HAL
open science

Le PAM, un Programme d'Analyse Métrique pour le français médiéval

Enzo Poggio, Timothée Premat

To cite this version:

Enzo Poggio, Timothée Premat. Le PAM, un Programme d'Analyse Métrique pour le français médiéval. Timothée Premat; Ariane Pinche. Actes des Rencontres lyonnaises des jeunes chercheurs en linguistique historique, Diachronies contemporaines, 2019, 10.5281/zenodo.3464477 . hal-02320550

HAL Id: hal-02320550

https://hal.science/hal-02320550v1

Submitted on 18 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open licence - etalab

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

Le PAM, un Programme d'Analyse Métrique pour le français médiéval

Enzo POGGIO* & Timothée PREMAT**
*Développeur pour les SHS, chercheur indépendant.
**Université Paris 8 (SFL UMR 7023)

Résumé

Les corpus de français médiéval existant ne proposent pas d'étiquetage des propriétés prosodiques et métriques des textes. Un tel étiquetage est pourtant nécessaire, non seulement pour celui qui étudie la métrique, mais aussi pour le phonologue. Avec le PAM, nous proposons un script Python capable de segmenter un texte en syllabes et d'associer à chaque syllabe un statut de proéminence prosodique à partir duquel on doit pouvoir déduire le gabarit métrique sous-jacent. Cet article détaille certains points du fonctionnement général du programme, toujours en cours de développement.

Mots-clefs : Linguistique de corpus ; annotation métrique et prosodique ; syllabation ; français médiéval.

DOI : 10.5281/zenodo.3464477

1. Introduction

Cet article présente les principes du fonctionnement général du PAM, le Programme d'Analyse Métrique. Le PAM est un programme d'annotation prosodique qui permet la déduction de la structure métrique du texte à partir de ses propriétés prosodiques. Il prend la forme d'un script Python exécutable dans une console, et propose différentes fonctions d'export des données annotées.

Après avoir brièvement exposé le panorama des pratiques numériques dans lequel s'inscrit le PAM (section 1), nous abordons les propriétés prosodiques qu'annote le PAM (2), la façon dont celles-ci peuvent être transposées en propriétés métriques (3), et enfin les perspectives dans le développement à venir du programme (4).

1.1. Panorama de la philologie numérique historique française

Depuis le début des années 80, des corpus informatiques de textes en français médiéval ont commencé à voir le jour. Les buts de ces corpus diffèrent : certains visent principalement à l'étude de la répartition dialectale des formes graphiques (Dees, 1980), d'autres s'intéressent à l'étude de la syntaxe diachronique (c'est le cas des fondements de la *Base de Français Médiéval* (BFM), dirigée à l'origine par Ch. Marcello-Nizia et aujourd'hui par C. Guillot-Barbance) ou à la constitution de bases de données pour des dictionnaires (*Dictionnaire du Moyen Français* (DMF), *Anglo-Norman Dictionnary* (AND)). Aujourd'hui, ces corpus tendent à utiliser les normes d'encodage de la *Text Encoding Initiative* (TEI), ce qui permet une interopérabilité optimale. Ainsi, le second corpus d'A. Dees (1987) a été ré-encodé aux normes de la TEI puis versé dans le *Nouveau Corpus d'Amsterdam* (NCA) dirigé par P. Kunstmann et A. Stein. Les outils d'interrogation des corpus sont eux aussi en voie de standardisation ; ainsi la dernière version du NCA est devenue

interrogeable dans la plateforme TXM (Heiden *et al.*, 2010), qui est aussi le moteur de consultation et d'interrogation privilégié par la BFM.

Aucun de ces projets ne vise à annoter l'ensemble des propriétés d'un texte, tâche qui confinerait à l'impossible. Ainsi, l'étiquetage principalement morphosyntaxique de la BFM correspond à l'orientation syntaxique de son origine, aujourd'hui renforcée par l'étiquetage en arborescences syntaxiques du sous-corpus externe *Syntactic Reference Corpus of Medieval French* (SRCMF) dirigé par S. Prévost et A. Stein. D'autres types d'annotations ont été ajoutés à la BFM par la suite : certains textes sont également lemmatisés, et les passages au discours direct sont également annotés. La tendance générale semble être à l'annotation des propriétés du discours et des modules 'élevés' de la grammaire. Sont donc naturellement privilégiées des approches lexicologiques, syntactiques et morphosyntaxiques.

Les chercheurs qui, au contraire, travaillent sur les modules 'bas' de la grammaire, ceux qui traitent des 'formes sonores' (prosodie, métrique, phonologie et phonétique) ne disposent pas d'outils idoines pour mener leurs recherches. Ils se retrouvent contraints d'utiliser des requêtes portant sur des graphèmes, ne pouvant que marginalement s'aider des propriétés encodées. Pourtant, les *guidelines* de la TEI proposent des outils pour annoter précisément ces propriétés des 'formes sonores'. On peut par exemple aller jusqu'à encoder l'ensemble des traits phonologiques des segments d'un texte grâce aux balises de *feature structure*[1]. Cependant, à notre connaissance, un tel encodage n'est mis en place par aucun corpus historique français. L'importante variation graphique du français médiéval et l'incertitude constitutive de la phonologie diachronique rendraient de toute manière une telle approche périlleuse. Enfin, il semble que l'objet encodé soit plus souvent appréhendé comme un texte écrit que comme un texte *oralisable*, transcription d'un texte oral et/ou destiné à une communication orale. En ce sens, il est naturel que les annotations se concentrent sur des aspects qui appartiennent autant à la linguistique textuelle qu'à la linguistique en général, délaissant par-là l'étiquetage des propriétés 'sonores'[2].

1.2. Annotation métrique et syllabique : projets existants

Dans le domaine de la métrique, on retrouve la même orientation, c'est-à-dire que les propriétés macrostructurales des textes sont régulièrement annotées. Par exemple, les textes versifiés de la BFM sont encodés en strophes et en vers. Un texte, les *Chroniques concernant la guerre d'Écosse* de Jordan Fantosme (éd. 1981), pousse l'annotation jusqu'à l'étiquetage des césures. Tous les autres textes, en revanche, n'annotent aucune propriété inférieure à l'échelle du vers : ni les césures ni les syllabes ne sont annotées.

Sans même parler des besoins des chercheurs travaillant sur les formes métriques, une telle annotation serait pourtant précieuse pour le phonologue. Il en va ainsi, par exemple, de l'analyse des rimes, tâche primordiale pour la phonologie diachronique. En l'absence d'une démarcation des syllabes à l'intérieur du texte, le phonologue se retrouve contraint d'extraire le dernier mot précédant la balise de fin de vers, ce qui complique considérablement le travail des données. Et encore, une telle interrogation n'est pas toujours possible ; par exemple, sur le corpus de la BFM, il n'est par défaut pas possible d'interroger la position d'un mot

[1] https://tei-c.org/release/doc/tei-p5-doc/en/html/FS.html
[2] L'on pourrait objecter qu'il existe pourtant une prosodie de l'écrit (Martin, 2011).

dans le vers[3]. Le phonologue ne dispose pas non plus d'outils lui permettant de différencier la syllabe tonique de la syllabe posttonique dans le cas d'une rime féminine.

Une des conséquences de cette focalisation de l'encodage sur les propriétés macrostructurales est qu'elle pousse les chercheurs qui travaillent sur les propriétés microstructurales à développer leurs propres outils. Ainsi, T. Rainsford, pour sa thèse de doctorat portant sur l'évolution entre accent de mot et accent de groupe en français médiéval (Rainsford, 2011b), a choisi de développer son propre outil lui permettant, sur 87 extraits d'environ 3000 mots chacun, de repérer la position des syllabes accentuables et des frontières de groupes prosodiques et métriques ainsi que leur interaction avec la structure syntaxique. L'auteur ne conseillant pas la réutilisation de son outil, nous avons été contraints d'en développer un nouveau. Par ailleurs, son programme tend à corriger les vers pour gérer les hiatus et diphtongues, or le PAM a été conçu, à l'origine, pour repérer les vers pouvant être mal formés, dans la cadre de la thèse de doctorat du second auteur de ce papier. Un tel mécanisme, qui simplifie considérablement l'architecture du programme et améliore son entropie, n'était donc pas approprié à cette recherche. Précisons néanmoins que certains des mécanismes du PAM sont inspirés, non directement du code de T. Rainsford, mais de la logique algorithmique qu'il a mise en place.

Un autre projet de T. Rainsford mérite d'être mentionné. En collaboration avec O. Scrivner (2014), T. Rainsford travaille à l'enrichissement d'arbres syntaxiques par des informations métriques. Le corpus qui leur sert à développer cette approche est constitué d'ancien occitan, ce qui leur impose de recourir à la lemmatisation pour placer l'accent. Nous n'avons pas adopté cette approche compte tenu de la facilité avec laquelle l'accent du français médiéval peut être prédit à partir de sa forme graphique.

Pour conclure cette introduction sur les corpus et les annotations prosodiques et métriques, mentionnons que d'autres corpus, comme le corpus historique des chansons néerlandaises *Nederlandse Liederenbank*, ont préféré exprimer les propriétés métriques dans les métadonnées du texte : la structure des strophes et des vers est détaillée sous forme d'étiquettes conventionnelles hors du texte. Une telle approche permet la mise en réseau des chansons entre elles en fonction de leur forme, mais se révèle peu appropriée pour analyser l'interface entre les énoncés linguistiques et les structures métriques. Enfin, citons le projet *Anamètre* dirigé par É. Delente et R. Renault qui annote et analyse en profondeur les propriétés 'sonores' dont nous parlions ci-dessus, mais qui le fait sur un corpus moderne du XVIIe au début du XXe siècle. Le fonctionnement général des algorithmes d'Anamètre n'est pas sans similarités avec celui utilisé par T. Rainsford ou par nous-mêmes, mais leur projet se trouve facilité par la stabilité des formes graphiques aux époques qu'ils étudient. Si nous ne nous sommes pas inspirés de ce projet pour la conception du PAM, il demeure néanmoins exemplaire pour le traitement des données textuelles versifiées dans le domaine français.

[3] Une telle interrogation n'est possible qu'à condition d'appliquer au corpus une préparation spécifique, comme celle qui a été mise en place pour le projet ANR *Oriflamms*. Une telle préparation n'étant pas livrée 'clefs en main' à l'utilisateur et ne faisant l'objet d'aucune publicité, sa diffusion demeure confidentielle.

2. Quelles propriétés annoter, et comment le faire ?

Dans cette partie, nous détaillons rapidement selon quelle approche de la prosodie et de la métrique a été développé le PAM, avant d'aborder de manière plus concrète la segmentation des unités que nous avons choisi d'utiliser et leur annotation.

2.1. Cadre théorique : métrique générative

Comme indiqué dans l'introduction, une démarche de constitution et d'annotation de corpus est toujours dépendante de l'intérêt scientifique de celles et ceux qui la mettent en place. Il faut également ajouter une autre forme de dépendance : elle dépend également des options théoriques favorisées par ces acteurs. En effet, le choix d'un appareil théorique peut orienter fondamentalement le traitement des données, que ce soit au niveau de leur préparation ou de leur interrogation. Il convient donc d'expliciter le cadre théorique dans lequel s'inscrit le PAM.

Ce cadre théorique est celui de la métrique générative ou post-générative, qui sous-entend une approche phonologique des unités linguistiques mises en jeu par la versification. Dès l'article séminal de Halle & Keyser (1966), cette approche s'inscrit en opposition à l'analyse traditionnelle, parfois qualifiée de scolaire, des spécificités formelles du discours versifié. En effet, la théorie traditionnelle est une théorie *surface-oriented*, ce qui signifie qu'elle ne postule aucune échelle de représentation sous-jacente et opère d'éventuelles transformations directement en surface. Ce faisant, la théorie traditionnelle ne peut pas répondre à l'objectif d'une grammaire générative : ses transformations ne sont pas suffisamment contraintes pour ne générer que des structures attestées (elles permettent des résultats non attestés) et, étant focalisée sur le niveau de la surface phonétique, cette approche dépend des structures accentuelles effectivement mises en place par un locuteur donné et des seules propriétés que ces structures expriment phonétiquement. Cette approche ne fait donc pas de différence claire entre *rythme* et *mètre*, alors même que la distinction entre ces deux objets n'est pas une idée nouvelle (cf. Platon, 1975 : 502c) et est capitale pour la métrique générative. Enfin, une telle approche dispose d'un pouvoir explicatif faible : elle se borne le plus souvent à décrire des faits de surface sans pouvoir rendre compte de la cohérence profonde qui unit ces faits de surface dans leur variété.

Il est à noter qu'une approche *de surface* présenterait pourtant un avantage pour le traitement automatique du langage : elle permettrait de faire une réelle annotation métrique. Au lieu de cela, ainsi que nous le verrons plus bas, le PAM ne fait pas d'annotation métrique : il produit une annotation prosodique qui permet, dans un second temps, d'effectuer une analyse métrique.

La méthode de la métrique générative consiste à observer des régularités prosodiques (nombre et disposition des frontières de groupe, nombre et disposition des syllabes en fonction de leur proéminence) pour en déduire une représentation sous-jacente appelée *gabarit* ou *patron métrique*. Chaque type de vers (octosyllabe, décasyllabe, etc.) dispose de son propre gabarit. L'association entre les unités (groupes et positions) du gabarit métrique et celles d'un vers donné est régie par une grammaire d'association, qui contient des règles ou contraintes exprimant la possibilité ou l'interdiction d'associer telle unité du gabarit à telle unité du vers.

Ces règles d'association doivent, pour le français, autoriser les configurations que sont la rime féminine et la coupe épique, la coupe lyrique et la coupe

enjambante. Elles autorisent également le libre placement des accents et frontières prosodiques à l'intérieur des vers et hémistiches, les contraintes d'association ne s'appliquant qu'en fin de constituant métrique.

2.2. Structure métrique et annotation de surface

Puisque le gabarit métrique correspond à un niveau de représentation sous-jacent, il ne peut être directement annoté dans le texte, qui est toujours un texte de surface. Le gabarit n'est pas nécessairement réalisé tel quel, les règles d'association permettant d'associer des structures prosodiques qui peuvent être en léger décalage avec la structure métrique. Étant par essence *absent* du texte, le gabarit métrique n'est immédiatement disponible ni pour le lecteur/auditeur, ni pour le chercheur, ni pour l'annotateur, que celui-ci soit humain ou machine.

C'est par une opération mentale que l'humain peut restituer le gabarit métrique qui sous-tend un ensemble de vers donnés. On suppose que les gabarits et les règles d'association sont acquis, sur le modèle de l'acquisition phonologique. Une fois cette acquisition réalisée, un sujet peut reconnaître qu'un énoncé est une instanciation bien-formée d'un gabarit métrique donné, lorsque le contexte sociolinguistique l'induit à faire cette analyse. D'autres approches, comme le travail de B. de Cornulier, insistent sur la comparaison contextuelle : pour lui, la structure métrique d'un vers ne peut être détectée que par la mise en relation avec la structure des autres vers qui l'entourent (de Cornulier, 1995 : 21). En comparant un vers avec les vers qui précèdent et suivent (ou qui se trouvent à la même position de la strophe dans le cas des strophes hétérométriques), le sujet isole les propriétés récurrentes et considère que ces propriétés récurrentes sont la réalisation des propriétés structurales du vers.

Cette option, qui s'inscrit dans le travail de B. de Cornulier dans une négation de l'existence autonome de gabarits métriques abstraits, ne correspond pas à notre ancrage théorique. Elle est en revanche très adaptée à une approche informatique et automatique des textes : il suffit d'annoter toutes les propriétés prosodiques des vers, de repérer celles qui sont récurrentes, et de considérer que ces propriétés récurrentes sont des propriétés métriques. Ensuite, le chercheur est libre de considérer que ces propriétés traduisent l'existence d'un gabarit métrique sous-jacent ou que, suivant B. de Cornulier, elles fondent le caractère métrique des vers dans une répétition de surface de certaines propriétés prosodiques.

Il est cependant à noter que, même dans l'approche de B. de Cornulier, il faut prendre en compte une certaine variation de surface comme n'étant pas structurale. Ainsi, par exemple, l'alternance entre vers masculins et féminins n'empêche pas de leur reconnaître une structure commune, dans la mesure où ce qu'il faut considérer, ce n'est pas le nombre de syllabes du vers, mais le nombre de ses syllabes *anatoniques* (i.e. situées avant ou sous l'accent de fin de vers) (Cornulier, 2010). Le même genre de mécanisme doit être mis en place pour l'alternance des coupes possibles en français médiéval. Ceci ne représente pas une difficulté majeure, dans la mesure où les principes de ces variations de surface sont bien connus pour le vers français, qu'ils soient exprimés dans l'anatomie du vers ou sous la forme des règles d'association de la métrique générative.

2.3. Unités prosodiques à annoter et étiquettes

C'est donc par l'annotation des unités prosodiques que le programme peut donner accès à la restitution des propriétés structurales et/ou gabaritiques des vers. Ces unités prosodiques sont les syllabes et les syntagmes phonologiques.

La première tâche du PAM est donc la segmentation du texte en syllabes. Cette segmentation est opérée par le module de fonction utilitaire `syllabise.py`, qui utilise la librairie NLTK (*Natural Language ToolKit* : Bird, Klein & Loper, 2009) pour le découpage en mots. Ce module est par ailleurs disponible de manière autonome sur GitHub. Le syllabeur a été pensé de manière à respecter au mieux la coupe syllabique (jonction des attaques branchantes et disjonction des groupes hétérosyllabiques, etc.), même si la syllabation d'un texte écrit est une tâche confinant à l'absurde, dans la mesure où le syllabeur traite des graphèmes alors que la syllabation humaine traite des phonèmes et que la correspondance graphèmes-phonème est loin d'être simplement bijective. Travaillant sur une base graphémique, le syllabeur respecte les élisions qui sont graphiquement notées (<qu'alors> → <qua.lors>), mais ignore celles qui ne sont pas notées (<amie et> → <a.mi.e.et>). En effet, l'élision du schwa dépend de paramètres suprasegmentaux qui ne peuvent pas être considérés à l'échelle strictement locale et interne au mot à laquelle opère le syllabeur. De plus, ces paramètres suprasegmentaux peuvent s'appliquer d'une manière différente en prose (écrite et/ou orale) et en vers, dans la mesure où la grammaire d'association est une paraphonologie (Kiparksy, 1977) apte à faire adopter à la phonologie des vers des comportements qui ne sont pas attestés en prose (orale et/ou écrite). La syllabation que propose `syllabise.py` est donc une syllabation sous-jacente, située *avant* l'application de l'élision des posttoniques.

Une fois cette syllabation effectuée, l'output du syllabeur sert d'input à l'annotateur du PAM. Celui-ci va associer une étiquette à chaque syllabe, qui représente son statut prosodique (dans l'optique de la paraphonologie du vers). Ces étiquettes sont les suivantes :

(1) Étiquettes prosodiques du PAM

a. **-1.** Est étiquetée « -1 » toute syllabe qui est non métrifiée (syllabes apocopées et rimes féminines) ;

b. **0.** Est étiquetée « 0 » toute syllabe métrifiée qui est posttonique ou qui appartient à un clitique :

pour, *-e* de *cunte* devant initiale consonantique, etc.,

c. **1.** Est étiquetée « 1 » toute syllabe métrifiée prétonique :

plai- dans *plaisir*, *par-* dans *parler*, etc.,

d. **2.** Est étiquetée « 2 » toute syllabe métrifiée qui peut porter un accent :

-sir dans *plaisir*, *-ler* dans *parler*, etc.

L'étape (1a) correspond à l'élision des schwas posttoniques présents dans la forme graphique. Elle est simplement calculée sur le principe suivant : chaque fois qu'un mot se termine par un <e> et n'appartient pas à une liste d'exception, celui-ci est étiqueté « -1 » si le mot suivant commence par une voyelle. Le traitement du statut disjonctif du <h> initial (qui n'autorise l'élision que lorsqu'il est dépourvu de représentation phonologique, ce qui n'est pas toujours le cas) est

provisoirement géré par l'édition de deux types de <h> dans le texte d'entrée. Ce processus sera amélioré ultérieurement.

Dans le terminal, la sortie de cette annotation (accompagnée par l'analyse métrique) prend l'apparence de la figure 1, ici sur les quatre premiers vers du *Chevalier au Lion* (Chrétien de Troyes, éd. 2009).

```
1 Analyse de ../../Desktop/YvainKu.txt:
1   2 |0 |  2  | 2  |0 | 1   2   -1  m:8[8]
ar·tus li boens rois dë bre·tain·gne num_l:1
0 | 0 | 1  2  0 | 0 |1   2  -1       m:8[8]
la cui pro·es·cë nos en·sei·gne      num_l:2
 0 | 0 |1   2  | 2  |0 | 1   2       m:8[8]
quë nos so·ιens preu et cor·tois     num_l:3
 2  | 2  |0 |2   0 |1  0 | 2         m:8[8]
tint cort si ri·chë co·më rois       num_l:4

1 ../../Desktop/YvainKu.txt  m:8
100.00% soit  4  /4 vers bien formés
00.00% soit  0  /4 vers mal formés
```

Figure 1. Exemple de sortie du PAM dans un terminal.

2.4. Insuffisance des formes graphiques

Le PAM étant conçu pour pouvoir fonctionner sur des corpus peu ou pas annotés en syntaxe et en partie du discours, il travaille essentiellement sur les graphèmes. Néanmoins, une telle approche est insuffisante : il faut par exemple que le PAM soit capable de différencier la finale *-ent* des troisièmes personnes du pluriel de la finale *-ment* adverbiale (← -MENTE). Dans la mesure où la finale verbale *-ent* peut être précédée d'une base se terminant par *-m-* (p. ex. *aiment*), une approche purement graphémique ne peut différencier ces deux morphèmes. Dans ce cas, le PAM utilise une liste de tous les adverbes attestés dans la partie annotée en parties du discours de la BFM : chaque fois qu'il rencontre une finale *–ment*, l'annotateur vérifie si l'ensemble de la forme graphique est attesté dans la liste des adverbes ; si oui il annote « 2 » la syllabe *-ment*, sinon il l'annote « 0 » et place un « 2 » sur la syllabe précédente. D'autres mots, en quantité négligeable, sont gérés dans une liste d'exception (p. ex. *argent*).

Dans le cas des adverbes et des finales verbales en *-ent*, la vérification manuelle des résultats du PAM n'a montré l'existence d'aucune paire homographique à même de provoquer l'identification d'un adverbe à la place d'une P6 verbale, ou vice-versa. En revanche, certaines formes graphiques d'adverbes ne sont pas représentées dans le corpus annoté en partie du discours, ce qui provoque leur identification en tant que P6 verbales. Il sera essentiel, lorsque le PAM sera à une étape plus avancée de son développement, de calculer la marge d'erreur générée par ce genre de résultats, de manière à pondérer les résultats des analyses statistiques. Notons que, lorsque l'adverbe mal identifié est situé en fin de vers, il provoque automatiquement une erreur métrique (le vers est détecté hypométrique, puisque sa phase anatonique s'arrête une syllabe avant là où est vraiment l'accent), ce qui permet de repérer très facilement ces erreurs et d'ajouter ces adverbes à la liste d'exceptions.

En revanche, la détection des mots qui, pour des critères grammaticaux, ne peuvent pas porter d'accent provoque un nombre plus élevé de mauvaises

analyses. En effet, le nombre d'homographes est ici beaucoup plus grand, notamment à cause de la très forte variation graphique médiévale. Ainsi <deus> peut être une graphie pour *deux* comme pour *Dieu*. Par ailleurs, si certains pronoms changent de forme lorsque, postposés, ils deviennent toniques (*me donne* ~ *donne-moi*), d'autres n'en changent pas (*nous donne* ~ *donne-nous*). En ce cas, en l'absence d'arborescence syntaxique, il n'est pas possible de savoir si une forme <nous> est atone ou tonique.

Le traitement de ces confusions peut s'opérer de deux manières, suivant le type de texte : sur les textes non annotés en parties du discours, une étiquette spéciale est prévue, qui permettra à l'analyste de repérer tous les vers contenant des occurrences incertaines soit pour les annoter lui-même, soit pour les exclure tout simplement de son corpus et ne pas biaiser ses analyses quantitatives. Sur les textes annotés en partie du discours, le PAM pourra, pour chaque occurrence douteuse, interroger l'étiquette du XML originel, et établir son annotation prosodique à partir de cette information. Ces deux fonctionnalités ne sont pas encore présentes dans le PAM au moment où nous écrivons ces lignes, mais le seront dans des versions ultérieures du programme.

2.5. Gestion des paramètres linguistiques : les configs

Comme toute langue naturelle, la langue médiévale française est sujette aux variations habituelles de la sociolinguistique. Parmi celles-ci, deux sont pertinentes pour le PAM : la variation diachronique et la variation diatopique. En effet, le rapport entre graphèmes et phonèmes peut varier suivant les dialectes ou suivant le moment de composition du texte. Il en est ainsi, par exemple, de la graphie <-ni-> devant voyelle qui, en wallon, sert à noter l'équivalent d'un <gn> dans les autres dialectes. Si le PAM n'est pas informé de cette différence, il va syllaber le mot <companie> (« compagne ») en quatre syllabes (<com.pa.ni.e>) alors que la représentation phonologique de ce mot n'en contient que trois, le <i> formant avec le <n> un digramme consonantique (<com.pa.nie>).

Les paramètres fondamentaux de syllabation sont présents dans un fichier `constants.yaml`, accompagné par les fichiers qui servent à gérer les exceptions et les adverbes (`special_syll.yaml` et `special_type.yaml`). Ces trois fichiers sont regroupés dans une configuration (*config*) générale. Lorsqu'il fait face à un texte dont la situation diachronique ou diatopique modifie les paramètres de syllabation, l'utilisateur peut dupliquer ces trois fichiers dans un autre répertoire et les modifier à sa guise, ce qui crée une nouvelle config, spécifique à un texte ou à un ensemble de textes. Lorsqu'il exécute le PAM, si l'utilisateur ne fournit aucune information relative à la config, c'est la config par défaut qui est exécutée ; s'il veut utiliser une config spécifique, il l'indique en ajoutant `-C [chemin de la config]` à la ligne de commande.

3. De la prosodie à la métrique

L'annotation syllabique pratiquée en 2.3 et modifiée telle qu'en 2.4 permet d'établir partiellement la structure prosodique du texte, puisqu'elle permet de repérer ce qui correspond aux groupes clitiques. Ceux-ci s'étendent d'une étiquette 2 (et l'éventuel -1 qui le suit) jusqu'à la précédente étiquette 2 (et l'éventuel -1 qui le suit). Les groupes clitiques ne sont cependant pas le constituant prosodique le plus intéressant pour l'analyse métrique. Mais, en l'absence de consensus sur l'extension du syntagme phonologique en français, et face à la

difficulté de reconstruire la structure prosodique d'une langue morte sans enregistrements audio, nous suivons une démarche analogue à celle de T. Rainsford (2011a, 270-272), en analysant de *possibles syntagmes phonologiques*. Ceux-ci sont constitués au minimum d'un seul groupe clitique[4] et au maximum de plusieurs groupes clitiques, la tête du syntagme phonologique étant toujours celle du dernier groupe clitique.

Notons par ailleurs que les règles d'associations étant capables d'invoquer une paraphonologie à même de contrecarrer la phonologie du langage naturel, il n'est pas nécessaire de s'interroger précisément sur l'extension du syntagme phonologique médiéval : on considère qu'une fin de constituant métrique (hémistiche ou vers) doit être alignée avec une fin de syntagme phonologique, mais sans avancer d'hypothèse sur le fait que l'ensemble du constituant corresponde à un seul syntagme phonologique. Dans notre analyse, à laquelle répond le fonctionnement du PAM, ce sont les frontières des constituants et non l'extension de ceux-ci qui sont ciblées par la grammaire d'association.

De fait, l'annotation du PAM propose des syllabes toujours atones (« -1, 0, 1 ») et des syllabes qui, si tel syntagme phonologique est réalisé par le locuteur, doivent en porter l'accent (« 2 »). Mais, si le locuteur choisit de réaliser un syntagme phonologique plus grand et qu'il y a un autre groupe clitique sur la droite du groupe clitique considéré, alors la tête du premier groupe clitique n'est pas une tête de syntagme phonologique et ne porte pas d'accent. Cela ne revient pas non plus à faire l'hypothèse que les têtes de syntagmes phonologiques associées avec des têtes de constituant métrique sont nécessairement accentuées lors de la diction du vers, la réalisation (surface) du vers pouvant être plus ou moins éloignée de sa structure métrique (profonde) (cf. Verluyten, 1989 : 51-53).

En revanche, selon l'approche de B. de Cornulier exposée en 2.2 et implémentée dans le fonctionnement du PAM, les seules têtes de groupes clitiques à être associées à des têtes de constituant métrique sont celles qui sont régulièrement disponibles de vers en vers. Pour la tête du vers, cela est assez simple à repérer. Pour la tête de l'hémistiche, il faut considérer la possibilité que la tête soit décalée d'une position vers la gauche (coupe lyrique). Il faut aussi considérer que la posttonique d'un premier hémistiche féminin peut être extramétrique (coupe épique) ou métrifiée dans l'hémistiche suivant (coupe enjambante). Ces paramètres sont gérés par le PAM qui propose, pour chaque ligne et en fonction du mètre de référence indiqué par l'utilisateur, différentes coupes possibles, et qui réunit ces possibilités dans les statistiques globales de l'extrait analysé. (2) représente les coupes analysées sur les trois premiers vers de *Li Ver del juïse* (Anonyme, 1982).

(2) Exemple de coupes analysées par le PAM

Sanior oiez raison glorïose et saintisme	6ma
Del ciel en est la voiz de paradis la vie	6ma
Deus la tremist en terre por amendeir noz vies	6épC

En (2), les deux premiers vers ont pour seule coupe possible une coupe masculine après la sixième position (6ma). En revanche, dans le troisième vers, la seule césure possible est une césure épique devant initiale consonantique (6épC). Les césures épiques devant initiale consonantique sont aisées à repérer, dans la

[4] Nous considérons que, dans le cas où un mot phonologique n'est pas précédé de clitiques, il forme à lui seul un groupe clitique de manière à respecter le caractère strict de la hiérarchie prosodique.

mesure où, avant l'analyse des césures, elles provoquent un vers extramétrique (si l'on métrifie le *-e* de *terre*, alors le vers fait 13 syllabes métriques). Devant initiale vocalique, le PAM propose toujours deux analyses : une césure épique devant voyelle ou une césure masculine par apocope de la posttonique. Par exemple, sur le vers (3) (chanson homonyme, dans l'édition de Rosenberg, Switten & le Vot, 1998), le PAM proposerait les deux coupes suivantes :

(3) Bele Doette as fenestres se siet 4ma/4épV/6na

Des configurations comme (3) sont traditionnellement analysées comme étant des cas d'apocope de la posttonique aboutissant à des césures masculines. C'est notamment l'opération couramment utilisée par les philologues qui cherchent à corriger toutes les coupes épiques résiduelles dans le corpus des trouvères. Mais, pour cette chanson comme pour quelques autres chansons, la notation musicale accompagnant ce vers semble indiquer que le *-e* de *Doette* est bien chanté (cf. Premat, 2017), ce qui n'invalide pas tout à fait l'hypothèse traditionnelle (mais demanderait alors à ce que la paraphonologie du vers considère comme apocopée une syllabe effectivement prononcée), mais permet également de soutenir qu'en fin de constituant métrique, une posttonique peut être maintenue même devant initiale vocalique. Le fait de proposer les deux analyses donne au PAM une neutralité, celui-ci laissant à l'analyste le dernier mot. Enfin, le PAM détecte également les césures lyriques et enjambantes, et indique aussi les coupes qui, compte tenu du mètre de référence indiqué par l'utilisateur, ne sont pas acceptables, comme à la sixième position en (3) (6na, pour *not applicable*), qui tomberait entre la prétonique et la tonique de *fenestres*.

4. Perspectives

4.1. Export : la perspective du XML

Au moment où nous écrivons ces lignes, le PAM produit son analyse dans la console où l'utilisateur l'exécute. Ayant produit l'annotation ligne par ligne, il produit également un résumé donnant accès aux statistiques générales de l'extrait interrogé : combien de vers bien formés vis-à-vis du mètre de référence indiqué par l'utilisateur, combien de vers mal formés, quantité des différents types et positions de coupes. Le PAM propose également un export en table .csv ou .xlsx, permettant à l'utilisateur d'appliquer des analyses statistiques ou simplement de sauvegarder l'annotation effectuée.

Ces exports ne permettent ni l'enrichissement des analyses du PAM par les informations supplémentaires éventuellement présentes dans le fichier d'entrée (tel que l'étiquetage en parties du discours) ni leur réutilisation ou leur croisement avec ces autres données. Le format de sortie du PAM, par ailleurs, ne répond pas aux normes de standardisation des corpus actuellement en vigueur. L'objectif, qui sera poursuivi dans les années à venir, est donc de réinjecter les informations du PAM dans les fichiers XML d'origine, en suivant les consignes de la TEI-P5. Ainsi, si les responsables des corpus en question s'y accordent, ces informations seront à terme disponibles dans les fichiers XML et pourront être interrogées dans des plateformes comme TXM, tant dans une perspective d'analyse métrique que phonologique. Enfin, comme indiqué plus haut, cela permettrait, sur les textes annotés en partie du discours, d'affiner l'annotation du PAM en cas d'homographie.

4.2. Accessibilité : sortir du terminal

Enfin, il est à considérer que l'installation et l'exécution du PAM dans la console de l'ordinateur réduit son accessibilité auprès des membres de la communauté scientifique qui ne seraient pas habitués à l'usage du terminal et/ou de commandes Python. Nous envisageons en ce sens la possibilité d'une interface HTML qui libérerait l'utilisateur du terminal et qui, si elle était disponible sur serveur, le dispenserait d'avoir à télécharger et installer le programme sur sa machine.

Conclusion

L'absence d'annotation phonologique, prosodique et métrique pousse les chercheurs travaillant sur ces domaines soit à travailler sans outils informatiques idoines soit à développer les leurs. Dans cette optique, le PAM est un Programme d'Analyse Métrique qui, après avoir effectué une syllabation et une annotation prosodique du texte, propose à l'utilisateur des analyses métriques. Il travaille sur une base essentiellement graphémique, augmentée de quelques dictionnaires de formes, de manière à pouvoir fonctionner sur des textes peu ou pas annotés. Toujours en développement, le projet vise à améliorer la précision des résultats fournis, à les réinjecter dans les fichiers XML-TEI d'origine pour l'amélioration de cette précision et pour diffuser l'information annotée auprès de la communauté, et éventuellement à sortir du terminal pour améliorer son accessibilité.

Bibliographie

Anonyme (1982). *Li ver del juïse* (éd. E. Rankka). Uppsala : Almqvist och Wiksell. Publié en ligne par la Base de français médiéval (dernière révision le 10/08/2010).
Url : http://catalog.bfm-corpus.org/Juise.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. Sebastopol, Beijing, Cambridge : O'reilly.

Chrétien de Troyes (2009). *Chevalier au Lion ou Yvain* (éd. P. Kunstamm). Ottawa, Nancy : Laboratoire de Français Ancien, ATILF.
Publié en ligne par la Base de français médiéval (dernière révision le 06/05/2009).
Url : http://catalog.bfm-corpus.org/YvainKu.

De Cornulier, B. (1995). *Art Poëtique, Notions et problèmes de métrique*. Lyon : Presses universitaires de Lyon.

De Cornulier, B. (2012). « Si le mètre m'était compté... Sur la notion fallacieuse de mesure du vers ». In L. de Saussure, A. Borillo & M. Vuillaume (dir.), *Grammaire, Lexique, Référence, Regards sur le sens, Mélanges offerts à Georges Kleiber pour ses quarante ans de carrière*, Bern : Peter Lang, p. 355-376.

Dees, A. (1980). *Atlas des formes et des constructions des chartes françaises du 13e siècle*. Tübingen : M. Neimeyer Verlag.

Dees, A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français*. Tübingen : M. Neimeyer Verlag.

Fantosme, J. (1981). *Chroniques concernant la guerre d'Écosse* (éd. R. C. Johnston). Oxford : Clarendon.
Publié en ligne par la Base de français médiéval (dernière révision 23/05/2009).
Url : http://catalog.bfm-corpus.org/Fantosme.

Halle, M. & Keyser, S. J. (1966). « Chaucer and the Study of Prosody ». *College English*, 28(3), p. 187–219.

Heiden, S., Magué, J.-P. & Pincemin, B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement ». In S. Bolasco, I. Chiari & L. Giuliano (dir.), *10th International Conference on the Statistical Analysis of Textual Data – JADT 2010*, Rome : Edizioni Universitarie di Lettere Economia Diritto, p. 1021-1032.

Kiparksy, P. (1977). « The Rhythmic Structure of English Verse ». *Linguistic Inquiry*, 8(2), p. 189–247.

Martin, P. (2011). « Ponctuation et structure prosodique ». *Langue française*, 172, p. 99–114.

Platon (1975). *Lysis Symposium Gorgias* (éd. W. R. M. Lamb). London : W. Heinemann.

Premat, T. (2017). *Conflits et synergies des patrons structuraux dans la musique des trouvères*, mémoire de master.
Url : https://scd-resnum.univ-lyon3.fr/out/memoires/langues/2017_premat_t.pdf.

Rainsford, T. (2011a). « Dividing Lines: The changing syntax and prosody of mid-line Break in medieval France octosyllabic verse ». *Transactions of the Philological Society*, 109, p. 265–283.

Rainsford, T. (2011b). *The Emergence of Group Stress in Medieval French*. Thèse de doctorat, Cambridge University, St. John's College.

Rainsford, T. & Scrivner, O. (2014). « Metrical Annotation for a verse treebank ». In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova & A. Przepiórkowski (dir.), *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, Tübingen : University of Tübingen, p. 149–159.

Rosenberg, S. N., Switten, M. & le Vot, G. (1998). *Songs of the Troubadours and Trouveres: An anthology of poems and melodies*. New York & London : Garland.

TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, version 3.6.0, modifiée le 16/07/2019.
Url : http://www.tei-c.org/Guidelines/P5/.

Verluyten, S. P. (1989). « L'analyse de l'alexandrin, mètre ou rythme ? ». In M. Dominicy (dir.), *Le souci des apparences : Neuf études de poétique et de métrique*, Bruxelles : Éditions de l'Université de Bruxelles, p. 31-74.