Old Gallo-Romance (OGR) Corpus : annotation phonologique et métrique des plus anciens textes gallo-romans

Thomas Rainsford*
Institut für Linguistik/Romanistik, Universität Stuttgart, Keplerstr. 17, 70174 Stuttgart, Allemagne.

Résumé : L'objectif du *Old Gallo-Romance Corpus* (corpus OGR) est de réunir tous les textes gallo-romans copiés avant 1130 dans une forme le plus fidèle possible au manuscrit de base et munis d'une annotation approfondie. En particulier, le corpus dispose d'une couche d'annotation phonologique et d'une couche d'annotation métrique. Dans cet article, nous présentons les innovations principales implémentées dans le corpus OGR, surtout le développement de la couche d'annotation phonologique et la création d'une infrastructure technique qui facilite la création de cette annotation et permet de l'exporter dans un format XML.

Abstract :*Old Gallo-Romance (OGR) corpus*: phonological and metrical annotation of the oldest Gallo-Romance texts. The goal of the *Old Gallo-Romance (OGR) Corpus* is to unite in a single corpus all Gallo-Romance texts copied before 1130 in a form as faithful as possible to the base manuscript and annotated in depth. In particular, the corpus contains both phonological and metrical layers of annotation. In this article, we present the main innovations implemented in the OGR corpus, with special focus on the creation of a technical infrastructure which assists in the creation of this annotation and exports it in an XML format.

1 Introduction

Les textes gallo-romans antérieurs au 12e siècle ont toujours occupé une position privilégiée dans la recherche linguistique et philologique. Parmi les plus anciens témoins de toutes les langues romanes, les « plus anciens monuments » du français et de l'occitan contiennent des traits linguistiques qui disparaissent avant l'essor des traditions textuelles, et pour cela ils sont incontournables pour toute recherche sur le développement des langues romanes.

Cependant, malgré une longue histoire d'études philologiques détaillées, ces textes ne sont réunis dans aucune base électronique adaptée à leur exploitation. Tout d'abord, la séparation de la philologie française et de la philologie occitane en deux disciplines — d'ailleurs regrettable pour une époque où les données dont nous disposons sont

* Corresponding author : thomas.rainsford@ling.uni-stuttgart.de

extrêmement limitées — est typiquement suivie par les compilateurs des corpus électroniques, et cela malgré le fait que plusieurs des plus anciens documents nous sont parvenus dans une forme qui mélange les deux langues. En effet, les plus anciens textes sont presque tous énigmatiques en tant que témoins uniques de trois siècles d'évolution linguistique mal documentés et paraissent de ce fait exceptionnels dans les bases où la majorité des textes datent du 13e siècle ou des siècles ultérieurs. Par ailleurs, au niveau technique, ils nécessitent un traitement à part : on n'obtient pas un résultat très satisfaisant si l'on essaie d'annoter un texte du 10e siècle qui mélange des formes françaises et latines avec un parser ou un lemmatiseur entraîné sur les données du 13e siècle. Enfin, il existe un décalage entre l'approche philologique des plus anciens textes, où l'édition est typiquement accompagnée d'un glossaire complet, une traduction, des notes critiques et une étude détaillée de la langue, et la pratique des bases électroniques, qui numérisent l'édition du texte mais suppriment l'apparat critique.

La création du *Old Gallo-Romance Corpus* (corpus OGR, Rainsford 2022) cherche à combler cette lacune et à compléter les ressources existantes. L'objectif du corpus OGR est de réunir tous les textes gallo-romans copiés avant 1130 dans une forme la plus fidèle possible au manuscrit de base et avec une annotation approfondie. En particulier, le corpus dispose d'une couche d'annotation phonologique et d'une couche d'annotation métrique, car l'extrême variabilité de la graphie des plus anciens textes fournit des données particulièrement précieuses pour la phonologie historique. Bien que le corpus ne soit pas encore complètement achevé (voir section 2), dans l'état actuel il contient déjà la majorité des données, car les textes qui restent à annoter sont plutôt courts. La version actuelle du corpus (v0.4) est publiée en libre accès avec une documentation détaillée de l'annotation sur le site http://www.ogr-corpus.org. Deux versions du corpus sont proposées : un portail en ligne qui s'appuie sur la plateforme ANNIS 3.6 (Krause et Zeldes 2016), et une version téléchargeable pour la plateforme TXM (Heiden 2010), qui se base sur l'export XML-TEI du corpus.

Dans cet article, nous présentons les innovations principales implémentées dans le corpus OGR. Nous discutons d'abord les enjeux philologiques soulevés par les plus anciens textes (§2) avant de traiter les schémas d'annotation développés pour le corpus (§3), avec une focalisation sur le schéma d'annotation phonologique. Dans la section 4, nous présentons la chaîne de traitement utilisée pour la création de l'annotation phonologique avant de donner quelques exemples des recherches rendues possibles par le corpus (§5).

2 Approche philologique

L'approche philologique du corpus OGR est conforme à ses objectifs scientifiques, c'est-à-dire la facilitation de la recherche linguistique sur la phonologie et la morphologie des premières étapes documentées du français et de l'occitan. La sélection de textes et les principes d'édition sont ainsi caractérisés par le primat du manuscrit. Cela se manifeste d'une part par le choix d'inclure seulement les textes préservés dans un manuscrit copié avant 1130 sans considération de la date de composition du texte lui-même, et d'autre part par l'établissement d'un texte de base fidèle au manuscrit, sans ajouts éditoriaux.

Tableau 1 : Textes gallo-romans copiés avant 1130. * = texte avec plus de mille mots tokenisés, fro = ancien français, pro = ancien occitan, frp = franco-provençal, lat = latin.

Sigle	**Titre**	**Langue**	**Date du manuscrit**	**OGR**
Alba	Aube bilingue de Fleury	pro, lat	*ca* 1000	en préparation
Alexis*	Vie de saint Alexis	fro	*ca* 1130	v0.4
AlexAlb	Fragment octosyllabique du Roman	frp	1er quart 12e s.	en préparation

Sigle	Titre	Langue	Date du manuscrit	OGR
	d'Alexandre			
AlexisProl	Prologue à la Vie de saint Alexis	fro	*ca* 1130	v0.4
BenClerm	Formules de bénédiction de Clermont	pro	10e s.	en préparation
Boeci*	Fragment d'une paraphrase de Boèce	pro	début 12e s.	v0.4
CantQuSol	Paraphrase de la Cantique des cantiques	fro	après 1100	en préparation
ChansLas	Deux strophes d'une chanson d'amour française	fro	3e tiers 11e s.	prévu
EpreuveJudic	Cérémonial d'une épreuve judiciaire	fro	début 12e s.	prévu
EpSEt	Épître farcie de saint Étienne	fro	*ca* 1130	prévu
Eulalie	Séquence de sainte Eulalie	fro	fin 9e s.	v0.4
Jonas	Fragment de Valenciennes (Sermon sur Jonas)	fro, lat	1re moitié 10e s.	en préparation
PassAug	Passion d'Augsbourg	fro/pro	10e s.	en préparation
Passion*	Passion du Christ	fro	*ca* 1000	v0.4
PrDieu	Prière à Dieu en vers : *Be deu hoi mais…*	pro, lat	avant 1100	v0.4
PrVierge2	Prière à la Vierge en vers : *Mei amic e mei fiel...*	pro, lat	avant 1100	v0.4
PrVierge3	Prière à la Vierge en vers : *O Maria, Deu maire…*	pro, lat	avant 1100	v0.4
Serments	Serments de Strasbourg	fro, lat	*ca* 1000	v0.4
SFoi*	Chanson de sainte Foi	pro	fin 11e..début 12e s.	v0.4
SGregPaint	Discours de saint Grégoire sur les images	fro	*ca* 1130	v0.4
SLeger*	Vie de saint Léger	fro	*ca* 1000	v0.4
Sponsus	Sponsus	pro, lat	avant 1100	v0.4

Le Tableau 1 montre les 22 textes dans 16 manuscrits qui répondent au critère de sélection[i], même si, puisque la datation des manuscrits est approximative, il est impossible de savoir si certains manuscrits de la première moitié du 12e s. sont vraiment antérieurs ou postérieurs à cette date[ii]. Cependant, une date limite fixe située dans la première moitié du 12e s. donne un *terminus ad quem* clair et relativement précoce pour tous les développements phonétiques et morphologiques attestés dans le corpus et évite les problèmes posés par un possible rajeunissement de la graphie dans une copie tardive. Parmi ces documents, seuls cinq comptent plus de mille mots tokenisés (signalés par un astérisque dans le Tableau 1) : un texte français (*Alexis*), deux textes occitans (*Boeci*, *SFoi*) et deux textes « français » qui montrent néanmoins de fortes influences occitanes (*Passion*, *SLeger*). L'attribution de ces derniers au domaine d'oïl repose sur la distinction traditionnelle de la langue de l'auteur (français) de celle du copiste (occitan), mais c'est une solution insatisfaisante : comme De Poerck (1963 : 21) l'a signalé il y a soixante ans, il s'agit de « textes […] nés à des carrefours culturels » à une époque antérieure au développement de traditions littéraires et scripturales distinctes dans les deux domaines. Dans un corpus d'ancien français, les traits occitans présents dans ces textes les rendent aberrants, mais dans un corpus d'ancien gallo-roman, la présence de textes occitans permet

une évaluation plus nuancée de ces traits linguistiques, même pour des études qui se focalisent uniquement sur le français. Notons aussi qu'ils sont loin d'être les seuls textes difficiles à localiser dans l'un des deux domaines, car la provenance de *Spons*, *AlexAlb* (peut-être franco-provençal), et *PassAugs* est également difficile à fixer. Ceci nous parait un argument puissant en faveur d'un corpus qui englobe les deux variétés présentes à cette époque précoce du développement de l'écriture vernaculaire.

En ce qui concerne l'édition des textes, le corpus OGR suit « le principe d'une fidélité maximale au manuscrit de base » adopté par l'équipe de la *Base de français médiéval* (BFM) pour l'édition des manuscrits (Guillot-Barbance et al. 2017b : 149) et cela pour les mêmes raisons : le point de départ de la recherche en linguistique est nécessairement le texte manuscrit primaire, et le corpus électronique doit permettre au chercheur d'accéder le plus directement possible à ce texte de base. Le texte de base du corpus OGR est une transcription légèrement normalisée du manuscrit de base associée à une transcription diplomatique. Puisque tous les textes sauf *Alexis* sont conservés dans un seul manuscrit, le texte des éditions publiées est généralement fiable ; cependant, toutes les leçons ont été soigneusement vérifiées, soit sur les images du manuscrit, soit sur les transcriptions diplomatiques de Foerster et Koschwitz (1932), si les images ne sont pas encore facilement disponibles. La transcription diplomatique des textes suit les principes énoncés par Guillot-Barbance et al. (2017b : 148–149) mais retient en outre les signes diacritiques médiévaux ainsi que la division des mots médiévale. Les modifications éditoriales, comme la résolution des abréviations paléographiques ou la correction des fautes claires et banales (par exemple, la répétition d'un mot à la fin d'une ligne au début de la prochaine), sont toujours signalées entre crochets dans la transcription.

3 Annotation : État de l'art et enjeux

3.1 Le mot : texte de base, lemmatisation et annotation morphosyntaxique

Sur le plan de l'annotation du mot, les enjeux sont bien connus et plusieurs schémas d'annotation ont été développés, par exemple par l'équipe de la BFM (Guillot-Barbance et al. 2017a) ou sur la base du schéma Penn pour le corpus arboré MCVF (Martineau et al. 2010). En général, le corpus OGR suit les normes de la BFM pour assurer une compatibilité maximale entre les deux corpus et adopte le jeu d'étiquettes Cattex (Guillot et al. 2013) et le *Dictionnaire du moyen français* (DMF 2020) comme source principale des lemmes. Pour l'annotation morphologique, le corpus adopte également les recommandations du jeu d'étiquettes Cattex-max (Prévost et al. 2013).

La constitution du corpus présente cependant quelques particularités qui exigent une solution différente de celle pratiquée par la BFM. Nous avons opté pour un texte de base le plus simple possible, qui contient le texte du manuscrit en minuscules, sans l'ajout de la ponctuation ou des signes diacritiques modernes. L'absence de signes supplémentaires simplifie les requêtes sur les formes et facilite de plus l'analyse phonologique de la graphie (voir section 4). Deux modifications ont cependant été introduites. Premièrement, puisque la distinction entre consonne et voyelle est essentielle pour l'annotation phonologique, le texte de base distingue ⟨j⟩ et ⟨v⟩ (consonnes) de ⟨i⟩ et ⟨u⟩ (voyelles ou semi-voyelles). Deuxièmement, la tokenisation en mots se base non pas sur les normes éditoriales de l'ancien français, conçues surtout pour rendre le texte plus accessible aux lecteurs habitués à la typographie moderne, mais sur une définition linguistique du token en tant qu'unité syntaxiquement indépendante. Les clitiques sont ainsi traités comme des mots indépendants.

La Figure 1, présentée à la fin du texte, montre le texte et l'annotation du mot pour un seul vers du texte occitan *Boeci* au moyen de la vue *words* dans la version ANNIS du corpus. Sur le plan de la tokenisation, le mot clitique non-syllabique *l* 'le' est séparé du mot d'appui dans le texte de base. Puisque la tokenisation standardisée ne correspond pas à la division des mots dans le manuscrit, le texte diplomatique est subordonné au texte normalisé et est encodé par moyen de deux étiquettes : *dipl*, qui montre les caractères dans le manuscrit qui correspondent au token, et *wd_div*, utilisé pour l'encodage de tout ce qui se trouve à droite du token dans le manuscrit, y compris les signes de ponctuation. Les caractères spéciaux '_' et '|' figurent dans *dipl* et *wd_div* et dénotent respectivement un espace blanc et un saut de ligne. L'agglutination de deux tokens dans le manuscrit est indiquée par le symbole '+' dans *wd_div*. De la Figure 1, on observe ainsi que *ki*, *et* et *in* sont agglutinés au mot suivant dans le manuscrit, c'est-à-dire ⟨kil⟩, ⟨&uius⟩ et ⟨iniutimẽ⟩. L'annotation morphosyntaxique est représentée dans les couches *pos_syn* et *morph*, qui suivent le schéma Cattex avec quelques modifications ponctuelles[iii].

Sur le plan de la lemmatisation, en raison des divergeances entre les traditions philologiques, les dictionnaires de l'ancien français ne tiennent pas compte des formes occitanes et vice versa, à moins qu'elles ne soient attestées dans les textes de provenance mixte ou incertaine. En outre, certaines formes attestées uniquement dans les plus anciens textes du français sont absentes du DMF, par exemple la particule négative *giens* (*Alexis*, v. 268). Nous avons donc choisi d'adopter une lemmatisation double, qui est illustrée dans la Figure 1. La propriété *lemma* est systématiquement renseignée mais s'appuie sur des sources différentes, notamment le *Dictionnaire d'Occitan Médiéval* (DOM)[iv] pour tous les textes occitans et éventuellement d'autres dictionnaires de l'ancien français pour les lemmes qui manquent dans le DMF. La source du lemme est documentée par la propriété *lemma_src* (non affichée dans la Figure 1). La propriété *lemma_dmf*, par contre, ne contient que les lemmes du DMF, qui sont, le cas échéant, spécifiés à côté du lemme du DOM dans les textes occitans pour les nombreux mots communs aux deux langues. Cependant, pour des mots apparentés qui ne figurent pas dans le DMF, ce champ reste vide.

3.2 Annotation métrique

Les corpus de français dotés d'une annotation métrique sont rares. Pour le français classique et moderne, le projet Anamètre (Delente et Renault 2015) a développé un corpus de vers annoté par un algorithme qui part de la forme graphique (standardisée), identifie les voyelles et les voyelles sujettes à l'élision, et procède ainsi à l'attribution d'une position métrique dans le vers à chaque syllabe. Pour l'ancien français et l'ancien occitan, Rainsford (2011b) et Rainsford et Scrivner (2014) ont développé un algorithme similaire capable de analyser automatiquement le décompte des syllabes dans le vers. L'algorithme exige en entrée :

1. un texte divisé en syllabes ;
2. information sur l'accentuabilité de chaque syllabe ;
3. information sur la possibilité pour chaque syllabe d'être élidée ou non, notamment de ⟨e⟩ inaccentué devant voyelle ;
4. information sur la forme du vers dans le texte : longueur du vers, position et nature possible de la césure, etc.

En sortie, l'algorithme attribue une position métrique à chaque syllabe dans le vers, identifie les syllabes élidées, et annote la position et le type de la césure (Rainsford et Scrivner 2014 : 153–154). Enfin, Poggio et Premat (2019) proposent un *Programme d'Analyse métrique*[v], également capable d'analyser les vers à partir de la forme graphique et d'en proposer plusieurs coupes possibles.

Dans le corpus OGR, nous avons choisi de poursuivre l'approche de Rainsford et Scrivner (2014), car l'algorithme intègre déjà une infrastructure capable d'exporter les

résultats vers la plateforme ANNIS. La Figure 2, présentée à la fin du texte, montre à nouveau le vers 17 de *Boeci* mais en utilisant la vue *meter (grid)* du corpus ANNIS, qui se focalise sur l'annotation métrique. Il est à noter que les unités étiquetées ici sont le vers, la syllabe et le segment. La propriété *line_met* encode la longueur du vers (les deux premiers caractères, c'est-à-dire '10' pour un vers décasyllabique), la nature de la rime (caractère 3, 'm' pour masculin) et de la césure (caractères 4 et 5, c'est-à-dire 'n4' pour une césure normale ou masculine à la quatrième syllabe). La propriété *syll_met* donne la position métrique de la syllabe dans le vers en comptant de gauche à droite (caractères 1 et 2) et en comptant de droite à gauche (caractères 3 et 4); le dernier caractère indique si la syllabe se trouve à la rime ('r') ou à la césure ('c').

3.3 Annotation phonologique

3.3.1 Enjeux principaux pour une annotation phonologique

La graphie des plus anciens textes gallo-romans se base pour la plupart sur le principe phonographique, tout comme la graphie du latin médiéval (Meisenburg 1996 : 81–83, Selig 2006: 1941). Si l'ancien français du 12e et surtout du 13e siècle sera marqué par le développement des *scriptas*, c'est-à-dire des conventions régionales d'écriture en partie indépendantes de la phonie (Gossen 1967 : 15), les textes de notre corpus relèvent d'une phase plus « experimentale » de l'écriture (Meisenburg 1996 : 59). Par conséquent, l'annotation phonologique du corpus cherche à respecter le plus possible la graphie, même si la nature « voyageuse » de certains textes conduit à ce que la langue ainsi représentée n'est pas homogène. En outre, en raison de l'interêt linguistique des plus anciens textes, l'interprétation phonique de la graphie est bien traitée dans de nombreuses études et de nombreux commentaires, ce qui donne une base sûre pour l'annotation.

La tokenisation de base repose sur le segment phonologique, typiquement représenté par un seul caractère mais qui peut regrouper deux caractères formant un digraphe tels que ⟨ch⟩, ⟨qu⟩, et ⟨ss⟩[vi]. Cependant, s'il est relativement simple de tokeniser le texte en segments phonologiques, l'établissement d'un étiquetage approprié pour les segments soulève plusieurs difficultés. Tout d'abord, l'association entre graphie et phonie est étroite mais la graphie n'est aucunement phonémique, d'une part à cause de la polyvalence de certains graphèmes (notamment ⟨e⟩, voir ci-dessous), et d'autre part à cause de la variabilité graphique, qui pourrait indiquer des graphies alternatives de la même séquence phonémique (p. ex. ⟨k⟩ et ⟨qu⟩), des variantes diatopiques (p. ex. le mélange des parfaits occitans en ⟨-et⟩/⟨-ed⟩ avec des parfaits français en ⟨-ad⟩ dans *Passion*) ou même des variantes diachroniques, par exemple la coexistence dans *Alexis* des graphies ⟨ie⟩ et ⟨e⟩ pour la diphtongue [ie], qui s'est réduite à [e] dans les dialectes de l'ouest. La graphie de chaque texte est donc fortement influencée par la phonie mais ne correpond pas toujours à un seul système phonologique. En outre, il existe une tension évidente entre la nécessité d'adopter une annotation cohérente pour tout le corpus et le fait que le système phonologique de chaque texte est différent.

3.3.2 Représentation sous-spécifiée des segments

Au niveau conceptuel, la solution adoptée dans le corpus OGR se base sur la possibilité d'une annotation plus ou moins sous-spécifiée de chaque segment mais avec une base commune de traits phonologiques. En suivant l'approche de la phonologie générative, nous modélisons chaque segment phonologique comme une matrice de traits binaires, qui peuvent être positifs, négatifs, ou *non-renseignés*[vii]. Chaque matrice de traits est associée à

un caractère unique, en général un symbole de l'API pour les segments pleinement spécifiés et une majuscule pour les segments sous-spécifiés.

Tableau 2: Traits vocaliques et voyelles non-postérieures non-fermées dans le corpus OGR

Symbole	API	cons	son	nas	LABIAL	round	DORSAL	high	low	back	atr	voice
V		-	+									
Æ		-	+	-			+	-		-		+
E		-	+	-			+	-	-	-		+
e	e	-	+	-			+	-	-	-	+	+
ɛ	ɛ	-	+	-			+	-	-	-	-	+
A		-	+	-			+	-	+	-		+
a	a	-	+	-			+	-	+	-	-	+
æ	æ	-	+	-			+	-	+	-	+	+

Le Tableau 2 montre les traits définis pour l'encodage des voyelles ainsi que les symboles disponibles pour la transcription des voyelles écrites ⟨e⟩ et ⟨a⟩. Quatre matrices sont pleinement spécifiées et sont dénotées par un symbole de l'API: /e/, /ɛ/, /a/, et /æ/. Les symboles /A/ et /E/ n'ont aucune valeur pour le trait [±ATR][viii] et regroupent alors /a, æ/ et /e, ɛ, ə/ respectivement. Un symbole très sous-spécifié, /Æ/, regroupe ensuite toutes les voyelles non-postérieures et non-fermées du français. Enfin, le symbole /V/ peut désigner toute voyelle, sans spécification du lieu d'articulation.

3.3.3 Conventions de transcription

Le fait que chaque symbole soit associé à une matrice de traits plus ou moins spécifique donne un grand degré de flexibilité d'annotation pour l'interprétation d'une graphie variable, ambigüe et elle-même sous-spécifiée. En même temps, cette flexibilité nécessite l'établissement des conventions de transcription claires valables pour tout le corpus.

En raison des différences phonologiques entre les variétés gallo-romanes, les conventions de transcription diffèrent pour chaque texte. Par exemple, dans les textes français, ⟨tel⟩ serait transcrit /tæl/, parce que la voyelle /æ/ (⟨e⟩) est en opposition avec /a/ (⟨a⟩). Dans les textes occitans, ⟨tal⟩ est transcrit /tAl/, car l'opposition entre /a/ et /æ/ n'existe pas en occitan. La transcription se sert ainsi toujours du symbole le *moins* spécifique possible capable de représenter les oppositions phonémiques présentes dans le système phonologique du texte concerné.

Si la graphie d'un phonème est variable, la solution adoptée varie selon le cas, et dépend de l'analyse du système graphique du texte et les interprétations proposées dans les éditions critiques et des commentaires. Certaines variantes, par exemple ⟨tiel⟩ 'tel' dans *SLeger*, se situent uniquement au niveau graphématique. Puisque le développement d'une diphtongue /ie/ est inconnu ici, le graphème ⟨ie⟩ doit correspondre simplement à /æ/ dans ce texte. D'autres variantes, telles que la notation variable de la syllabe finale inaccentuée dans plusieurs de nos textes (p. ex. ⟨karl**o**⟩, ⟨karl**e**⟩ dans *Serments*), relèvent de la difficulté de transcrire des sons inconnus en latin, dans ce cas-ci, la voyelle /ə/. Ce qui est plus difficile à traiter, c'est la variabilité graphique qui relève de la variabilité phonologique, et en règle générale la transcription phonologique respecte la graphie de chaque forme individuelle, par exemple dans le cas de la monophtongaison de /ie/ à /e/ dans *Alexis*, indiquée par la concurrence des graphies ⟨ie⟩ et ⟨e⟩. Il arrive parfois, mais rarement, qu'une règle contextuelle puisse être identifiée, notamment dans le cas de ⟨an⟩ et ⟨en⟩ pour /eN/

étymologique dans *Alexis*, où la variation se limite aux syllabes fermées inaccentuées : on peut comparer, par exemple *amfant* /ÆN'fANT/, où le copiste écrit ⟨a⟩ pour /E/ inaccentué, avec le nominatif *emfes* /'ENfəS/, où ⟨e⟩ est systématiquement maintenu en position tonique. On peut noter ici l'emploi de /Æ/ sous-spécifié pour la neutralisation de l'opposition entre /E/ et /A/.

Enfin, il existe quelques cas de variation ou de sous-spécification graphiques qui sont impossibles à interpréter de manière satisfaisante. Le cas le plus difficile est le passage de [k] à [ʧ] devant [a], qui a une distribution diatopique restreinte dans le nord du domaine gallo-roman et est typiquement représenté dans la graphie par ⟨ch⟩. Les difficultés proviennent d'une part de l'existence d'un mélange des graphies dans le même texte, par exemple ⟨**ch**er⟩ < *carum* mais ⟨**c**artre⟩ < *carcer* dans *Alexis*, et d'autre part de la polyvalence du graphème ⟨ch⟩, qui peut également dénoter [k], par exemple dans ⟨chi⟩. Ici, l'annotation reste très sous-spécifiée, et nous utilisons le symbole /ċ/ associée à une matrice de traits qui l'identifie simplement comme une obstruante dorsale non-continue et sourde, ce qui regroupe une variété de réalisations phonétiques possibles ([k], [c], [ʧ], etc.).

Cette approche permet ainsi d'utiliser un seul schéma d'annotation, avec des matrices de traits et des symboles valables pour tous les textes, mais en l'adaptant au système phonologique de chaque texte tel qu'il est représenté par la graphie. En outre, la représentation de l'annotation dans la version ANNIS du corpus permet des requêtes sur des matrices de traits plus ou moins spécifiques, ce qui fait que l'hétérogénéité des transcriptions phonologiques n'empêche pas la formulation des requêtes valables pour tous les textes (voir ci-dessous, section 5).

4 Implémentation technique de l'annotation phonologique : la chaine de traitement

Pour pouvoir intégrer une annotation phonologique et métrique, la création du corpus OGR a exigé le développement d'une infrastructure technique, qui prend la forme d'une série de scripts, publiés sur Sourceforge sur dans le cadre du projet *Syllabic Verse Analysis (SylVA)*[ix]. Ces scripts servent notamment à gérer le passage entre une forme de l'annotation phonologique lisible pour l'annotateur, c'est-à-dire la transcription dans la forme d'une chaine de caractères, et la réalité phonologique sous-jacente, c'est-à-dire une série de matrices de traits binaires qui représentent les segments annotés.

La chaine de traitement comprend plusieurs aller-retours entre l'annotateur et les scripts, et les étapes principales sont les suivantes :

1. *transcribe.py* : transformation de la forme graphique des mots (entrée) dans une transcription phonologique en forme des chaines de caractères (sortie) ;
2. correction manuelle des transcriptions, ajout des distinctions manquantes dans la graphie, modifications basées sur la phonologie du texte individuel ;
3. *syllabize.py* : transformation de la transcription définitive des mots (entrée) dans une chaine de caractères représentant la syllabation des mots (sortie) ;
4. correction manuelle de la syllabation, intégration de l'annotation dans les fichiers source du corpus ;
5. *export_text.py* : conversion des sources du corpus vers deux formats XML (XML-TEI P5 et PAULA-XML), y compris l'ajout de l'annotation métrique pour les textes en vers ;
6. ajout des métadonnées et conversion des fichiers PAULA-XML vers le format relANNIS et chargement du corpus dans le portail ANNIS sur http://www.ogr-corpus.org ; import des fichiers XML-TEI P5 dans TXM.

Les étapes 1 à 4 appartiennent à la phase de l'annotation et nécessitent une intervention manuelle. Le script SylVA *transcribe.py* (étape 1) facilite l'établissement de la transcription

en transformant les caractères graphiques en segments phonologiques selon des règles définies par l'annotateur, ce qui permet la reconnaissance des digraphes (tels que ⟨qu⟩ pour /k/) et la désambiguïsation partielle des graphèmes polyvalents au moyen du contexte (par exemple, ⟨c⟩ devant ⟨i⟩ ou ⟨e⟩ représente /ʦ/). Pour une transcription générique qui distingue simplement les voyelles des consonnes, ce script fournirait des résultats satisfaisants ; cependant, pour l'annotation du corpus OGR, une étape de correction et d'enrichissement manuel de l'annotation (étape 2) est essentielle. Avec la lemmatisation manuelle des textes, cette étape s'est avérée la plus chronophage du processus de la création du corpus.

La syllabation des transcriptions (étapes 3 et 4 de la chaine de traitement) implique non seulement l'identification des divisions syllabiques mais aussi de la structure interne de la syllabe, en distinguant l'attaque consonantique du noyau vocalique et de la coda. Le script SylVA *syllabize.py* propose un algorithme de syllabation en quatre étapes :

A. identification des noyaux possibles (les voyelles) ;
B. division du mot en syllabes possibles en respectant le principe selon lequel chaque syllabe doit contenir au moins un noyau, et analyse de la structure interne des syllabes produites (attaque, noyau, coda) ;
C. classement des candidats générés à l'étape 2, qui se base sur (i) une préférence pour les attaques remplies et des codas vides et (ii) la sonorité des segments, qui doit augmenter de l'attaque vers le noyau et redescendre dans la coda ;
D. sélection du meilleur candidat suite à l'élimination des candidats impossibles par moyen des contraintes phonotactiques hiérarchisées définies par l'annotateur et spécifiques à l'ancien gallo-roman[x].

Par exemple, pour la syllabation de *fortment* /fɔrtmENT/, l'algorithme identifie d'abord deux noyaux, /ɔ/ et /E/ (étape A). Ensuite, toutes les divisions possibles entre les noyaus sont identifiées : ici, il s'agit des candidats *fɔ.rtmENT*, *fɔr.tmENT*, *fɔrt.mENT* et *fɔrtm.ENT* (étape B). À l'étape C, les candidats sont classés selon les principes généraux de la bonne syllabation et la forme préférée sera alors /fɔr.tmENT/, avec une coda simple et une attaque branchante dont la sonorité augmente vers le noyau, et ensuite *fɔrt.mENT* (coda branchante mais sonorité respectée), *fɔ.rtmENT* (coda vide mais la sonorité de l'attaque descend de /r/ à /t/ et puis augmente à nouveau), et finalement *fɔrtm.ENT* (coda longue sans sonorité décroissante) (étape C). L'étape C produira souvent un résultat correct, mais dans cet exemple l'attaque /tm/ n'est pas conforme à la phonotactique du gallo-roman, car les seules attaques branchantes possibles sont du type obstruante plus liquide. Nous avons donc formulé une contrainte qui pénalise tout candidat où le deuxième élément de l'attaque branchante n'est pas /r/ ou /L/, ce qui éliminera la syllabation *fɔr.tmENT* à l'étape D et sélectionnera alors le deuxième candidat proposé à l'étape C, *fɔrt.mENT*. La syllabation finale proposée par l'algorithme marque aussi la division en attaque, noyau et coda par moyen du balisage '/\' autour du noyau, c'est-à-dire *f/ɔ\rt.m/E\NT*.

Cette représentation simple et lisible est la forme de l'annotation phonologique encodée dans les sources du corpus, qui prennent la forme d'un tableau CSV (voir Tableau 3)[xi]. Chaque ligne du tableau contient un seul mot-token et les colonnes sont utilisées pour les étiquettes.

Tableau 3: Extrait du tableau source (*Eulalie*, v. 2)

wd_id	line_id	word	phon_map	syllabified	tei_left	tei_right
5	2	bel	b.e.l	b/ɛ\l	<l n="2">	
6	2	auret	a.u.r.e.t	'/a\u.r/ə\ð		
7	2	corps	c.o.r.p.s	k/ɔ\rPS		

À l'étape 5 de la chaine de traitement, le script interprète activement plusieurs étiquettes dans le tableau source et génère alors des couches d'annotation portant sur les unités autres

que le mot. L'annotation phonologique est établie à partir des étiquettes *syllabified* et *phon_map*, et ce dernier sert à encoder la division des caractères dans le texte en segments phonologiques. Pour les textes versifiés, l'étiquette *line_id* permet à l'algorithme métrique d'identifier les vers à analyser. Enfin, les colonnes *tei_left* et *tei_right* donnent la possibilité d'ajouter des balises TEI avant (*left*) ou après (*right*) le token, ce qui permet l'annotation des unités supérieures au mot (phrase, paragraphe, strophe) et des sauts de ligne et de page dans le manuscrit. En général, il importe de souligner que l'export implique un passage d'une entrée où l'annotation est représentée d'une manière concise et lisible pour l'annotateur à une sortie étoffée où toute l'information est codée explicitement et n'est lisible que par l'ordinateur. Un extrait de la sortie XML-TEI se trouve en appendice et montre seulement une partie de l'encodage du mot *bel* du Tableau 3 (l'extrait qui correspond à tout le Tableau 3 occuperait plusieurs pages). Il faut surtout noter l'ajout des segments et syllabes au moyen de la balise <seg> et l'annotation injectée dans l'attribut *ana* de chaque élément, y compris la matrice des traits phonologiques du segment (*#plus:cons,LABIAL,voice*, *#minus:cont,lat,nas,son,strident*)[xii].

Enfin, à l'étape 6 de la chaine de traitement les fichiers XML sont convertis dans le format binaire exigé par ANNIS et par TXM. Pour la version ANNIS du corpus, la conversion des sources PAULA-XML se fait par l'outil Pepper (Zipser et Romary 2010). Pour la version TXM, l'import du fichier XML-TEI dans la plateforme TXM est assuré par le module d'import XML-TEI zéro avec un fichier XSL de pré-traitement adapté au corpus OGR. Nous avons également adapté le paramétrage des visualisations dans ANNIS et développé des fichiers XSL pour la génération des éditions HTML dans TXM pour améliorer le retour à l'édition.

5 Recherches sur la phonologie de l'ancien français : un exemple

Seule la version ANNIS du corpus permet d'exploiter au maximum l'annotation phonologique et métrique, car on peut profiter de la capacité du logiciel à gérer plusieurs niveaux de tokenisation indépendants et les visualiser dans un format tabulaire. Il est ainsi possible d'extraire en quelques minutes des données qui sont inaccessibles dans d'autres corpus.

L'annotation phonologique peut servir de base pour l'étude de la structure syllabique du très ancien français, comme celle que nous avons effectué sur *Alexis* et la *Chanson de Roland* (Rainsford 2020), et le corpus OGR permet d'accéder rapidement aux données nécessaires à ce type d'analyse. Par exemple, dans le domaine d'oïl, les codas internes ont tendance à chuter : les obstruantes (sauf /S/) sont généralement éliminées avant les premiers textes et les sonantes et /S/ au cours de l'ancien français (Scheer et al. 2020 : §295). Les textes du corpus confirment-ils cette observation ? La requête ANNIS suivante (1) permet l'extraction de tous les mots (nœud #1, *word*) qui contiennent une coda interne (nœud #3, *onc="C"*) qui contient une obstruante (nœud #4, *seg_minus=/.*son.*/*) autre que /S/ (nœud #5, *seg_phoneme !="S"*):

(1)
```
word & onc & onc="C" & seg_minus=/.*son*./ & seg_phoneme !="S"
& #1 _r_ #2
& #1 _o_ #3
& #3 . #2
& #3 _o_ #4
& #4 _=_ #5
```

La fonctionnalité *Frequency Analysis* du portail ANNIS permet alors l'export des résultats dans la forme d'un tableau synoptique avec l'ajout des métadonnées, comme par exemple le texte dans lequel se trouve la forme concernée.

Tableau 4: Extrait du tableau des résultats ANNIS, mots qui contiennent une obstruante autre que /S/ dans la coda interne.

#1|word	meta|text	count
alexis	Alexis	25
ciptet	Alexis	3
baptizet	Alexis	1
conpta	Alexis	1
sanctet	Alexis	1
sedme	Alexis	1
vedve	Alexis	1
afflictiuns	Alexis	1
suzlevet	Alexis	1
…		

Le Tableau 4 montre un extrait du tableau ainsi obtenu, qui se limite aux formes identifiées dans la *Vie de saint Alexis*. Outre le nom propre *Alexis*, les codas internes avec une obstruante se trouvent surtout dans des latinismes (*ba**p**tizet*, *affli**c**tiuns*, *san**c**tet*) ou dans des mots composés (*su**z**levet*). Le maintien de ⟨p⟩ dans les graphies *ci**p**tet* et *con**p**ta*, plus fréquent dans les textes occitans du corpus, est inattendu dans un texte du nord du domaine d'oïl, et une requête sur le lemme *cité* dans *Alexis* révèle que seulement 3 des 15 graphies contiennent ce ⟨p⟩, ce qui suggère que la lettre ne correspond en réalité à aucun segment phonologique dans cette variété du français. Enfin, *vedve* et *sedme* pourraient être de vraies exceptions dans le vocabulaire natif, avec un maintien exceptionnel de /ð/ devant consonne suite à la syncope de la voyelle suivante. En conclusion, en vue du petit nombre de formes ainsi identifiées, on peut conclure que la chute des obstruantes dans la coda est confirmée.

Les couches d'annotation phonétique et métrique aident à l'extraction des données qui servent à l'étude d'autres phénomènes. Puisque les consonnes et les voyelles sont facilement identifiables (*seg_plus=/.*cons.*/* vs. *seg_minus=/.*cons.*/*), il devient facile d'étudier les phénomènes sandhi qui dépendent de la nature vocalique ou consonantique du segment initial (ou final) du mot suivant (ou précédent). On peut ainsi utiliser le corpus pour étudier la distribution de la consonne finale des conjonctions telles que *et* (⟨e⟩/⟨ed⟩) ou *à* (⟨a⟩/⟨ad⟩) devant voyelle ou devant consonne (voir Chasle 2008, Russo 2013), ou bien pour étudier la distribution des pronoms enclitiques (Rainsford 2014). En outre, l'annotation métrique permet d'étudier la nature de la coupe d'une manière similaire aux résultats générés par le PAM de Poggio et Premat (2019). Ensuite, en formulant une requête basée sur l'étiquette qui indique la position de l'accent lexical de chaque mot (*syll_lstress*) et l'annotation de la position métrique de la syllabe dans les vers (*syll_metpos*), il est possible d'étudier le rythme du vers en calculant combien de fois, par exemple, la quatrième syllabe d'un vers octosyllabique est accentué (Rainsford 2011a,b). Enfin, les autres couches d'annotation (lemmatisation, partie du discours et morphologie vérifiées) ainsi que la transcription diplomatique du manuscrit de base constituent une base très fiable pour les études lexicologiques ou morphosyntaxiques sur les plus anciens textes.

6 Conclusion

Le corpus OGR comble une lacune dans le domaine des corpus historiques des langues romanes en réunissant les plus anciens textes du français et de l'occitan dans une seule base, une solution qui, à notre avis, est mieux adaptée à la réalité linguistique d'une période située avant le développement des traditions textuelles distinctes dans les deux langues.

Avec l'ajout des couches d'annotation phonologique, métrique et morphologique, le corpus cherche aussi à répondre aux besoins particuliers des philologues et des chercheurs en phonologie historique. En même temps, l'annotation de la partie du discours et la lemmatisation ainsi que l'utilisation de la plateforme TXM bien connue des spécialistes du français médiéval font que l'utilité du corpus n'est pas limitée aux phonologues. Enfin, la publication sous licence libre du corpus compilé, des sources du corpus et des scripts SylVA qui facilitent l'annotation phonologique vise à assurer non seulement le partage des données actuelles mais également le développement de la ressource à l'avenir.

Références bibliographiques

Chasle, N. (2008). Manifestation de la latence en ancien français aux Xème et XIème siècles: liaison et redoublement syntaxique. In : *Congrès Mondial de Linguistique Française 2008*. Paris : EDP Sciences, p. 1645–1656. <https://doi.org/10.1051/cmlf08175> [consulté le 14 mars 2022].

DMF (2020). *DMF : Dictionnaire du Moyen Français*, version 2020. ATILF - CNRS & Université de Lorraine. <http://www.atilf.fr/dmf> [consulté le 5 janvier 2022].

De Poerck, G. (1963) Les plus anciens textes de la langue française comme témoins de l'époque. *Revue de linguistique romane*, 27, 1–34.

Foerster, W., et Koschwitz, E. (1932). *Altfranzösisches Übungsbuch*. 7e éd. Leipzig : Reisland.

Gossen, C. T. (1967). *Französische Skriptastudien: Untersuchungen zu den nordfranzösischen Urkundensprachen des Mittelalters*. Vienne : Böhlau.

Guillot, C., Prévost, S., et Lavrentiev, A. (2013). Manuel de référence du jeu Cattex09, version 2.0. *BFM - Base de Français Médiéval* [En ligne]. Lyon : ENS de Lyon, <http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf> [consulté le 5 janvier 2022].

Guillot-Barbance, C., Heiden, S. et Lavrentiev, A. (2017a). Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique. *Diachroniques*, 7, 168–184.

Guillot-Barbance, C., Lavrentiev, A. Rainsford, T. Marchello-Nizia, C, et Heiden, S. (2017b). La « philologie numérique » : tentative de définition d'un nouvel objet éditorial. In : R. Trachsler, F. Duval, et L. Leonardi *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 13 : Philologie textuelle et éditoriale.* Nancy : ATILF/SLR, p. 143–154.

Gussenhoven, C., et Jacobs, H. (2005). *Understanding Phonology*. 2nd ed. Londres : Hoddor Arnold

Heiden, S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In : R. Otoguro, K. Ishikawa, H. Umemoto, K. Yoshimoto et Y. Harada. *24th Pacific Asia Conference on Language, Information and Computation, Nov 2010, Sendai, Japan. Institute for Digital Enhancement of Cognitive Development, Waseda University*, p. 389-398. <halshs-00549764> [consulté le 5 janvier 2022].

Krause, T. et Zeldes, A. (2016). ANNIS3: A new architecture for generic corpus query and visualization. In : *Digital Scholarship in the Humanities 2016*, 31. <http://dsh.oxfordjournals.org/content/31/1/118> [consulté le 5 janvier 2022].

Meisenburg, T. (1996). *Romanische Schriftsysteme im Vergleich : Eine diachrone Studie*. Tübingen : Narr.

Poggio, E. et Premat, T. (2019). Le PAM, un Programme d'Analyse Métrique pour le français médiéval. In : *Zenodo*. <https://doi.org/10.5281/ZENODO.3464477> [consulté le 5 janvier 2022].

Pope, M. K. (1934). *From Latin to Modern French with especial consideration of Anglo-Norman: Phonology and Morphology*. Manchester : Manchester University Press.

Prévost, S., Guillot, C., Lavrentiev, A., et Heiden, S. (2013). Jeu d'étiquettes morphosyntaxiques CATTEX2009, version 2.0. *BFM - Base de Français Médiéval* [En ligne]. Lyon : ENS de Lyon, <http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf> [consulté le 5 janvier 2022].

Rainsford, T. (2011a). Dividing lines : The changing syntax and prosody of the mid-line break in medieval French octosyllabic verse. *Transactions of the Philological Society*, 109, 265–283. <https://doi.org/10.3352/jeehp.2013.10.3> [consulté le 14 mars 2022].

Rainsford, T. (2011b). The Emergence of Group Stress in Medieval French. Thèse de doctorat, University of Cambridge. <https://doi.org/10.17863/CAM.16503> [consulté le 5 janvier 2022].

Rainsford, T. (2014). Sur la disparition de l'enclise en ancien français. In : W. Ayres-Bennett et T. Rainsford (éd.) *L'Histoire du français. État des lieux et perspectives*. Paris : Garnier. p. 21–44. <http://doi.org/10.15122/isbn.978-2-8124-2986-6.p.0021> [consulté le 14 mars 2022].

Rainsford, T. (2022). *Old Gallo-Romance Corpus*, version 0.4. Stuttgart: Institut für Linguistik/Romanistik. <http://www.ogr-corpus.org> [consulté le 18 mars 2022].

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

Russo, M. (2013). Il raddoppiamento fonosintattico dell'italiano. Tratti prosodici e struttura fonologica. In : F. Sánchez Miret et D. Recasens i Vives (éd.) *Studies in Phonetics, Phonology and Sound Change in Romance*. Munich : LINCOM Europa, p. 145–178.

Scheer, T, Ségéral, P., Gess, R. S., Jacobs, H., et Laks, B. (2020). Partie 3: Phonétique historique. In : C. Marchello-Nizia, B. Combettes, S. Prévost, et T. Scheer (éd.) *Grande grammaire historique du français (GGHF)*. Berlin: De Gruyter Mouton, p. 151–490.

Selig, M. (2006). Die Anfänge der Überlieferung der romanischen Sprachen : Quellentypen und Verschriftungsprinzipien. In : G. Ernst, M.-D. Gleßgen, C. Schmitt, W. Schweickard et J.-P. Chambon (éd.) *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen*. Berlin : de Gruyter, p. 1924–1944.

Vennemann, T. (1988). *Preference Laws for Syllable Structure and the Explanation of Sound Change: With Special Reference to German, Germanic, Italian, and Latin*. Berlin: Mouton de Gruyter.

Zipser, F. et Romary, L. (2010). A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010*. Malta. <http://hal.archives-ouvertes.fr/inria-00527799/en/> [consulté le 5 janvier 2022].

Appendice: Extrait de la représentation XML-TEI du texte du tableau 3

```
<w
    xml:id="Eulalie_w_5"
    ana="#line_id:2
         #phon_map:b.e.l
         #syllabified:b/ɛ\l
         #prosody:1ox
         #metpos:01.-
         #soptem:10.-
         #line_met:10---"
    met="01.-">
    <!-- Segments phonologiques dans le mot -->
    <seg
        xml:id="Eulalie_seg_22"
        type="segment"
        ana="#plus:cons,LABIAL,voice
             #minus:cont,lat,nas,son,strident
             #phoneme:b
             #onc:O">
        <!-- Caractères correspondants au segment -->
        <c>b</c>
    </seg>
    ...
</w>
...
<!-- Définition des syllabes à la fin du texte -->
<standOff>
...
<seg
    xml:id="Eulalie_syll_10"
    type="syllable"
    ana="#is_monosyllable:y
         #has_onset:y
         #lstress:y
         #form:b/ɛ\l
         #metpos:0110-"
    met="0110-">
<!-- Référence aux segments qui forment l'attaque, le noyau et la coda -->
    <span target="#Eulalie_seg_22" ana="#onset" />
    <span target="#Eulalie_seg_23" ana="#nucleus" />
    <span target="#Eulalie_seg_24" ana="#coda" />
</seg>
...
</standOff>
```

⊟ words

line_ref	Boeci, l. 17								
word	ki	l	mort	et	vius	tot	a	in	jutjamen
dipl	ki	l	mort	[et]	uius	tot	a	in	iutiam[en]
wd_div	+	_	_	+	|	_	_	+	.
lemma	qui	lo2	morir	e	viu	tót	avér	én	jutjamén
lemma_dmf	qui	le	mourir	et	vif	tout	avoir	en	jugement
pos_syn	PRNrel	DETdef	NOMcom	CONcoo	ADJqua	PRNind	VERcjg	PRE	NOMcom
morph	smn	smr	smr	--	pmr	snr	indpst3	--	smr

Fig. 1: Visualisation *words* du vers 17 de *Boeci*, version ANNIS du corpus.

⊟ meter (grid)

line_ref	Boeci, l. 17																									
line_met	10mn4																									
syll_metpos	0110-			0209-				0308-	0407c				0506-			0605-	0704-		0803-		0902-			1001r		
syll_lstress	n			y				n	y				y			y	n		n		n			y		
seg_phoneme	k	i	L	m	ɔ	r	T	e	v	i	u	S	t	o	T	A	i	N	ʤ	y	t	ʤ	A	m	E	n
tok	k	i	l	m	o	r	t	et	v	i	u	s	t	o	t	a	i	n	j	u	t	j	a	m	e	n

Fig. 2: Visualisation *meter (grid)* du vers 17 de *Boeci*, version ANNIS du corpus.

i Les plus anciennes chartes en occitan sont exclues du corpus mais font cependant partie du corpus des *Plus anciens documents linguistiques de la France* dirigé par Martin-D. Gleßgen : https://www.rose.uzh.ch/phoenix/workspace/web/ [consulté le 4 janvier 2022]. Ce corpus a également le mérite d'inclure les textes du domaine d'oc et du domaine d'oïl.

ii L'exemple le plus évident, pour l'instant exclu du corpus, est le manuscrit d'Oxford de la *Chanson de Roland*, qui date du 2e quart du 12e siècle (DEAFBiblél).

iii Nous distinguons notamment les pronoms clitiques (PRC) des pronoms nominaux (PRN) et les personnes verbales sont annotées de 1 à 6 au lieu de 1s à 3p.

iv http://www.dom-en-ligne.de, consulté le 5 janvier 2022.

v https://github.com/EPgg92/pam_project, consulté le 5 janvier 2022.

vi Très rarement, un seul caractère, par exemple ⟨x⟩, représente deux segments phonologiques. Dans ce cas, le deuxième segment est associé à un espace insécable ajouté au texte de base.

vii Nous avons adopté avec quelques modifications des matrices de traits proposées dans l'introduction à la phonologie générative de Gussenhoven et Jacobs (2005).

viii Nous utilisons le trait [±ATR] pour indiquer l'opposition entre les voyelles tendues (/e/, /æ/ et les voyelles relâchées correspondantes (/ɛ/, /a/).

ix https://sourceforge.net/projects/syllabic-verse-analysis, consulté le 5 janvier 2022.

x Du point de vue théorique, l'étape 3 s'inspire des lois de préférence proposées par Vennemann (1988) : l'algorithme attribue un coefficient de bonne formation à chaque noyau, attaque et coda en les additionnant pour arriver au classement final. L'algorithme à l'étape 4 suit les même principes de la théorie de l'optimalité, où les contraintes de bonne formation sont hiérarchisées et la forme sélectionnée est soit celle qui respecte toutes les contraintes, soit celle qui provoque seulement une infraction à des contraintes vers le bas de la hiérarchie. Notons cependant que les contraintes que nous avons développées à l'étape 4 sont plutôt *ad hoc* et ne servent qu'à exclure des résultats non voulus.

xi Les fichiers source du corpus sont publiés sur *GitHub* : https://github.com/rainsfordtm/ogr.

xii Un·e relecteur·trice a observé que le stockage des annotations dans la propriété *ana* pourrait poser problème aux outils. Cette méthode est pourtant la seule conforme aux normes de la TEI, qui ne prévoient pas l'ajout des propriétés supplémentaires aux éléments. Les annotations dans la chaîne des caractères *ana* peuvent être interprétées au moyen d'une feuille de style XSLT avant ou bien lors de l'import dans une plateforme, ce qui est la solution adoptée pour la création du corpus TXM.