HAL
open science

Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison : à propos de l'étude stylistique des lettres de Pierre Damien

Valérie Thon, Laurent Vanni, Dominique Longrée

To cite this version:

Valérie Thon, Laurent Vanni, Dominique Longrée. Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison : à propos de l'étude stylistique des lettres de Pierre Damien. JADT 2022 - Proceedings of the 16th International Conference on Statisical Analysis of Textual Data, 2022. hal-03892792

HAL Id: hal-03892792

https://u-paris.hal.science/hal-03892792

Submitted on 10 Dec 2022

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.

Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison : à propos de l'étude stylistique des lettres de Pierre Damien

Valérie Thon[1], Laurent Vanni[2], Dominique Longrée[3]

[1]Université de Paris – valerie.thon@u-paris.fr

[2]Université Côte d'Azur – laurent.vanni@unice.fr

[3]Université de Liège – dominique.longree@uliege.be

Abstract

To carry out a complete and reliable morphosyntactic labeling of Latin texts is a particularly time-consuming task. It is therefore necessary to choose wisely the texts to be included in a labelled comparison corpus when one wishes to study the intertextual distances between a given author, in particular a medieval one, and his predecessors. A stylistic research on the letters of Peter Damian (11th century) was the occasion to question the methods to be implemented to operate this selection. The intertextual distances were first computed on the forms using additive tree analysis. The results were then compared to the predictions of the deep learning, attributing with variable recognition rates passages of Damian to various authors of the comparison corpus. Where ADT relies primarily on the lexicon, the Convolutional Neural Network takes into account morphosyntactic parameters, with strong areas of activation suggesting a recognition of linguistic patterns that Damian shares with some of his predecessors.

Keywords: ADT, deep learning, prediction, lemmatization, morphosyntax, Peter Damian

Résumé

Réaliser un étiquetage morphosyntaxique complet et fiable de textes latins est une tâche particulièrement chronophage. Il s'agit dès lors de choisir à bon escient les textes à intégrer à un corpus de comparaison étiqueté lorsque l'on désire étudier les distances intertextuelles entre un auteur donné, en particulier un auteur médiéval, et ses devanciers. Une recherche stylistique sur les lettres de Pierre Damien (XIe siècle) a été l'occasion de s'interroger sur les méthodes à mettre en œuvre pour opérer cette sélection : les distances intertextuelles ont été d'abord calculée sur les formes à l'aide d'analyses arborées ; les résultats ont été ensuite comparés aux prédictions du deep learning, attribuant, avec des taux de reconnaissance variables, des passages de Pierre Damien à divers auteurs du corpus de comparaison : là où l'ADT semble s'appuyer essentiellement sur le lexique, le Convolutional Neural Network prend mieux en compte des paramètres morphosyntaxiques, les zones d'activation fortes suggérant une reconnaissance de motifs linguistiques que Damien partagerait avec certains de ses prédécesseurs.

Mots clés : ADT, deep learning, prédiction, lemmatisation, morphosyntaxe, Pierre Damien

1. Introduction

Le deep learning offre-t-il des instruments pour compléter les méthodes de l'ADT lorsqu'il s'agit d'établir un corpus de comparaison en vue d'une analyse stylistique ? Nous tenterons de répondre à cette question dans le cadre d'une étude textométrique plus large de la langue et du style de Pierre Damien (1007-1072/73), ermite 'réformateur' du Moyen Âge central, auteur de 180 lettres latines. Dans le cadre du projet doctoral de V. Thon, 12 d'entre elles (33.396 occurrences), couvrant la totalité de sa carrière ecclésiastique, ont été traitées à cette fin selon les méthodes du LASLA (lemmatisation et analyse morphosyntaxique semi-automatique, avec

une vérification manuelle complète). Complétant les méthodes de la stylistique traditionnelle, ce traitement permet non seulement une analyse textométrique fine de l'évolution de la langue et du style de l'auteur à travers sa carrière, mais aussi de le confronter à ses devanciers pour identifier ceux qui auraient influencé sa technique d'écriture. La mesure de cette 'influence diachronique' implique d'aller au-delà du domaine lexical, trop sensible à la thématique des divers textes, et d'examiner les caractéristiques morphosyntaxiques qui présentent une plus grande indépendance vis-à-vis de celle-ci. Pour ce faire, il s'agit de disposer d'un corpus de comparaison lemmatisé et étiqueté morphosyntaxiquement. Pour la période classique, la banque de textes du LASLA fournit un corpus étiqueté important, mais qui ne couvre pas l'ensemble des textes susceptibles d'avoir influencé Pierre Damien. Il est également nécessaire de pouvoir confronter ses lettres aux écrivains tardo-antiques ou médiévaux, tels les Pères de l'Église ou les auteurs carolingiens. Or ce corpus « tardif » est de taille trop considérable pour être traité intégralement. Le défi principal est dès lors d'identifier –sans étiquetage préalable– les auteurs tardo-antiques et médiévaux avec lesquels Pierre Damien présente des affinités et qui seraient ainsi des candidats par excellence à un étiquetage fin. Pour ce faire, nous proposons de croiser ADT et deep learning (module de deep learning – CNN fondé sur Hyperbase Web) sur un grand corpus de textes non-traités (40 auteurs au total).

2. Les méthodes du LASLA : fiables, mais chronophages

Le traitement des textes latins avec les méthodes du LASLA consiste à procéder à une lemmatisation et à une analyse morphosyntaxique complète de chacune des formes rencontrées, en utilisant une procédure semi-automatique : les analyses que propose le logiciel font l'objet d'une sélection et d'une vérification systématique par un philologue confirmé qui peut compléter ou corriger les informations fournies. Ce travail a le mérite d'assurer une très grande fiabilité aux données contenues dans les fichiers du LASLA, mais, même si cette opération se fait en ligne grâce à l'interface LEI (LASLA Encoding Initiative), celle-ci requiert un temps de travail important (de l'ordre de 100 à 150 heures pour 20.000 mots). Le travail est d'autant plus chronophage quand le texte comprend des termes tardo-antiques ou médiévaux inconnus du dictionnaire classique. En collaboration avec le projet Collatinus, le LASLA a pu trouver une alternative à la procédure LEI : en s'appuyant sur un dictionnaire enrichi, Collatinus-LASLA permet de traiter un peu plus rapidement les textes, mais sans pour autant éliminer une intervention manuelle importante (Verkerk *et alii*, 2020). Le LASLA s'est par ailleurs tourné à diverses reprises vers l'utilisation de lemmatiseurs ou étiqueteurs complétement automatiques, mais malheureusement sans arriver à une fiabilité suffisante pour une étude stylométrique fine (Longree et Poudat, 2010). Dans ce contexte, il s'agit donc de choisir avec circonspection les textes à retenir pour constituer le ou les corpus de comparaison, textes devant faire l'objet d'un traitement avec la procédure LEI. Cette sélection peut heureusement s'appuyer sur des analyses préalables portant sur des textes non lemmatisés. En effet, le latin étant une langue flexionnelle, les formes contiennent dans leurs terminaisons une information morphosyntaxique importante que les méthodes de l'ADT peuvent prendre en compte, mais seulement jusqu'à un certain point. Le recours au Deep Learning permet, comme on va le voir, d'aller un peu plus loin.

3. Pour une confrontation méthodologique : ADT et deep learning

3.1. Méthodologie : la constitution du corpus de comparaison

Pour étudier le style de Pierre Damien, nous avons constitué un corpus de 40 auteurs susceptibles de l'avoir influencé. En nous appuyant sur les informations fournies dans l'édition

critique de Reindel (1983-1993) et sur le catalogue (fin XIe – début XIIe siècle) de la bibliothèque de Fonte Avellana, monastère où Pierre Damien fut prieur à partir de 1043 (Martini, 2002), nous avons veillé à ce que le corpus comprenne à la fois des auteurs que Pierre Damien a certainement connus (voire lus), et des auteurs avec lesquels aucun lien particulier n'est supposé a priori. Ceux-ci sont répartis de façon équitable sur trois périodes, correspondant à 3 modèles *Hyperdeep* entraînés (Tableau 1) : Antiquité (« Période 1 »), Antiquité tardive et haut Moyen Âge (« Période 2 ») et Moyen Âge central jusqu'à l'époque de Pierre Damien (« Période 3 »).

	Nombre d'occurrences	Nombre d'auteurs	Précision du modèle *Hyperdeep*
Période 1 (IIIe av. – I^e apr.)	1.001.582	12	98.05%
Période 2 (IIe apr. – VIIIe apr.)	1.820.007	15	95.78%
Période 3 (IXe apr. – XIe apr.)	951.511	13	95.56%

Tableau 1 : caractéristiques du corpus de comparaison divisé en 3 périodes

Cette sélection obéit à un certain nombre de contraintes : au sein d'une même période, les textes retenus présentent des tailles approximativement identiques, pour éviter que des tailles trop différentes puissent biaiser les résultats ; seuls des textes comprenant plus de 30.000 mots ont été retenus ; chaque modèle entraîné ne dépasse pas la limite de 20 classes. Pour construire un corpus riche et varié, les œuvres choisies se distinguent les unes des autres non seulement d'un point de vue chronologique, mais également par le genre et le style : les textes de comparaison sont donc théologiques, ecclésiastiques, épistolaires, rhétoriques, 'romanesques', historiographiques, philosophiques.... Nous avons toutefois écarté pour le moment tout texte poétique du corpus, parce que la structuration linguistique plus libre des textes poétiques par rapports aux textes prosaïques risquait d'altérer les résultats du deep learning. Enfin, nous avons harmonisé les graphies, ponctuations et mises en pages (les ponctuations faibles ont été supprimées et toutes les ponctuations fortes homogénéisées sous forme de points).

3.2. Les résultats de l'ADT

Dans un premier temps, nous avons soumis les trois corpus de comparaison à un calcul de distances intertextuelles. Dans le cadre imparti ici, nous ne pourrons présenter que les résultats de la « Période 2 ». Les distances ont été successivement calculées selon la méthode Brunet (fréquence des formes) puis la méthode Jaccard (présence/absence des formes) telles qu'implémentées dans Hyperbase Web Edition (avec représentation arborée d'après le procédé de regroupement Neighbour Joining appliquée à une représentation de type radial).

La première méthode, celle de Brunet (Figure 1), rapproche Pierre Damien de Jérôme (IXe – Ve siècle) et de Grégoire le Grand (VIe – VIIe siècle). C'est avant tout le genre pastoral qui prédomine chez eux : Jérôme et Grégoire ont tous les deux rédigé un commentaire sur les prophéties bibliques d'Ézéchiel, textes inclus dans le corpus, et Grégoire est également l'auteur d'une Règle Pastorale, destinée à montrer aux évêques de son temps comment ils doivent guider les fidèles dont ils ont la charge. La proximité de Pierre Damien avec ceux-ci n'est pas surprenante. Celui-ci fait figure de maître spirituel à l'esprit pastoral, veillant à ce que ses moines s'orientent toujours vers Dieu, et n'hésitant pas à adresser divers conseils à ses contemporains, hommes et femmes, clercs et laïcs.

La méthode « Jaccard » (Figure 2) situe les lettres de Pierre Damien sur une branche qui regroupe Tertullien (IIe – IIIe siècle), Hilaire (IVe siècle) et Alcuin (VIIIe – IXe siècle). Les textes

Figure 1 : calcul des distances Brunet

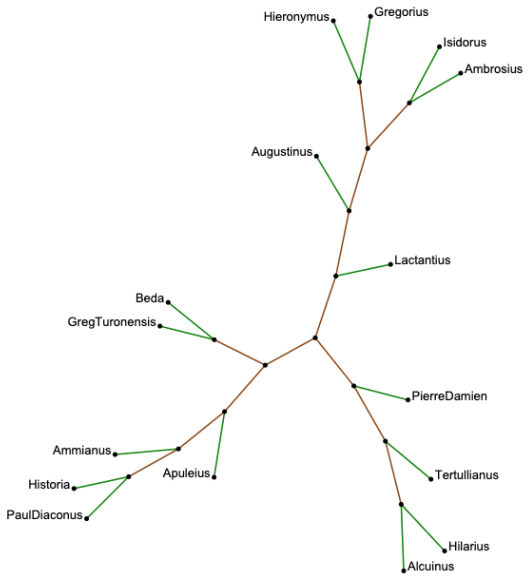

Figure 2 : calcul des distances Jaccard

de ces trois auteurs s'organisent tous autour de l'apologie : Tertullien est représenté par ses *Prescriptions contre les hérétiques*, son *Apologétique* (sa grande défense de la religion chrétienne) et son *De l'idolâtrie* ; Hilaire est présent dans le corpus par son *De Trinitate*, défense du dogme chrétien de la trinité ; Alcuin est essentiellement représenté par un traité voué au même sujet (*De fide trinitatis*). Le fait que Pierre Damien se range parmi ces auteurs pourrait de prime abord ne pas surprendre : défenseur de l'idéologie « grégorienne », il lance régulièrement des attaques contre toute pratique à son avis inacceptable. Mais ce déplacement thématique d'une méthode à l'autre pose néanmoins question. Or les textes rapprochés par la méthode Jaccard ont un autre point commun : les textes pris en compte pour Tertullien, Alcuin et Pierre Damien présentent une taille plus réduite que les autres textes du corpus de la période 2. Les résultats obtenus peuvent donc aussi s'expliquer par la grande sensibilité de la méthode Jaccard à la taille des textes : ce ne sont donc pas nécessairement les apologètes qui se rapprochent de Damien en raison d'un vocabulaire partagé, mais plutôt les autres textes qui repoussent tant les apologètes que Damien de par des absences lexicales liées à la faible taille de leurs partitions respectives. Il n'en reste pas moins que, dans les deux méthodes les spécificités lexicales jouent un rôle déterminant, alors que le rôle de la morphosyntaxe est quant à lui totalement occulté. Or des travaux antérieurs ont bien mis en évidence (Vanni et al., 2018a et 2018b) la sensibilité du deep learning à la morphosyntaxe. Nous avons donc vérifié s'il en était de même ici.

3.3. Les résultats du deep learning

Les 12 lettres de Pierre Damien faisant partie de la thèse de V. Thon, –supposées former un échantillon représentatif–, ont d'abord été proposées une par une au modèle entraîné sur le corpus « Période 2 », en lui donnant comme tâche d'attribuer ces lettres à un des auteurs du corpus. Afin de vérifier la stabilité des résultats sur une population plus large, nous avons ensuite procédé de même avec la quasi-totalité de ses lettres en un seul fichier (édition Migne ; 158 lettres étaient connues à l'époque de cette édition). Les 12 lettres ont été souvent attribuées en premier lieu à Bède le Vénérable (VIIe – VIIIe siècle), avec des taux de reconnaissance particulièrement élevés, puis à Grégoire le Grand, proche de Damien avec la méthode Brunet.

Ambroise (IVe siècle) et Augustin (IVe – Ve siècle) sont également susceptibles d'avoir un taux important, mais sont globalement moins présents. Le résultat pour l'ensemble des 158 lettres semble confirmer ce constat : Grégoire le Grand (23%) y occupe la première position, suivi de près par Bède (22%) et Ambroise (13%) ; les autres auteurs du corpus de comparaison ne s'y élèvent pas au-dessus de 10%.

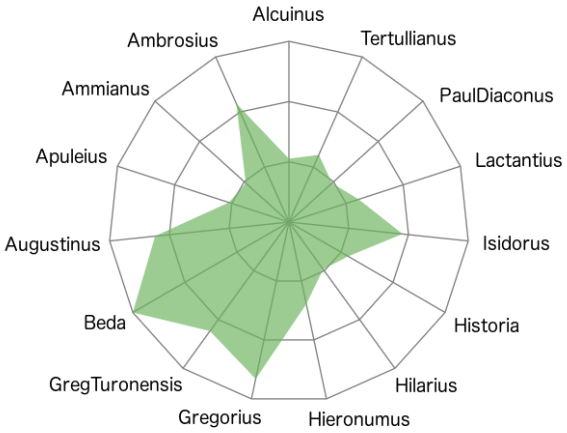

Figure 3 : attribution de la Lettre 114

L'attribution des textes de Pierre Damien à Grégoire, Ambroise ou Augustin n'est pas nécessairement étonnante, étant donné que ces trois derniers sont tous reconnus comme de grands Pères de l'Église dont les écrits ont exercé une influence indéniable sur tout le Moyen Âge latin. Le lien avec Bède, en revanche, est plus surprenant, d'autant qu'il n'est pas suggéré par les arborées de l'ADT. Il est représenté dans le corpus par deux textes plutôt historiques : *Histoire ecclésiastique du peuple anglais* et *De la raison du temps*. La reconnaissance semble s'appuyer non seulement sur l'idée d'un vocabulaire partagé (le champ sémantique du religieux et du temporel, par exemple, est présent chez tous les deux), mais également sur des structures d'ordre morphosyntaxique. Remarquons en effet que le deep learning semble avoir une sensibilité à la morphosyntaxe. Parmi les motifs linguistiques repérés pour Bède, par exemple, nous trouvons régulièrement des expressions du type « *qui nimirum, qui dudum, qua uidelicet, quae nimirum, quae statim, cui uidelicet,…* ». La figure 4 fournit un exemple des taux d'activation important mis en évidence pour la séquence *cui uidelicet ad corrigendum* grâce au calcul du TDS (Vanni et al. 2018b).

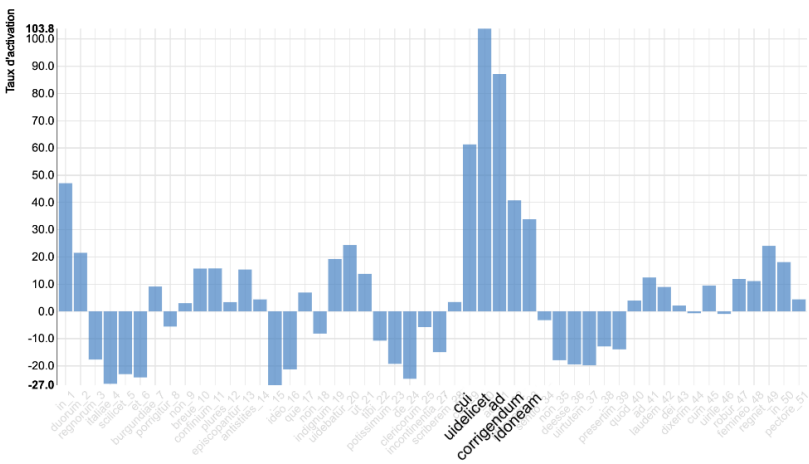

Figure 4 : Exemple de motifs repéré par le deep learning de la Lettre 114

Une recherche sur la distribution des deux expressions *qui uidelicet* et *cui uidelicet* dans le corpus « Période 2 » montre que celles-ci sont fortement spécifiques de Bède. Le deep learning aurait-il reconnu plus largement au sein des lettres de Pierre Damien une structure syntaxique « pronom relatif + adverbe », qui serait également caractéristique du style de Bède ? Lemmatiser un ou plusieurs textes de Bède permettrait de vérifier cette hypothèse par les méthodes éprouvées de l'ADT.

4. Évaluation

Cette recherche visait à croiser ADT et méthode prédictive du réseau de neurones dans le cadre d'une étude textométrique du corpus épistolaire de Pierre Damien. L'objectif était de détecter d'éventuels modèles stylistiques parmi les textes de ses devanciers, en vue d'identifier des candidats à une lemmatisation semi-automatique fine. D'une part, l'ADT a perçu un lien étroit entre Pierre Damien et le genre pastoral (Grégoire le Grand et Jérôme), mais suggérait également un rapport avec l'apologie (Tertullien, Hilaire, Alcuin), alors que celui-ci est essentiellement dû à la trop grande sensibilité de la méthode Jaccard à la taille des textes et donc aux absences de formes. De son côté, le deep learning, ne prend en compte lui que des présences et surmonte cette difficulté. Ainsi, il a identifié une proximité non seulement avec Grégoire le Grand, mais aussi avec Bède le Vénérable. Ce résultat était inattendu : les zones d'activations suggèrent que l'attribution s'appuie non seulement sur des critères lexicaux, mais aussi morphosyntaxiques, le CNN semblant capable d'identifier des catégories grammaticales comme celles du relatif ou de l'adverbe, en ne s'appuyant que sur les formes seules. Cette recherche montre une nouvelle fois tout l'intérêt d'un dialogue entre les méthodes, dialogue qui pourra se prolonger une fois les textes de Grégoire et de Bède ayant été lemmatisés et étiquetés pour permettre une analyse textométrique fine.

Références

Brunet, É. et Vanni, L. (2019). *Deep learning* et authentification des textes. *Texto ! Textes et cultures*, vol. 24.1 : 1-34.

Longrée, D. et Poudat, C. (2010). New Ways of Lemmatizing and Tagging Classical and post-Classical Latin: the LATLEM project of the LASLA. In P., Anreiter et M., Kienpointner (Eds.), *Proceedings of the 15th International Colloquium on Latin Linguistics. Innsbruck.*

Martini, P. S. (2002). L'inventario del secolo XII della biblioteca di Santa Croce di Fonte Avellana. In Gatto L. et Martini P. S. éditeurs, *Studi sulle società e le culture del Medioevo per Girolamo Arnaldi*, All'Insegna del Giglio.

Reindel, K. (éditeur, 1983-1993). *Die Briefe des Petrus Damiani. Teil 1 – 4*. Monumenta Germaniae Historica, Die Briefe der deutschen Kaiserzeit.

Vanni, L, Mayaffre, D. et Longrée, D. (2018a). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables. *JADT 2018* : 459-466.

Vanni, L., Ducoffre, M., Mayaffre, D., Precioso, F., Longrée, D., Elango, V., Santos, N., Gonzalez, J., Galdo, L. et Aguilar, C. (2018b). Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) : 548-557.

Verkerk, P., Ouvrard, P., Fantoli, M. et Longrée, D. (2020). L.A.S.L.A. and Collatinus: a convergence in lexica. In L., Tesconi (Ed.), Studi e saggi linguistici 2020(1) (pp. 1-26). Edizioni ETS.