation nominale

Le modèle pour l'ancien français a été entraîné sur un panel relativement large à partir de corpus constitués en grande partie à l'École nationale des chartes. Il comprend 100 000 mots issus du corpus « *Geste* » (Camps, Cochet, Ing & Albarran, 2019), 100 000 tokens[6] du *Corpus Juris Civilis* annotés par F. Duval et L. Ing, 200 000 tokens issus du corpus de Chrétien de Troyes fourni par P. Kunstmann, et alignés avec nos référentiels, 15 000 tokens du *Lancelot en prose* annotés par L. Ing et enfin 45 000 tokens issus du corpus hagiographique de Wauchier de Denain annoté par mes soins. Des corpus en occitan ont également été ajoutés comme le corpus Monferrand (30 000 tokens) annoté par J.-B. Camps, G. Couffignal et M. Mazars et le corpus Flamenca (20 000 tokens), d'après l'édition de P. Meyer et annoté par O. Scrivner. Malheureusement si, aujourd'hui l'intégralité du corpus comporte les informations au niveau du lemme et des POS, ce qui nous permet d'atteindre un taux de fiabilité de 95 %, seuls 30 % du corpus (environ 150 000 tokens) possèdent des informations morphosyntaxiques. Toutefois, nous arrivons à un taux de fiabilité avoisinant les 93 %.

Si les deux modèles proposés dans l'interface Pyrrha ne sont pas pertinents pour un projet ou demandent à être enrichis, l'utilisateur est libre d'utiliser ses propres lemmes et étiquettes de morphosyntaxe, et même de proposer ses propres modèles entraînés après demande aux administrateurs. Il est également possible d'importer et de soumettre un corpus déjà annoté au format TSV[7] comportant, associées à chaque mot les rubriques suivantes : *Form*, *Lemma*, *POS* et *Morph*.

2.3. Lemmatiser son corpus à l'aide de Pyrrha

Pyrrha offre à ses utilisateurs un accès aisé, sans ligne de commande, au lemmatiseur-annotateur Pie grâce à une interface en ligne[8] qui permet d'importer au format texte un corpus sans annotation. La seule condition préalable est de s'enregistrer avec une adresse mail pour créer un compte utilisateur qui permet de stocker son corpus et donc d'y avoir accès à chaque connexion au service.

Pour créer un nouveau corpus, il faut importer sur la page de création de corpus son texte via un simple copier-coller dans la première case de la section *data*. Grâce à la section *metadata,* on peut nommer le corpus et définir le nombre de mots à afficher avant et après le terme traité pour le contexte.

Grâce à la fonction de lemmatisation située en dessous de la case texte de la section data, on peut lemmatiser à l'aide de l'un des deux modèles le texte à

[6] Un *token* est un élément du texte obtenu suite à la tokenisation qui peut aussi bien être un mot qu'un signe de ponctuation.
[7] Format texte où les différents champs sont séparés par des tabulations.
[8] https://dh.chartes.psl.eu/pyrrha.

importer. Enfin, on peut ajouter des listes de contrôle[9] en choisissant parmi celles proposées ou bien en important la sienne[10]. Ces listes viendront aider à l'annotation en répertoriant les annotations qui dérogent aux valeurs de contrôle, mais aussi en proposant une aide à l'annotation depuis l'interface de post-correction grâce à un système d'autocomplétion. Une fois le corpus lemmatisé et les listes choisies, le corpus doit être enregistré afin de pouvoir utiliser l'interface de post-correction.

3. Correction et export des données

Même si le taux de réussite de Pie est proche des 95 % pour l'ancien français, il peut être intéressant pour le linguiste d'améliorer encore son corpus. Grâce à Pyrrha, une relecture peut être opérée via une interface de post-correction qui permet d'annoter aisément et plus rapidement un texte.

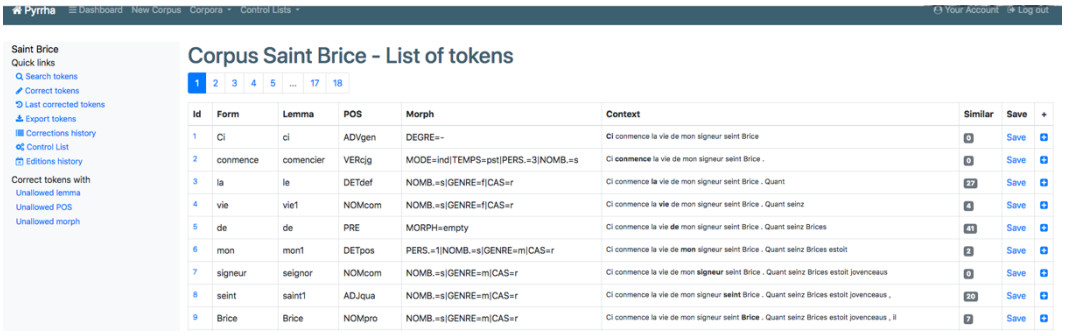

Figure 1. Interface de correction, capture d'écran
(source : Clérice, Pilla & Camps, 2018)

3.1. Fonctionnalités de base : relecture et édition des corrections

L'accès à l'interface de post-correction se fait via l'onglet *corpora* pour sélectionner le texte à traiter. L'interface affiche un tableau avec neuf catégories différentes.

1) *Id* : numéro attribué à chaque token pour l'identifier.
2) *Form* : terme tel qu'il apparaît dans le texte.
3) *Lemma* : lemme attribué à chaque terme permettant ainsi de l'associer à une forme normalisée.
4) *POS* : nature du mot.
5) *Morph* : annotation morphosyntaxique.
6) *Context* : contexte textuel.
7) *Similar* : nombre de termes dans une situation comparable.
8) *Save* : sauvegarde des modifications opérées sur l'annotation.
9) « + » : accès vers les options de modification du token : correction, suppression, ajout.

Il est possible d'intervenir directement dans les catégories *Lemma*, *POS* et *Morph* pour en changer le texte. On peut ainsi corriger la valeur d'un lemme à condition

[9] La liste standard pour l'ancien français est « Ancien Français – École des Chartes », pour le latin : « LASLA latin ».
[10] Pour créer ses propres listes de contrôle, il suffit de respecter les champs suivants : *Form*, *Lemma*, *POS* et *Morph*.

que la nouvelle valeur existe dans la liste de contrôle. L'annotation morphosyntaxique peut également être directement modifiée si les étiquettes utilisées correspondent à celles définies dans la liste de contrôle des éléments de morphosyntaxe. Une fois la modification opérée, elle peut être sauvegardée grâce au *Save* (8) présent en fin de ligne. Si l'une des catégories comporte des informations divergeant de celles des référentiels, une coloration rouge apparaît et la sauvegarde est empêchée.

Deux catégories ont été ajoutées pour simplifier les corrections. La première, *Similiar* (7) donne le nombre de cas similaires du corpus et la possibilité d'effectuer des corrections en chaîne grâce à un lien cliquable qui apparaît à la sauvegarde des modifications. La catégorie « + » (9) donne accès à trois opérations différentes : la modification d'une forme fautive, la suppression du token ou l'addition d'un nouveau token dans le cas où le texte importé serait fautif. Ces options sont à utiliser avec beaucoup de précautions au risque de ne plus pouvoir aligner le texte importé avec le texte résultant de l'annotation dans Pyrrha. Si le texte est exporté par la suite en XML-TEI, la liste des opérations effectuées sur les formes du texte sera répertoriée au début du fichier.

3.2. Fonctionnalités avancées : corrections en fonction de filtres de recherche

Pyrrha possède des fonctionnalités d'exploration du corpus accessibles via la liste de liens dans la barre de navigation de gauche.

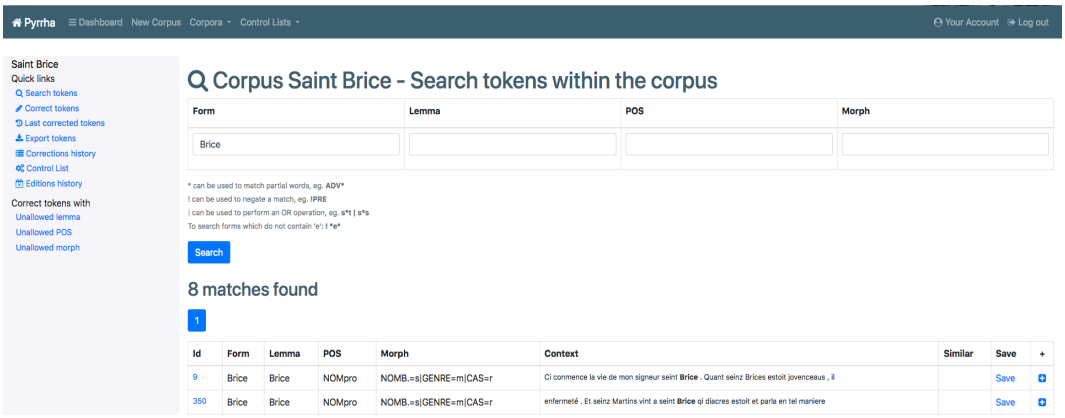

Figure 2. Interface de requêtes, capture d'écran
(source : Clérice, Pilla & Camps, 2018)

Il est possible de chercher une forme (*search token)* en particulier pour effectuer des corrections en série ou bien vérifier que l'annotation est homogène. La recherche peut porter sur le lemme, la nature et la morphologie.

On peut ne traiter que les données pour lesquelles les catégories *Lemma*, *POS* et *Morph*[11] possèdent des valeurs non valides afin d'effectuer un « nettoyage » rapide de son corpus grâce aux fonctionnalités proposées par la catégorie « *correct token with* ».

Afin d'assurer de la qualité des corrections apportées, un accès aux modifications permet à tout moment d'avoir un historique des actions effectuées grâce au lien *corrections history* avec en jaune l'ancienne version et en bleu la

[11] Dans l'ordre : Unallowed lemma, Unallowed POS et Unallowed morph.

nouvelle version. Les tokens sont classés dans l'ordre chronologique des changements, de la modification la plus ancienne à la plus récente.

Enfin, les listes de contrôle (*control list)* sont accessibles et ordonnées selon les trois catégories suivantes : *Lemma*, *POS*, *Morphologies*. Si les listes officielles sont soumises à modération avant de pouvoir être modifiées, on peut utiliser ses propres listes et les modifier en fonction de ses besoins. Les listes sont des fichiers CSV ou TSV basiques. Pour la liste des lemmes, chaque valeur est séparée par un retour à la ligne, pour les POS par une virgule. La liste des étiquettes morphosyntaxiques contient une colonne pour le label et une colonne de description de la signification du label (séparée par une tabulation). Chaque nouvelle étiquette occupe une nouvelle ligne.

3.3. Export des données

À l'issue des corrections, les données peuvent être intégralement sauvegardées et exportées dans un fichier CSV au format Pie ou en XML-TEI pour être interrogées. Le format CSV se prête aisément à des analyses statistiques, tandis que le fichier XML-TEI est davantage un fichier de conservation pour assurer la pérennité de l'annotation morphosyntaxique et à terme son partage. Le fichier TEI est très simple. Il possède un teiHeader vide qu'il conviendra de compléter. Le choix d'encodage est le suivant :

```
<w xml:id="t1" n="1" lemma="ci" type="POS=ADVgen|DEGRE=-" >Ci</w>.
```

Chaque token est englobé dans une balise <w> et toute l'annotation est contenue dans les attributs.

1) @*xml:id* pour l'identifiant du mot qui correspond à l'identifiant du mot dans l'interface Pyrrha, soit son numéro, précédé de « t ».
2) @*n* correspond à la position du mot dans le texte[12].
3) @*lemma* correspond au lemme.
4) @*type* correspond à la fois à la catégorie POS et à la catégorie Morph. Chaque subdivision de cette catégorie est séparée par « | »

4. Exploitation des données : deux cas pratiques

Les exemples proposés sont issus des Vies de saint Martin, saint Brice, saint Gilles, saint Jérôme et saint Alexis, ainsi que des *Dialogues sur les vertus de saint Martin*[13] de Wauchier de Denain. L'annotation a été générée à partir d'une lemmatisation automatique par Pie entraîné sur le modèle pour l'ancien français de l'École nationale des chartes qui était alors en cours de développement. Les corrections ont été opérées au moyen de l'interface Pyrrha. L'enjeu sera d'étudier certains

[12] Attention, la position correspond à la position du terme à l'export, elle peut être différente de celle de l'import en cas d'addition ou de suppression de tokens. Voir le modèle de données : https://github.com/hipster-philology/pyrrha/blob/dev/app/templates/tei/geste.xml. L'intégralité des modifications opérées sur les tokens est consignée dans le préambule du XML-TEI.

[13] Ces textes sont issus de mon corpus de thèse : Ariane Pinche, Édition nativement numérique du recueil hagiographique « Li Seint Confessor » de Wauchier de Denain d'après le manuscrit 412 de la Bibliothèque nationale de France, dir. C. Pierreville et B. Bureau, thèse en cours de rédaction. Les *Seint Confessor* de Wauchier de Denain sont un recueil de neuf textes composé au début de XIII^e^ siècle dans le contexte de la cour de Flandre pour Philippe de Namur (1174-1212), comte de Hainaut et de Namur.

traits de la scripta picarde, choisis parmi les traits caractéristiques décrits dans la *Petite Grammaire de l'ancien Picard* de Charles Théodore Gossen (Gossen, 1951), dont nous avons confronté les résultats avec ceux de l'étude d'A. Dees dans l'*Atlas des formes linguistiques des textes littéraires de l'ancien français* (Dees, Dekker, Huber & Van Reenen-Stein, 1987).

4.1. Réalisation de yod + ATA > pic. *-ie* (franc. *-iée*)

En Picard, la triphtongue descendante « -iee » s'est réduite en « -ie » (Gossen, 1951, p. 41). Ce trait est le trait dialectal qui connaît le moins d'exceptions dans les corpus et apparaît essentiellement dans la terminaison des participes passés au féminin. Il peut donc servir d'indicateur fiable pour déterminer l'appartenance dialectale d'un texte. En outre, ce phénomène touchant une catégorie de mots restreinte, il est aisé de déterminer les éléments à interroger. Nous avons limité notre étude au cas des participes féminins singuliers au cas sujet afin de s'assurer de l'accord.

À partir d'un export en XML-TEI de l'annotation[14], nous avons pu déterminer que le texte possède 169 participes passés au féminin employés avec l'auxiliaire être, dont 32 formes se terminant en « -ie » telles que « trenchie »[15] , tandis qu'une seule forme se termine par « -iee », à savoir la forme « sechiee ». Parmi toutes les formes en « -ie », seules les formes de participes passés du premier groupe sont touchées par le phénomène de réduction de la triphtongue, ainsi nous avons affiné la requête en ne sélectionnant que les verbes de ce groupe[16]. Sur les 32 participes passés en « -ie », tandis que 15 formes sont le résultat d'un radical en « -i » suivi du « -e » flexionnel, 17 termes présentent bien le picardisme à savoir :

> acouchie, aidie, amenuisie, apareillie, apeisie, apesie, aprochie, baillie, enforcie, esmaie, essaucie, mollie, negie, noncie, trenchie.

Ainsi, l'accord des participes féminins est respecté dans le texte et, pour les verbes concernés, dans 95 % des cas, la règle de simplification picarde de la triphtongue est effective.

4.2. Alternance graphique à l'initiale de <c> et <ch>

En ancien picard, C + e/i en position forte aboutit au son [ʃ] (Gossen, 1951 : 71) qui peut se noter à la fois <c> et <ch>. La proportion de la répartition des formes peut varier en fonction des régions (Dees *et al.*, 1987). De même, alors que C + A à l'initiale aboutit au son [k], la scripta picarde laisse parfois apparaître des variations en graphique en <ch> (Gossen, 1951 : 75-78). Nous avons donc essayé d'étudier ce phénomène complexe d'alternance sur notre corpus[17] afin de voir si

[14] Nous avons interrogé le corpus grâce à une feuille XSLT pour extraire toutes les formes de participes passés féminins au cas sujet, puis nous avons classé les formes en trois catégories différentes : les participes se terminant en « -iee », les participes se terminant en « -ie », et les autres.

[15] Liste de formes apparaissant dans le corpus : *acouchie, aemplie, aidie, amenuisie, apareillie, apeisie, apesie, aprochie, baillie, convertie, deguerpie, departie, enforcie, esmaie, essaucie, establie, florie, garantie, garie, guerie, mollie, negie, noncie, raemplie, raverdie, trenchie.*

[16] Le xpath comprend une expression régulière qui permet de trier en fonction de la terminaison du lemme en *-er* ou non.

[17] Nous avons interrogé le corpus grâce à une feuille XSLT pour comparer les graphies initiales des lemmes en *ce*, *ci*, *ca* ou *che*, *chi* ou *cha* et la graphie du terme tel qu'il apparaît dans le corpus afin de déterminer si une variation apparaissait pour certains termes et le nombre de fois où le phénomène pouvait être observé.

une région d'origine plus précise pouvait se dégager. L'étude se limite, ici, au phénomène en position initiale.

Sur les 1 432 fois où un terme possède un lemme avec C + e/i à l'initiale, seuls trois cas de graphie divergente en <ch> se distinguent : *chainte* (1) pour *ceindre*, *cheinture* (2) pour *ceinture*. Ces faibles statistiques invitent à situer le texte dans le nord de la Picardie : Nord, Hainaut, mais également en Wallonie (Dees *et al.*, 1987 : 4) ce qui est cohérent avec le corpus. Toutefois les occurrences en <ch> méritent d'être observées plus précisément et pourraient être moins marginales que les chiffres ne semblent le montrer, car toutes les occurrences des démonstratifs *ce/cel/cest* qui représentent tout de même près de 85 % des termes[18] sont graphiées en <c>, ce qui peut relever d'une habitude scribale sur un terme d'origine populaire extrêmement courant et dont la graphie se serait complètement normalisée. Autre fait intéressant, l'ensemble de ces graphies en <ch> n'apparaît que dans la *Vie de saint Alexis*. Il est difficile d'aboutir à des hypothèses fermes sur le corpus suite à cette enquête, mais cela peut nous amener à nous demander pourquoi la *Vie de saint Alexis* est différente du reste du corpus. Fait-elle bien partie du recueil ? Wauchier de Denain a-t-il été influencé par un substrat différent de celui des autres Vies pour sa translation ? Il faudrait poursuivre l'étude sur d'autres phénomènes linguistiques pour évaluer si ce phénomène est marginal ou non.

Si on regarde à l'inverse les termes qu'on attendrait avec la graphie <ch> à l'initiale, le texte ne possède aucune marque dialectale sur les 98 occurrences avec un vocalisme en « e ». L'alternance graphique en <c>/<ch> devant <i> n'apparaît que dans une unique occurrence de *cier* pour *chier*[19].

Pour C + A, la graphie en <ch> apparaît 4 fois sur 600 termes dont le lemme commence par <ca>, à savoir *chartage* (1) pour « Cartage », *Chaton* (2) pour « Caton », *chariole* (1) pour « cariole ». À l'inverse, seule une occurrence de *cascun*[20] pour « chacun » est à observer, ce qui nous éloigne des pourcentages attendus pour le terme dans le Hainaut ou le nord de la France d'après l'atlas d'A. Dees (Dees *et al.*, 1987 : 48) qui suggère au contraire une très grande domination de la graphie en <c>[21]. Enfin, le corpus présente également sept occurrences de *ceaille* pour « chaille[22] », montrant bien que la graphie <c> marque le son [ʃ].

Selon C. T. Gossen la scripta picarde présente le système suivant : c/ch + e, i = [ʃ], exceptionnellement [k], c + a = [k] et ch + a = [ʃ], exceptionnellement [k]. Cette théorie semble fonctionner sur l'ensemble de notre corpus. Toutefois cette courte étude ne nous ne permet pas encore de préciser à quel sous-ensemble de la scripta nos textes pourraient appartenir, car les marques dialectales sont trop rares pour ce phénomène et laissent penser que la langue littéraire est relativement normée et ne laisse apparaître que rarement des traits dialectaux.

[18] 1 195 occurrences.
[19] Sur 176 termes dont le lemme commence par « ci » et 33 termes dont le lemme commence par « chi ».
[20] Sur 28 occurrences du terme dans le corpus.
[21] L'observation est toutefois à nuancer par le fait que seuls trois textes ont été utilisés pour les statistiques sur ce phénomène dans l'atlas.
[22] Sur 183 termes dont le lemme commence par « cha ».

5. Conclusion

Nous espérons que ce court exposé permettra de montrer les avantages d'une annotation linguistique à travers ces quelques exemples et comment l'interface Pyrrha peut réellement faciliter la génération de l'annotation, mais aussi assurer le contrôle de la lemmatisation et des informations morphosyntaxiques pour permettre au chercheur de contrôler l'intégralité de la chaîne de génération de ses données. En outre, l'annotation, certes chronophage, mais automatisée en partie et vérifiée manuellement assure la qualité des informations et leur pérennité grâce à une sauvegarde en XML-TEI. L'étiquetage linguistique fin permet aussi de faire des analyses précises et variées sur un corpus donné qui ne relèvent plus de l'extraction d'un phénomène hors contexte, mais qui prennent en compte l'ensemble des occurrences. Enfin, la production en masse de tels corpus permettrait à terme de mettre en place des modèles de lemmatisation de plus en plus performants, mais aussi, un jour peut-être, de mettre à jour les grammaires et les atlas grâce à des études statistiques plus fiables à partir de masses de données suffisamment importantes pour que les phénomènes marginaux ne viennent plus fausser les résultats.

Bibliographie

Camps, J.-B., Cochet, A., Ing, L. & Albarran, E. (2019). *Jean-Baptiste-Camps/Geste : Geste : un corpus de chansons de geste, 2016-…*
Url : https://doi.org/10.5281/zenodo.2630574.

Clérice, T. (2019). *Deucalion Latin Lemmatizer*.
Url : https://doi.org/10.5281/zenodo.2707476.

Clérice, T., Pilla, J. & Camps, J.-B. (2018). *hipster-philology/pyrrha: 1.0.1*.
Url : https://doi.org/10.5281/zenodo.2325428.

Clérice, T. (2019). *chartes/deucalion-model-af: 0.2.0*.
Url : https://doi.org/10.5281/zenodo.3237455.

Dees, A., Dekker, M., Huber, O. & Van Reenen-Stein, K. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*, De Gruyter : Berlin (reprint 2014).

Forcellini, E., Furlanetto, G., Corradini, E. & Perin, J. (1965). *Lexicon totius latinitatis*. Gregoriana : Padoue.

Gossen, C. T. (1951). *Petite grammaire de l'ancien picard : phonétique, morphologie, syntaxe, anthologie et glossaire*. C. Klincksieck : Paris.

Guillot, C., Prévost, S., & Lavrentiev, A. (2013). *Manuel de référence du jeu Cattex09*.
Url : http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf.

Manjavacas, E., Kádár, Á. & Kestemont, M. (2019). « Improving Lemmatization of Non-Standard Languages with Joint Learning ». In : *Proceddings of NAACL-HLT 2019*, Minneapolis : Association for Computational Linguistics, p. 1493-1503.

Manjavacas, E., Kestemont, M. & Clérice, T. (2019). *emanjavacas/pie v0.2.3*.
Url : https://doi.org/10.5281/zenodo.1637878.

Mellet, S. (2002). « La lemmatisation et l'encodage grammatical permettent-ils de reconnaître l'auteur d'un texte ? ». *Médiévales*, 21(42), p. 13-26.

Tobler, A. & Lommatzsch, E. (1952). *Altfranzösisches Wörterbuch*. E. Steiner : Wiesbaden.