Lemmatisation de l'ancien français : Présentation du modèle et des outils de l'École des chartes

Frédéric Duval, Lucence Ing, Jean-Baptiste Camps, Naomi Kanaoka, Ariane Pinche, Thibault Clérice

Lemmatisation de l'ancien français : Présentation du modèle et des outils de l'École des chartes[1]

0. Introduction

Depuis les années 1960, l'annotation de corpus a été automatisée par les linguistes en vue de recherches quantitatives. La lemmatisation est ainsi devenue une condition sine qua non de l'exploitation linguistique de vastes corpora. Ses applications computationnelles sont multiples : outre les progrès de la description linguistique auxquelles elle contribue de plus en plus massivement, la lemmatisation fournit des données essentielles aux enquêtes scriptométriques, si précieuses pour l'identification des lieux de production des manuscrits et donc pour l'histoire des textes ; elle s'avère désormais indispensable en stylométrie, notamment pour la critique d'attribution, en étude des genres de discours ou dans celle des champs sémantiques. Enfin, elle est un préalable à la collation automatique de séquences textuelles en français peu ou pas standardisé (en particulier en ancien français), tant le 'bruit' généré par la variation graphique interdit des classements valides avant ce prétraitement.

Depuis 2015, l'École des chartes-PSL a développé des outils de lemmatisation et des modèles, librement disponibles (Camps et al., 2021b). Nous nous concentrerons ici sur la Old French Corpus Collection of the École des chartes [OF3C], disponible depuis 2021[2] : après une rapide présentation de l'environnement de lemmatisation proposé (outils et données), nous reviendrons sur l'élaboration d'OF3C. Enfin, la contribution livrera les résultats d'une évaluation du modèle, ce qui permettra de juger de ses performances actuelles, mais aussi de dégager des pistes d'amélioration.

1. L'environnement proposé

Les lemmatiseurs traditionnels comme TreeTagger reposent sur des ensembles de règles, un lexique avec des formes connues et un algorithme basé sur un arbre de décision. Ils fonctionnent très bien sur des états de langue standardisés, mais sont moins adaptés à des états de langues non-standardisés. Il est toutefois possible de multiplier et d'affiner les règles afin d'intégrer une dose de variation, comme l'a fait Gilles Souvay avec LGeRM, qui s'appuie sur les vastes données lexicographiques et textuelles du *Dictionnaire du moyen français*[3].

Partant d'états antérieurs moins standardisés et désireux d'aboutir à des résultats non-ambigus pour chaque forme – LGeRM propose fréquemment plusieurs possibilités par occurrence –, nous avons fait le choix de l'apprentissage profond (*deep learning*) et de la technologie des réseaux de neurones récurrents (RNN), particulièrement adaptée au traitement de la variation linguistique de par sa capacité d'apprentissage et de prédiction : grâce au RNN, le lemmatiseur n'a plus besoin de comparer la forme du texte à une base de connaissance, mais peut prédire le lemme, séquence par séquence, caractère par caractère.

---

1 Les outils présentés ici sont le fruit d'un travail d'équipe, qui a rassemblé de nombreux contributeurs : Thibault Clérice et Julien Pilla pour le développement informatique ; Jean-Baptiste Camps, Frédéric Duval, Lucence Ing, Naomi Kanaoka et Ariane Pinche pour la lemmatisation de textes intégrés au modèle OF3C.

2 Les données du corpus sont librement accessibles à l'adresse <github.com/chartes/OF3C>.

3 *LGeRM*, <stella.atilf.fr/LGeRM/>.

Le service de lemmatisation, baptisé *Deucalion*, proposé par l'École des chartes[4], utilise *Pie*, le lemmatiseur et annotateur de parties du discours (POS)[5] développé par Enrique Manjavacas[6]. Cet outil a été complété, sous le nom de *Pie extended*[7], par des développements réalisés par Thibault Clérice et Julien Pilla visant à intégrer des services de prétraitement (tokenization, autrement dit la conversion d'un texte en une liste d'unités à lemmatiser) et de post-traitement.

Ce dernier point a fait l'objet d'une attention particulière, car, si un lemmatiseur peut atteindre des scores élevés, toute lemmatisation automatique d'une langue non-standardisée est imparfaite. Afin d'améliorer les résultats du lemmatiseur, l'intervention humaine est nécessaire. L'enjeu était d'offrir un service de post-correction ergonomique, rapide et collaboratif. L'environnement *Pyrrha*, disponible depuis 2018, mais constamment amélioré depuis, répond à ces exigences[8]. Aisé à prendre en main, ne nécessitant aucune connaissance informatique, il permet des corrections par lots et l'intervention simultanée de plusieurs utilisateurs sur un même corpus.

Après sélection de l'un des modèles de langue disponible, *Deucalion* permet de lemmatiser toute séquence de texte importée ou saisie au format texte en UTF-8. Quant à *Pyrrha*, il offre la même possibilité, augmentée de services de post-correction.

2. Constitution du corpus OF3C

Le lemmatiseur a été entraîné sur l'ancien français grâce à la constitution progressive d'un corpus annoté. Les premières expérimentations, commencées en 2015 par Jean-Baptiste Camps, ont porté sur la *Chanson d'Otinel* éditée dans le cadre de sa thèse de doctorat[9]. Les référentiels de lemme, de POS et de morphologie furent alors choisis. Pour les lemmes, le choix s'est porté sur ceux du Tobler-Lommatzsch, moyennant quelques adaptations et compléments. Pour les POS et la morphologie, le guide d'annotation Cattex09 a été suivi[10], là encore avec quelques menues adaptations[11]. Pour chaque forme, l'information donnée en POS est donc fine, car, outre le lemme, la catégorie grammaticale est précisée : on n'indique pas simplement « possessif », mais « déterminant possessif », « adjectif possessif » ou « pronom possessif » ; non pas simplement « verbe », mais « verbe conjugué », « infinitif » ou « participe passé »). En outre, l'annotation morphologique, renseignée pour une partie seulement du corpus d'entraînement, indique les cas, genre et nombre des déterminants, adjectifs et substantifs ; les modes, temps grammaticaux et personnes des verbes conjugués.

L'élaboration d'un corpus d'entraînement annoté aussi finement a requis beaucoup de temps et d'énergie. Par un processus circulaire, l'annotation a conduit à la mise au point

---

4 *Deucalion, Modèle ancien français*, <dh.chartes.psl.eu/deucalion/>.

5 Seule l'acronyme POS (pour *part-of-speech*) sera utilisé par la suite.

6 Manjavacas, Kádár et Kestemont (2019).

7 Thibault Clérice, *Pie Extended, an extension for Pie with pre-processing and post-processing*, <doi.org/10.5281/zenodo.3883589>

8 Thibault Clérice *et al.*, 2019.

9 Camps (2016).

10 Guillot, Prévost et Lavrentiev (2013).

11 Les principes d'adaptation des référentiels de lemmes (TL) et de la morpho-syntaxe (Cattex) sont explicités dans le wiki du dépôt Geste, à l'adresse suivante : <github.com/Jean-Baptiste-Camps/Geste/wiki>.

d'outils permettant d'accélérer la correction des sorties brutes de lemmatisation. C'est ainsi qu'est né *Pyrrha*.

Le corpus d'ancien français annoté compte au 1er septembre 2022 1 132 849 tokens. Pierre Kunstmann a partagé son corpus lemmatisé des romans de Chrétien de Troyes, qu'il a fallu adapter à nos référentiels. Pour le reste, la lemmatisation a été entièrement prise en charge par notre équipe. Le fait qu'elle soit assez resserrée a permis de rester cohérent dans les choix de lemmatisation, même si une marge laissée à l'interprétation personnelle subsiste inévitablement. Faute d'une dotation de départ très importante, la construction du corpus a été pragmatique, s'appuyant sur les projets menés par les collaborateurs : après les trois manuscrits de la *Chanson d'Otinel*, d'autres textes épiques ont été annotés, aujourd'hui intégrés au corpus *Geste*[12]. Deux thèses ont apporté des récits hagiographiques rédigés par Wauchier de Denain[13] et une partie du *Lancelot* en prose[14]. Quant au projet de recherche portant sur les traductions françaises du *Corpus juris civilis*[15], il a conduit à l'intégration de plusieurs de ces textes de droit savant.

À côté de ces gros blocs de données, nous nous sommes efforcés d'enrichir la couverture chronologique, diatopique et générique du corpus, afin d'augmenter son efficience sur les textes les plus divers. Naomi Kanaoka a ainsi encodé un extrait du *Pèlerinage de l'âme* de Guillaume de Digulleville, alors que le 14e siècle n'était jusqu'à présent pas représenté, non plus que la poésie religieuse allégorique. De même ont été intégrées des pièces de poésie lyrique (Thibaut de Champagne, Conon de Béthune, Gace Brulé) et des chartes sélectionnées dans DocLing de façon à représenter l'arc scripturaire le plus large.

Dataset	**Source**	**Morphologie**	**Nombre de tokens**
Chrétien de Troyes	Kunstmann 2009	non	252774
Corpus juris civilis	Miroir des classiques	partielle	160007
Chartes	DocLing	complète	68317
Geste	Camps 2016	complète	195303
Lancelot	Ing (en cours)	non	286095
WauchierSConf	Pinche 2021	complète	113694
Varia		complète	56659

Figure 1 : Données actuelles d'OF3C

12 Camps (2016).
13 Pinche (2021).
14 Ing (en cours).
15 Frédéric Duval, Lucence Ing, Naomi Kanaoka ; Projet intégré à Duval (2007-), financé par l'Equipex Biblissima.

Même s'ils n'ont pas encore servi à l'entraînement d'une nouvelle version du modèle *Old French*, de nouveaux textes sont venus récemment augmenter le corpus annoté. Ont été traités par Naomi Kanaoka de larges extraits de l'œuvre de Rutebeuf, ce qui permettait d'intégrer des genres absents comme les fabliaux ou le théâtre (*Miracle de Théophile*). Enfin, il a semblé intéressant d'enrichir le modèle d'un extrait du *Roman de la Rose*, tiré de la partie rédigée par Jean de Meun.

Nous en sommes maintenant à un tournant important, car les financements obtenus pour la constitution du modèle sont désormais épuisés[16]. Le moment est donc propice à son évaluation et à sa promotion. En effet, si l'on sait que le modèle entraîné à partir d'OF3C obtient en général de bons résultats, il est nécessaire d'en discerner les failles et les points faibles afin d'être en mesure de les combler par des actions plus ou moins ponctuelles. Pour ce qui est de la promotion, il faut insister sur le fait que le service offert par l'École des chartes n'est pas destiné uniquement à une utilisation computationnelle, mais qu'il est susceptible d'intéresser un public assez large de philologues et de linguistes. Ainsi, notre solution de lemmatisation de l'ancien français peut rendre de nombreux services à l'ensemble des éditeurs de textes par l'aide qu'il apporte à la rédaction du glossaire ou à la préparation d'une introduction linguistique. Mentionnons également les applications pédagogiques de *Pyrrha*, qu'il s'agisse de lemmatisation collective d'un texte en cours ou à distance ou bien de l'étude du marquage des sujets, des formes des démonstratifs, etc.

3. Évaluation du modèle

3.1. Évaluation générale

L'évaluation du modèle *Old French* entraîné sur le corpus OF3C, avant ajout des derniers textes mentionnés *supra*, permet de repérer les aspects à améliorer, c'est-à-dire le type de données qu'il faudrait ajouter au corpus pour que le modèle devienne plus performant. Deux types d'évaluation ont été réalisées. Tout d'abord, une évaluation, dite *in-domain*, qui est effectuée en observant les résultats du modèle sur une partie des données annotées du corpus, qui n'ont pas servi d'entraînement au modèle. Le modèle prédit les étiquettes et cette prédiction est comparée à l'annotation réalisée par l'humain. Les résultats du modèle sur les lemmes et les parties du discours (POS) se trouvent dans le tableau suivant[17] :

---

[16] Le DIM « Sciences du textes et connaissances nouvelles » de la région Île-de-France a généreusement financé l'emploi de Naomi Kanaoka.

[17] Pour des questions de place, nous traiterons principalement des résultats du modèle sur les lemmes, et de manière succincte des résultats sur les POS. Les résultats en morphologie ne seront pas abordés ici. Précotera néanmoins qu'ils varient entre 94.57% pour le cas et 98.97% pour le mode verbal.

	lemmes			POS	
	résultat[18]	nombre d'unités		résultat	nombre d'unités
total	0.9747	73507	total	0.9739	73507
unités connues	0.9831	71726	unités connues	0.9775	71726
unités inconnues	0.6367	1781	unités inconnues	0.8293	1781

Figure 2 : Résultats du modèle entraîné sur le corpus OF3C, pour les lemmes et les POS

Pour les deux catégories, le modèle obtient de bons résultats, avec plus de 97 % de résultats positifs. Néanmoins, il est possible de constater une divergence nette, pour les lemmes, entre les unités connues (plus de 98 % de résultats positifs) et les unités inconnues (moins de 64 %). Les unités inconnues sont des unités qui n'ont jamais été rencontrées par le modèle, qu'il s'agisse du lemme ou de formes graphiques. Cela signifie toutefois que, dans 64% des cas, le modèle réussit à attribuer un lemme à une forme graphique qu'il ne connaît pas : il s'agit d'un avantage notable d'une approche de type *deep learning*, par rapport à une approche basée sur un lexique (dans 7,5% des cas, le modèle réussit même à deviner un lemme inconnu). La différence entre les deux types d'unités est moindre pour les POS, avec un résultat positif qui reste supérieur à 80 %. Cela s'explique en partie par le nombre très faible d'étiquettes de POS et par leur nomenclature close.

La seconde évaluation réalisée est dite *out-of-domain* : elle porte sur des textes qui n'ont pas servi à l'entraînement. Réaliser une telle évaluation est intéressant car elle permet d'évaluer le modèle sur un corpus neutre, contenant des données hétérogènes, parfois radicalement différentes de celles qui ont servi à l'entraînement. Pour cette évaluation, un ensemble d'extraits tirés du *Nouveau corpus d'Amsterdam* (NCA)[19] a été confectionné[20]. Le choix de ce corpus a été motivé par plusieurs facteurs. Tout d'abord, sa qualité et disponibilité ; ensuite, le niveau d'information des métadonnées associées à chaque extrait, qui permet de multiplier les critères d'évaluation du modèle ; enfin, la diversité présumée des témoins présents dans le corpus. Ce corpus test contient 278 extraits de 50 unités de texte, sélectionnés de manière aléatoire et tirés de 150 témoins différents, représentant 106 textes distincts. La longueur des extraits a été déterminée de manière à disposer d'extraits courts, ce qui permettait de les mul-

[18] Les résultats retenus dans cet article sont ceux de l'*accuracy* du modèle. L'*accuracy*, exactitude en français, est ici calculée comme le nombre de prédictions correctes divisé par le nombre total de prédictions. Elle prend en compte chaque prédiction de manière égale. Il ne s'agit donc pas d'une moyenne qui donnerait le même poids à chaque classe, mais d'une mesure qui considère de manière identique toutes les occurrences, sans aucune pondération. D'autres métriques (précision, rappel, F-score) Ils disponibles, avec des informations complémentaires à l'adresse suivante : <github.com/chartes/deucalion-model-af>.

[19] Stein, Achim *et. al.* (2006).

[20] Les textes ont été annotés par Naomi Kanaoka, qui a également effectué en amont un long travail de toilettage des extraits.

tiplier et donc d'accroître la représentativité du corpus-test. Toutefois, ils devaient être assez longs pour être sémantiquement interprétables de façon autonome. Dans le tableau suivant sont donnés les résultats obtenus sur les lemmes et les POS du corpus test.

catégorie	**résultat**	**nombre d'unités**
lemme	91.11	35888
POS	94.74	35888

Figure 3 : Résultats du modèle sur le corpus *out-of-domain*

Les résultats sur le corpus *out-of-domain* confirment l'efficacité du modèle d'annotation, puisque plus de 91 % des résultats sont positifs sur les lemmes et près de 95 % sur les POS. L'étude des erreurs commises par le modèle confirme que les résultats sont bons, puisqu'une partie importante des erreurs est due à des problèmes d'homogénéisation des lemmes dans le référentiel de lemmes : il s'agit par exemple de l'étiquetage de *dieu* pour *Dieu*, mais aussi d'autres noms propres, comme *jüif, sarrasin*. On trouve également dans le référentiel, et dans les différents textes annotés, des lemmes identiques mais présentant des graphies différentes, par exemple *Jésus/Jhesu/Jhesus*. Régler ce type de problème demanderait un travail important d'homogénéisation des données dans le référentiel et dans les données du corpus. Bien que cet effort ait été mené en amont de l'étiquetage du corpus, notamment lors de la constitution même du référentiel de lemmes, l'annotation du corpus par des personnes aux pratiques différentes, sur des textes variés et sur un temps assez long, rend inévitable ce type d'erreurs. Outre les erreurs dues à l'homogénéisation des données, d'autres types d'erreurs sont fréquentes, comme le montre le tableau ci-dessous :

lemme correct	**Total erreurs**	**lemme prédit**	**Fréquence**
Dieu	148	dieu	146
que2	117	que4	77
il	59	le	38
que4	42	sque2	33
avoir	28	a3	11
le	26	il	20
estre1	22	ester	8
que1	21	que4	17
son4	19	soi1	6

Figure 4 : Lemmes présentant le plus d'erreurs, avec leur confusion la plus fréquente

Les erreurs les plus fréquentes se produisent dans des contextes syntaxiques difficiles, en présence d'homographes : le plus grand contingent d'erreurs vient de l'interversion de *que2*, pronom relatif, avec *que4*, conjonction de subordination, accompagnée de l'interversion de *que1*, comparatif, et *que4*.L'homographie entre deux formes peut également se produire de manière occasionnelle et amener à une confusion ponctuelle, comme l'interversion entre *avoir* conjugué à la troisième personne du singulier de l'indicatif présent et *a3*, préposition, ou celle entre *il* employé au cas régime et l'article défini *le*.

Enfin, les formes inconnues produisent également des erreurs, même si le lemmatiseur propose un certain nombre de prédictions exactes. Ces prédictions sont intéressantes à examiner car elles donnent l'occasion d'observer la manière dont le lemmatiseur étiquette les différentes occurrences. Par exemple, la forme *arrieregarde,* inconnue du modèle, est étiquetée *regarder*, selon une logique de proximité de formes entre l'occurrence trouvée dans le texte et le lemme.

L'évaluation sur le corpus *out-of-domain* a aussi été menée sur les lemmes proprement lexicaux. En effet, les résultats obtenus sur l'ensemble des lemmes ne permettent pas de comprendre finement quel type de lexique fait défaut à notre modèle, car un certain nombre de mots-outils vient perturber les résultats. Ainsi, les résultats ont été évalués sur l'ensemble exclusif substantifs/verbes/adjectifs qualificatifs :

	Lemmes					POS			
POS	**Acc.**	**Error contrib**	**Support relatif**	**Diff contrib support**	**POS**	**Acc.**	**Error contrib**	**Support relatif**	**Diff contrib support**
NOMpro	0.0757	0.2596	0.0249	0.2346	ADJqua	0.8059	0.1219	0.0330	0.0888
NOMcom	0.8360	0.2513	0.1362	0.1151	PROrel	0.8627	0.0550	0.0210	0.0339
PROrel	0.7913	0.0495	0.0210	0.0284	NOMcom	0.9364	0.1646	0.1362	0.0283
ADJqua	0.8479	0.0565	0.0330	0.0234	PROind	0.7649	0.0353	0.0079	0.0274
VERppe	0.8719	0.0415	0.0288	0.0127	VERppe	0.9026	0.0533	0.0288	0.0245
DETcar	0.7346	0.0086	0.0028	0.0057	ADVgen	0.9304	0.0842	0.0637	0.0205
DETdef	0.9830	0.0086	0.0453	-0.0367	DETdef	0.9752	0.0213	0.0453	-0.0239
PONfrt	0.9992	0.0003	0.0420	-0.0417	CONcoo	0.9722	0.0275	0.0521	-0.0245
CONcoo	0.9880	0.0069	0.0521	-0.0451	PONfrt	0.9992	0.0005	0.0420	-0.0415
PROper	0.9670	0.0279	0.0754	-0.0474	VERcjg	0.9637	0.0938	0.1361	-0.0423
PRE	0.9803	0.0166	0.0752	-0.0585	PRE	0.9783	0.0308	0.0752	-0.0443
PONfbl	0.9974	0.0023	0.0817	-0.0794	PONfbl	0.9989	0.0016	0.0817	-0.0800

Figure 5 : Résultats du modèle en lemmes et POS par partie du discours (pour les classes d'occurences supérieures ou égales à 100), classées en fonction de leur sur ou sous contribution à l'erreur totale (6 premières et dernières lignes seulement).

Concernant les lemmes, le constat le plus important concerne la contribution démesurée des noms propres aux erreurs (26% des erreurs pour 2.5% des occurrences !). À l'inverse, ponctuations et mots-outils tendent à être des sous-contributeurs à l'erreur totale, qu'il s'agisse de la prédiction des lemmes comme des POS, leur très grande fréquence affectant très favorablement le score global. D'un point de vue général, les hésitations du modèle entre adjectifs, noms et verbes, d'une part, et pronom relatif et conjonction de subordination d'autre part constituent le nœud d'erreur le plus important (fig. 6).

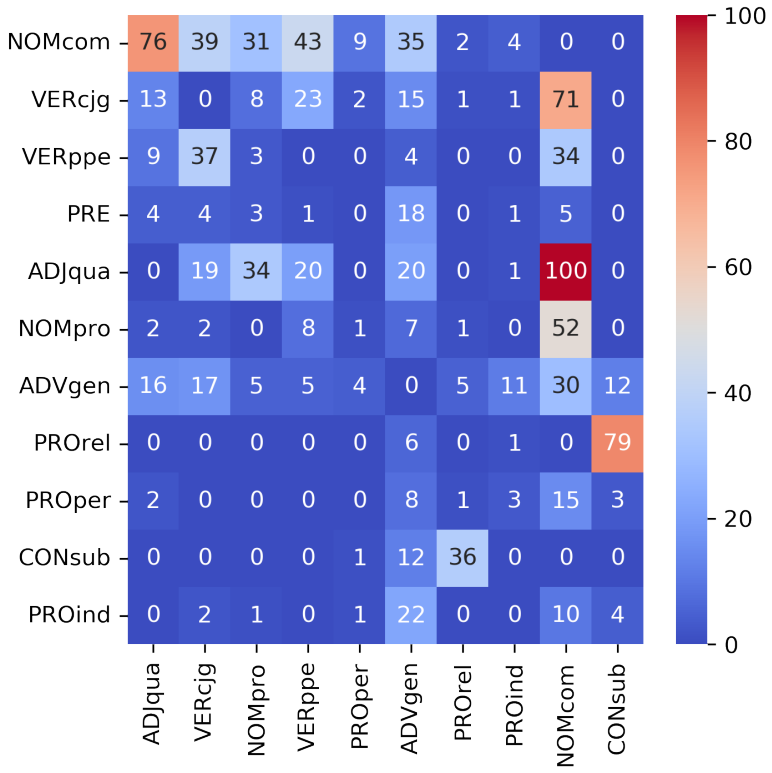

Figure 6: carte de chaleur des confusions les plus fréquentes dans la prédiction des POS

3.2. Évaluation du modèle en fonction de différents critères

Grâce aux différentes métadonnées associées à chaque témoin du corpus du NCA, il est possible d'observer si des facteurs tels que le lieu de rédaction du manuscrit, la date de composition du texte et le genre ont une influence sur les résultats. D'autres facteurs pourraient être explorés, mais ici seuls ces trois facteurs ont été retenus.

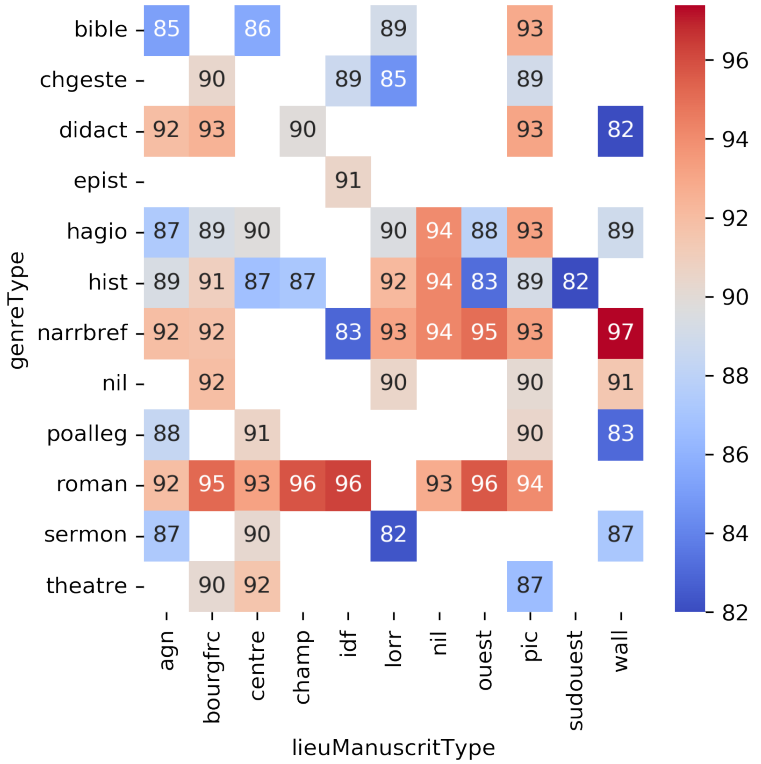

Figure 7 : Résultats (*accuracy*) sur les lemmes en fonction des *scriptae des manuscrits et des genres des textes*

Globalement, romans et formes narrratives brèves (étonnamment, pas les chansons de geste) obtiennent les meilleurs scores, tandis que sermons et textes bibliques sont les moins bien reconnus. Les différences entre *scriptae* semblent moins marquées, point encourageant en ce qui concerne la diversité du modèle, même si l'on note que les résultats de la *scripta* du Sud-Ouest, très peu représentée, sont mauvais en l'état (fig. 7)[21]. Par ailleurs, la *scripta* centrale se révèle particulièrement homogène : les témoins qui la représentent ne se signalent ni par des résultats extrêmement faibles ni par de très bons résultats. Les autres *scriptae* se caractérisent par une assez grande dispersion des résultats, qui ne sont donc pas tributaires d'elles seules.

[21] Pour parvenir à une représentation lisible, les informations du NCA relatives aux *scriptae* et aux dates ont été uniformisées. Les genres textuels ont été directement attribués par nos soins.

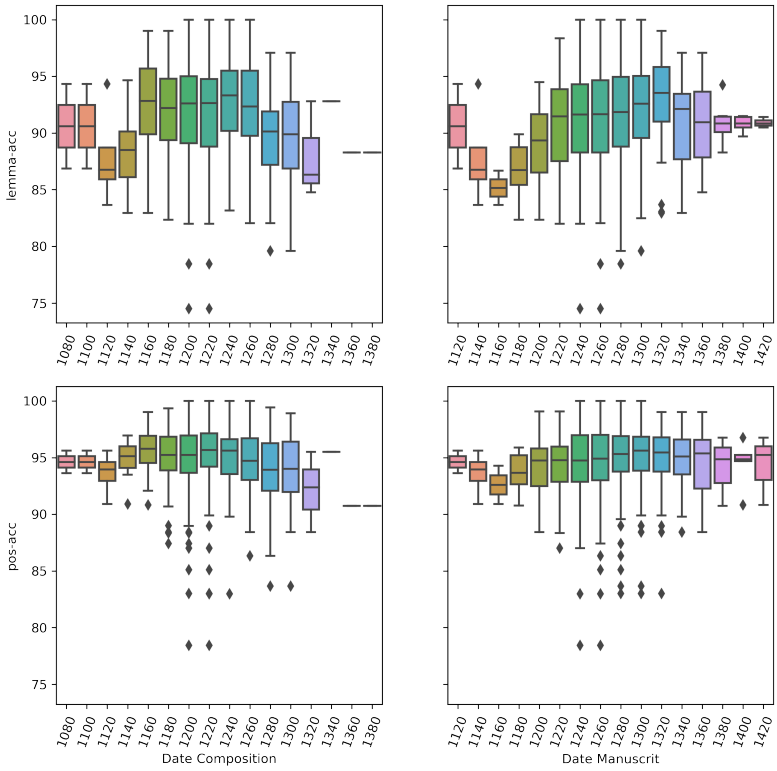

Figure 8 : Résultats sur les lemmes et POS en fonction des genres textuels

Le dernier graphique retenu ici représente les résultats du modèle en fonction des dates[22]. Si la variation chronologique est presque insensible pour les POS, systématiquement plus robustes que les lemmes, les résultats de ces derniers, en accord avec nos données d'entraînement, ont une tendance à la progression jusqu'au milieu du 13e siècle (date des textes), avant de décliner.

Attardons-nous pour finir sur les témoins dont les résultats de lemmatisation sont les plus faibles :

[22] De la même manière que les *scriptae*, les dates ont été sommairement réduites à des catégories homogènes.

sigle du témoin[23]	***scripta* (texte)**	**date de composition**	**résultat**
BestGuillR	ouest	1211	0.75
Turp1M	sudouest	1217	0.78
PerNeslesTabJ	pic.	1289	0.80
Turpin1A	sudouest	1217	0.82
MédLiégH	wall.	1275	0.82

Figure 9 : Les témoins aux résultats les moins bons

La présence, parmi ces cinq textes, de deux des trois seuls œuvres composées dans une *scripta* du Sud-Ouest parle pour elle-même. Pour le BestGuillR (miracle dramatisé en vers), PerNeslesTabJ (table des matières rimée), et MédLiégH (recueil de recette médicinales, récits de songes, etc.), c'est clairement le genre atypique de ces textes qui semble en cause.

4. Conclusion

L'évaluation du modèle *out-of-domain* a permis de repérer les points faibles du modèle. Notons qu'aucune variable n'explique à elle seule les résultats faibles du modèle, mais que chacune d'elle y contribue. Des tendances ont pu être observées, selon les types de *scriptae*, les datations et surtout les genres textuels, qui semblent tirer les résultats vers le haut ou contraire contribuer à leur affaiblissement. Ces observations montrent à quel point un modèle d'apprentissage profond est dépendant des données sur lequel il a été entraîné. Pour parvenir à obtenir un modèle hautement performant sur des textes variés, il faut que cette hétérogénéité soit présente dans les témoins utilisés. On notera tout de même la robustesse de la prédiction des parties du discours, qui semble moins affectées – si elles le sont même – par ces variations que les lemmes.

Plus généralement, on notera pour les lemmes que la modélisation par Pie (Manjavacas et al. , 2019) de la variation graphique semble porter ses fruits : la variation de *scripta* n'importe pas tant que celle de genre textuel, la chronologie se situant entre les deux. Finalement, c'est la variation de lexèmes plus que de graphies qui affecte les performances du lemmatiseur.

5. Recherches futures

Le présent article visait surtout à présenter à la communauté scientifique différents outils développés à l'École des chartes : le modèle *Old French* pour étiqueter les textes, une plateforme de correction de l'annotation linguistique, non dépendant de la langue, un corpus anno-

[23] Les sigles sont ceux de Frankwakt Möhren, *Dictionnaire étymologique de l'ancien français, Complément bibliographique* (DEAFBiblEl), <www.deaf-page.de/fr/bibl_neu.php>. Le sigle ElesB_T réfère à témoin manuscrit particulier, le témoin T tel que défini dans l'étude de la tradition manuscrite.

té librement réutilisable. Il marque un point d'étape, car les pistes à suivre demeurent nombreuses. Dans l'immédiat, il faudrait entraîner le modèle avec les textes nouvellement annotés et réévaluer le modèle en comparant les résultats avec ceux présentés ici. Cette nouvelle évaluation sera l'occasion d'explorer la piste du *fine-tuning*, qui permet d'ajuster le modèle à un type de données textuelles particulier, à partir d'un faible ajout de données étiquetées. Autrement dit, l'ajout au modèle de l'étiquetage de quelques centaines de vers du *Roman de la Rose* suffit-il à améliorer sensiblement les résultats sur l'ensemble du roman ? À plus long terme, il faudrait enrichir le corpus avec des données textuelles permettant de résorber les failles repérées dans le modèle, notamment en étiquetant des témoins relevant de genres et de *scriptae* qui lui sont mal connues ou inconnues. L'augmentation du nombre de textes étiquetés morphologiquement est également nécessaire. Se pose également la question de produire un modèle sur le moyen français, afin de rendre possibles des enquêtes en diachronie longue, puisque les outils proposent déjà un modèle pour le français classique et qu'un autre (« Français Classique (non modernisé) »), est entraîné sur un corpus allant du 16e au 18e siècle[24].

École nationale des chartes-PSL

Frédéric DUVAL*
Lucence ING*
Jean-Baptiste CAMPS
Naomia KANAOKA
Ariane PINCHE
Thibault CLÉRICE

5. Références bibliographiques

Camps, Jean-Baptiste, 2016. *La* Chanson d'Otinel *: édition complète du corpus manuscrit et prolégomènes à l'édition critique*. Thèse de doctorat, dir. Dominique Boutet, Paris, Paris-Sorbonne, <halshs.archives-ouvertes.fr/tel-016649>.

Camps, Jean-Baptiste (ed.), 2016-. *Geste : un corpus de chansons de geste*, version 02 d'avril 2019, Paris, École nationale des chartes, <doi.org/10.5281/zenodo.2630574>.

Camps, Jean-Baptiste / Gabay, Simon / Fièvre, Paul / Clérice, Thibault / Cafiero, Florian, 2021. « Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre », *Journal of Data Mining and Digital Humanities*, <doi.org/10.46298/jdmdh.6485>.

Camps, Jean-Baptiste / Clérice, Thibault / Duval, Frédéric / Ing, Lucence / Kanaoka, Naomi / Pinche, Ariane, 2021b. « Corpus and Models for Lemmatisation and POS-tagging of Old French », arXiv:2109.11442 [cs], 23 septembre 2021, http://arxiv.org/abs/2109.11442.

Clérice, Thibault / Pilla, Julien / Camps, Jean-Baptiste / Jolivet, Vincent / Pinche, Ariane, 2019. *Pyrrha, A language independant post correction app for POS and lemmatization*. doi : 10.5281/zenodo.2325427.

Duval, Frédéric (dir.), 2007-. *Miroir des classiques*, Paris, École nationale des chartes, <elec.enc.sorbonne.fr/miroir_des_classiques>.

Gabay, Simon / Clérice, Thibault / Camps, Jean-Baptiste / Tanguy, Jean-Baptiste / Gille-Levenson, Matthias, 2020. « Standardizing Linguistic Data: Method and Tools for Annotating (pre-

[24] Gabay, Clérice *et al.* (2020) et Camps, Gabay *et al.* (2021).

* Auteurs principaux.

orthographic) French », in : *Proceedings of the 2nd International Conference on Digital Tools & Uses Congress*, New York, Association for Computing Machinery, <doi.org/10.1145/3423603.3423996>.

Guillot, Céline / Prévost, Sophie / Lavrentiev, Alexei, 2013. *Manuel de référence du jeu Cattex09*, Lyon, École normale supérieure de Lyon. <bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf. Version 2.0 – 8 avril 2013>.

Ing, Lucence, en cours. *Disparitions lexicales en diachronie : traitements automatiques sur le* Lancelot *en prose*. Thèse de doctorat, dir. Frédéric Duval et Jean-Baptiste Camps, Paris, École nationale des chartes-PSL.

Manjavacas, Enrique / Kádár, Ákos / Kestemont, Mike, 2019. « Improving Lemmatization of Non-Standard Languages with Joint Learning », in : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologie*s, Minneapolis, Association for Computational Linguistics, Volume 1 (Long and Short Papers), p. 1493-1503, <doi.org/10.18653/v1/N19-1153>.

Pinche, Ariane, 2021. *Édition nativement numérique du recueil hagiographique* Li Seint Confessor *de Wauchier de Denain d'après le manuscrit fr. 412 de la Bibliothèque nationale de France.* Thèse de doctorat, dir. Corinne Pierreville et Bruno Bureau, Lyon, Université Jean-Moulin.

Stein, Achim / Kunstmann, Pierre / Gleßgen, Martin-D. (ed.), 2006. *Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350)*, établi par Anthonij Dees (Amsterdam 1987), Stuttgart, Institut für Linguistik/Romanistik, <stella.atilf.fr/gsouvay/nca/>.