Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, Serge Heiden

▶ **To cite this version:**

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev, Serge Heiden. Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. Corpus, 2024, La constitution de corpus en diachronie longue. Méthodologies, objectifs et exploitations linguistiques et stylistiques, 25, pp.8538. 10.4000/corpus.8538 . hal-04681591

HAL Id: hal-04681591
https://hal.science/hal-04681591

Submitted on 29 Aug 2024

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.

CORPUS

Corpus

25 | 2024
La constitution de corpus en diachronie longue

Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval

The Profiterole corpus of parsed Medieval French

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev and Serge Heiden

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/8538
DOI: 10.4000/corpus.8538
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Provided by Ecole Normale Supérieure Paris

ISTEX
L'excellence documentaire pour tous

Electronic reference
Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev and Serge Heiden, “Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval”, *Corpus* [Online], 25 | 2024, Online since 30 January 2024, connection on 29 August 2024. URL: http://journals.openedition.org/corpus/8538 ; DOI: https://doi.org/10.4000/corpus.8538

This text was automatically generated on February 1, 2024.

The text and other elements (illustrations, imported files) are “All rights reserved”, unless otherwise stated.

Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval

The Profiterole corpus of parsed Medieval French

Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev and Serge Heiden

Introduction

1 Alors que la plupart des changements morpho-syntaxiques et syntaxiques du français se sont déroulés durant la période médiévale (9e-15e s.), la période du moyen français n'était jusqu'ici guère équipée en données enrichies morpho-syntaxiquement et syntaxiquement, et celle de l'ancien français l'était encore insuffisamment, limitant de fait les études sur de vastes ensembles de données enrichies. La constitution du corpus Profiterole[1] a permis de remédier à cette situation, ouvrant un possible renouveau des études, qui pourront désormais se fonder sur une masse de données significative. La présente contribution s'articulera en trois points. Dans une première section, après avoir rappelé le contexte et les enjeux de la création du corpus Profiterole, nous présenterons les modalités de sa constitution pour ce qui concerne le choix raisonné des textes et leur intégration au corpus. Dans une seconde section, nous présenterons les modalités d'annotation du corpus en parties du discours et en dépendances syntaxiques, ainsi que le recours subséquent à un algorithme de vote appliqué aux prédictions de plusieurs modèles d'analyses syntaxiques automatiques pour faciliter les procédures de correction. Enfin, dans une dernière section nous évoquerons la distribution du corpus et les modalités de son exploration/interrogation via la plateforme TXM.

1. Présentation générale

1.1. Contexte de constitution du corpus Profiterole

2 Le corpus Profiterole a été conçu dans le cadre du projet ANR du même nom. Ce projet avait un triple objectif : il s'agissait d'une part de constituer des ressources pour le français médiéval (un corpus annoté et des lexiques), d'autre part de concevoir des analyseurs syntaxiques pour le français médiéval (mais réutilisables pour d'autres états de langue), et, enfin, d'esquisser la modélisation de certains aspects (morpho-)syntaxiques de l'évolution du français. C'est sur la constitution du corpus et sur le processus de *parsing* (et donc aussi sur les analyseurs syntaxiques) que nous concentrons notre attention dans cet article.

1.2. Motivations du projet

3 Le français médiéval, qui s'étend du 9e au 15e siècle et que l'on a coutume de subdiviser entre ancien français (9e-13e) et moyen français (14e-15e)[2], constitue une période décisive pour les changements morpho-syntaxiques et syntaxiques qui se sont produits en français, en particulier pour ce qui touche aux changements dans le domaine de la morphologie verbale, à la disparition progressive de la déclinaison casuelle, ou bien encore à la fixation de l'ordre des constituants majeurs. Il serait erroné de considérer que l'état de langue moderne est acquis à la fin du 15e s. : la langue continuera évidemment d'évoluer au-delà (ainsi, par exemple, de la montée du clitique 'je le veux faire', qui perdurera jusqu'au 17e s., ou des structures de l'interrogation qui connaîtront encore différentes transformations). Il n'en demeure pas moins que ces quelques siècles voient se produire des changements essentiels bien plus nombreux que ceux qui adviendront dans les cinq siècles suivants.

4 Si les linguistes n'ont pas attendu les corpus enrichis linguistiquement (et les outils adéquats pour les exploiter) pour mettre au jour et analyser les changements linguistiques, tous sont désormais convaincus que la mise à disposition de données massives et diversifiées, exploitables autrement que « à la main », contribue largement à un renouvellement des études, dans la mesure où ces données permettent de confirmer, ou non, avec davantage de certitude, les changements déjà identifiés, et de préciser leurs chronologies et leur corrélation éventuelle. La variation diachronique, à laquelle on peut ramener le changement linguistique, ne peut, pour être rigoureuse, se passer d'une quantification des données, même si celle-ci n'est que le point de départ d'analyses qualitatives.

5 Or le français manquait jusqu'ici de données médiévales enrichies morpho-syntaxiquement et syntaxiquement, même s'il existe déjà certaines ressources. La *Base de Français Médiéval* (9-15e s.) comprend 7,3 millions de mots, dont 1,1 million sont enrichis avec des étiquettes morpho-syntaxiques vérifiées, et 700 000 le sont avec des lemmes vérifiés. Mais elle ne comprend pas d'annotation syntaxique. Il existe par ailleurs le corpus SRCMF[3], qui comprend 250 000 mots, enrichis en morpho-syntaxe et en syntaxe (avec des étiquettes vérifiées), selon un modèle dépendanciel propre[4]. Une partie de ce corpus (170 000 mots) a été convertie au schéma d'annotation UD[5]. Aussi utile que soit cette ressource, elle reste néanmoins d'une taille limitée et ne contient pas de données de moyen français. Il existe enfin le Corpus MCVF-PPCHF *Modéliser le*

changement : les Voies du Français & Penn-BFM Parsed Corpus of Historical French[6], corpus enrichi selon un modèle d'annotation syntaxique en constituants (d'inspiration générativiste), d'un usage désormais plus limité dans la communauté. Ce corpus contient 411 000 mots pour l'ancien français et 760 000 pour le moyen français, qui correspondent respectivement à 20 et à 8 textes intégraux, le choix ayant été fait d'inclure des textes intégraux, ce qui, les forces d'annotation n'étant pas illimitées, a réduit la diversité des textes.

6 Le corpus Profiterole est donc venu combler un vide relatif en matière de données de français médiéval enrichies, sans pour autant, évidemment, constituer une fin en soi.

1.3. Choix des textes

7 Le corpus Profiterole comprend un million de mots et 63 textes ou extraits de textes, qui se répartissent comme suit.

Tableau 1. Répartition des données du corpus Profiterole

Siècle	9e-12e	13e	14e	15e
Nombre de textes	19	15	15	14
Nombre de mots	280 369	244 750	253 595	241 282

8 Le choix des textes a répondu à plusieurs critères, d'ordres différents. En premier lieu, tous les textes sont issus de la *Base de Français Médiéval*, qui contient des éditions fiables, libres de droit, et dont certains des textes étaient déjà enrichis en étiquettes morpho-syntaxiques et en lemmes. Le corpus SRCMF a été intégré au nouveau corpus Profiterole.

9 Il s'agit d'un corpus équilibré sur le plan quantitatif, puisqu'il contient environ 250 000 mots par siècle ou 500 000 mots par période (ancien ou moyen français)[7], le total de 1 million de mots nous semblant être un compromis raisonnable entre les exigences de représentativité et les contraintes de faisabilité, en particulier pour ce qui est de la vérification des annotations. Il s'agit en outre d'un corpus diversifié sur le plan qualitatif. Pour ce qui est de la forme, il contient 53 % de textes en prose, 40 % de textes en vers et 7 % de textes mixtes, avec, nécessairement, une proportion plus élevée de textes en vers dans la période la plus ancienne, puisque la prose en langue française ne commence à véritablement se développer qu'au 13e s. Pour ce qui est du domaine, les textes littéraires représentent 43 % de l'ensemble, les textes religieux 18 %, les textes historiques 18 %, les textes didactiques 17 %, et les textes juridiques 4 %[8]. La représentation est en revanche plus inégale en ce qui concerne le dialecte, avec une prévalence de l'anglo-normand, du normand, du champenois et du picard (un tiers des textes ne relevant pas d'un dialecte particulier), répartition qui correspond pour une large part à la réalité des données qui nous sont parvenues. Dans une perspective de mutualisation des données, il a aussi été donné priorité à l'intégration de textes présents dans d'autres projets (Corpus PaLaFra[9], DEMOCRAT[10]). Le choix a été fait d'échantillonner à 40 000 mots les textes dépassant cette taille, afin d'avoir davantage de textes (ou extraits) et donc une diversité plus grande[11]. Sans prétendre être un

parfait reflet de la langue médiévale, le corpus Profiterole en offre néanmoins une représentativité inégalée jusqu'ici pour un corpus annoté en syntaxe.

1.4. Enrichissement morpho-syntaxique

10 Une partie seulement des textes ont actuellement des étiquettes morpho-syntaxiques et des lemmes totalement vérifiés, respectivement 39 et 21 des 63 textes du corpus. Il s'agit de textes dont l'étiquetage automatique avait été vérifié préalablement à la constitution du corpus Profiterole, dans le cadre de campagnes de correction de la BFM. Ces textes bénéficient d'un double étiquetage morpho-syntaxique : le jeu d'étiquettes Cattex, développé dans le cadre de la BFM[12], et le jeu d'étiquettes UD[13], obtenu par conversion du jeu Cattex. Le reste des étiquettes morpho-syntaxiques (étiquettes UD attribuées automatiquement en même temps que les annotations syntaxiques) et des lemmes sera corrigé ultérieurement.

11 Pour ce qui est de l'annotation syntaxique, le corpus a été annoté selon le modèle dépendanciel du projet *Universal Dependencies*[14]. L'annotation a été réalisée par quatre analyseurs syntaxiques, puis vérifiée/corrigée en partie par un annotateur humain, après le passage d'un algorithme de vote (voir 2.3). Dans la mesure où il s'agit d'une démarche reproductible, et qui a prouvé son intérêt et ses bénéfices, il nous semble intéressant de l'exposer ci-dessous de manière quelque peu détaillée.

2. Analyseurs syntaxiques et votes

2.1. Analyseur HoPS

12 HoPS (Grobol & Crabbé 2021) est un analyseur syntaxique (aussi appelé *parseur*) en dépendances, fonctionnant sur le principe d'analyse par arbre couvrant maximal[15] (McDonald *et al.* 2005). Il utilise l'architecture neuronale à attention bi-affine[16] de Dozat et Manning (2016). Ce type d'analyseur est particulièrement adapté au traitement de langues qui, comme l'ancien français, présentent un ordre des mots flexible, puisqu'il n'impose aucune contrainte sur la projectivité[17] des arbres prédits, et n'est pas directement influencé par la longueur des dépendances syntaxiques.

13 HoPS diffère de l'implémentation originale de Dozat et Manning en ce que, au lieu d'utiliser comme entrées des représentations statiques des formes de surface des mots et de leurs parties du discours, il utilise des représentations arbitraires des formes uniquement, et prédit les parties du discours au moment de l'analyse, ce qui permet de travailler sur des textes ne comportant pas d'annotations préalables. Par ailleurs, les performances obtenues sur la prédiction des parties du discours suggèrent qu'une grande partie de cette information serait superflue en entrée, puisque le système est capable d'apprendre à la prédire.

14 Étant donné le manque relatif de données utilisables pour l'ancien français, l'utilisation de représentations arbitraires pour HoPS permet de tirer parti du maximum d'informations disponibles : des représentations vectorielles au niveau des caractères (qui permettent de tenir compte dans une certaine mesure de la variation orthographique), des formes lexicales et des représentations de type FastText[18] (Bojanowski *et al.* 2017) entraînées directement avec l'analyseur lui-même, mais également des représentations contextuelles de type BERT pré-entraînées sur un

corpus d'ancien et de moyen français (Grobol *et al.* 2022). L'association de ces différentes représentations permet d'obtenir un analyseur robuste. Grobol, Prévost et Crabbé (2021) rapportent ainsi des scores avoisinant les 91 % de LAS (*Labelled Attachment Score*, exactitude de la prédiction conjointe des têtes et des types de dépendances syntaxiques) pour le modèle utilisé ici, entraîné sur le corpus SRCMF. D'un point de vue qualitatif, cette même analyse montre de plus des performances encourageantes pour une pré-annotation automatique, en particulier s'agissant d'ordres de mots non-canoniques.

2.2. Autres analyseurs

15 En plus de HoPS, présenté ci-dessus, trois autres analyseurs syntaxiques[19] ont été utilisés pour produire des arbres en dépendances. L'objectif dans l'utilisation de plusieurs parseurs est de comparer leurs prédictions, car différents modèles tendent à faire des erreurs différentes. Avant de nous intéresser à la question de la comparaison des prédictions des différents modèles, nous commençons par présenter les trois autres parseurs utilisés pour la pré-annotation du corpus Profiterole.

16 MetaMOF (Regnault 2019) est un analyseur syntaxique symbolique basé sur un ensemble de règles de réécriture d'arbres appelé « métagrammaire ». Il s'agit de l'adaptation au français médiéval du parseur FRMG (Villemonte de La Clergerie 2005) pour le français contemporain. Il a été développé par M. Regnault pendant sa thèse de doctorat dans le cadre du projet Profiterole.

17 DYALOG-SRNN (Villemonte de La Clergerie *et al.* 2017), est un analyseur syntaxique dit « à transitions » dont les scores de transitions sont calculés par un réseau neuronal récurrent à partir d'une représentation vectorielle des mots de la phrase et de l'état courant du parseur comprenant, entre autres, une représentation des sous-arbres les plus récemment produits ainsi que de la suite de transitions ayant mené à l'état courant du parseur.

18 DYALOG-SRNN/MetaMOF est un analyseur hybride partageant l'architecture du modèle DYALOG-SRNN mais prenant en entrée, en plus des représentations de la phrase et de l'état du parseur, une représentation de la structure proposée par MetaMOF – quand celui-ci en propose une – pour que le modèle statistique puisse aussi apprendre à se reposer sur les points forts de la métagrammaire.

19 Ces quatre analyseurs se sont appuyés à différents degrés, pour l'entraînement de leurs modèles, sur le corpus SRCMF, annoté en dépendances.

2.3. Votes entre analyseurs syntaxiques

20 Bien que les analyseurs syntaxiques deviennent de plus en plus performants (cf. section 2.1 et Grobol *et al.* 2021), ils restent toujours perfectibles et commettent encore des erreurs.

21 La principale difficulté rencontrée dans l'utilisation de parseurs pour assister les annotateurs humains est que, contrairement au cas de données déjà annotées utilisées pour entraîner et évaluer lesdits parseurs, dans le cas des données à annoter l'on ignore où se situent les erreurs.

22 Une manière d'estimer la position de ces erreurs est d'utiliser plusieurs analyseurs en parallèle.

23 Dans la littérature du traitement automatique du langage (TAL), des systèmes de vote ont été étudiés dans l'optique d'agréger les prédictions de plusieurs modèles, l'objectif étant d'obtenir de meilleurs résultats exploitant les forces de chaque modèle.

24 En effet, des parseurs construits sur des architectures différentes et entraînés séparément tendent à produire des erreurs différentes, alors qu'en général il n'y a qu'une analyse canonique ; de plus, en partant du principe qu'il n'y a qu'une analyse correcte, quand deux analyseurs sont en désaccord, on a la certitude qu'au moins l'un des deux se trompe. Ainsi, en comparant les prédictions de plusieurs analyseurs (en les faisant voter pour les dépendances qu'ils produisent), non seulement l'on peut proposer une pré-annotation plus robuste à l'annotateur humain, mais l'on peut également détecter les dépendances de ladite pré-annotation les plus susceptibles d'être erronées. C'est dans cette optique que nous avons fait voter quatre analyseurs syntaxiques pour produire des pré-annotations pour le corpus Profiterole.

25 Pour le vote, nous distinguons les dépendances (choix d'un gouverneur pour chaque mot) de leur relation (étiquette indiquant le type de relation existant entre un mot et son gouverneur), c'est-à-dire que pour chaque mot l'on effectue deux votes : un pour le choix du gouverneur et un pour le type de relation syntaxique qu'il entretient avec celui-ci.

26 Comme nous utilisons quatre analyseurs, il arrive que l'on ait une égalité parfaite entre plusieurs analyses (2 contre 2 ou 1 contre 1 contre 1 contre 1) ; dans ces cas nous donnons l'avantage à HoPS car il obtient de meilleurs résultats que les autres modèles sur le corpus de français médiéval SRCMF, dont Profiterole est une extension.

27 Face à la diversité des textes constitutifs du corpus Profiterole, tant en termes de période que de genre, les différents analyseurs ne sont peut-être pas aussi compétents sur tous les textes. Cependant, il est difficile de se faire une idée précise sur ce point (au-delà des désaccords entre analyseurs) avant d'avoir commencé à corriger la pré-annotation à la main. En revanche, une fois qu'une partie d'un texte a été corrigée, il devient possible de comparer les résultats des votes avec ladite correction et de potentiellement pondérer les votes des analyseurs. Nous appelons cette étape l'analyse des coalitions car, en plus des votes, l'on prend aussi en compte les lignes de désaccords et les types de désaccords.

28 Cette méthode itérative permet ainsi non seulement de proposer une pré-annotation à l'annotateur, mais également de lui montrer les pré-annotations les moins convaincantes ainsi que les propositions faites par les différents modèles en cas de désaccords. L'annotateur est ensuite libre de choisir une des pré-annotations proposées ou de ne suivre aucun des parseurs et de corriger la structure d'une manière différente.

29 Le corpus se subdivise donc actuellement en un corpus « Gold » (212 000 mots), dont la pré-annotation a été vérifiée et corrigée si nécessaire, et un corpus « Ore » (non « Gold »), en cours de vérification, et destiné à disparaître, à l'issue des vérifications, au profit du seul corpus Gold.

3. Diffusion et exploitation du corpus

3.1. Diffusion des données, des logiciels et de la documentation

30 Les données et logiciels produits par le projet Profiterole sont diffusés sous des licences ouvertes via plusieurs canaux. Le corpus peut être interrogé en ligne directement sur le portail de la Base de français médiéval (voir la section suivante). Ce même portail permet de télécharger le corpus au format « binaire » (.txm) pouvant être exploité avec le logiciel TXM pour poste (version 0.8.3 ou ultérieure) grâce à l'extension « Annotation syntaxique » développée pour le projet et installable depuis l'application (section 3.3).

31 Les fichiers annotés au format CoNLL-U sont disponibles dans l'entrepôt du projet Profiterole sur le Gitlab d'Huma-Num[20] dans deux répertoires distincts correspondant aux corpus « Gold » et « Ore », laissant chacun libre de les exploiter avec les outils de son choix. Ces répertoires sont mis à jour au fur et à mesure de la vérification et de la correction des annotations et de l'amélioration de l'analyse syntaxique automatique.

32 Le corpus Gold est déposé sur le site de treebanks Universal Dependencies (UD), dans deux répertoires distincts (Old French et Middle French). Les répertoires seront enrichis tous les 6 mois.

33 La documentation du projet accompagne les ressources diffusées : l'entrepôt Gitlab contient des informations sur le format des annotations et les outils d'analyse automatique impliqués. Le portail BFM-TXM donne accès à un tutoriel d'interrogation du corpus et le site web de la plateforme TXM donne accès au manuel de l'extension « Annotation syntaxique » fournissant des exemples issus du corpus Profiterole mais aussi d'autres corpus intégrés dans UD. Des hyperliens permettant de naviguer entre les différentes ressources et plateformes sont bien entendu fournis.

3.2. Exploitation avec le portail BFM-TXM

34 Le portail de la Base de français médiéval http://txm.bfm-corpus.org est le principal point d'accès à l'interrogation du corpus Profiterole (version 1.0 actuellement). Le corpus est accessible à tous les utilisateurs inscrits à la BFM. Le sous-corpus « gold » permet de limiter les recherches aux annotations vérifiées.

35 Les annotations syntaxiques peuvent être exploitées à l'aide de l'ensemble des propriétés de mots interrogeables à travers des requêtes CQL[21] avec les commandes « Index » et « Concordance ». Pour chaque mot on connaît sa position dans la phrase (ud-id), la position de son gouverneur (ud-head, « 0 » si le mot est la racine de l'arbre de dépendances), son étiquette morphologique et ses traits additionnels (ud-upos, ud-feats), son étiquette syntaxique (ud-deprel), ainsi que les étiquettes de son gouverneur et de ses dépendants directs (ud-head-deprel, ud-dep-deprel). Ainsi, la requête suivante permet d'extraire les pronoms possessifs sujets :

```
[ud-upos="PRON" & ud-feats=".*Poss=Yes.*" & ud-deprel="nsubj"].
```

36 Pour extraire les déterminants possessifs dont le gouverneur est un sujet, il suffit de modifier la requête légèrement :

```
[ud-upos="DET" & ud-feats=".*Poss=Yes.*" & ud-head-deprel="nsubj"].
```

37 La recherche de séquences de mots nécessite des requêtes CQL plus complexes, puisqu'il faut s'assurer que les motifs extraits ne dépassent pas les limites d'une phrase.

En l'absence de cette contrainte, une requête sur un objet suivi d'un sujet peut donner comme résultat un objet situé à la fin d'une phrase suivi d'un sujet situé au début de la suivante. Les phrases complexes ajoutent un niveau de difficulté supplémentaire, car un objet d'une principale peut se trouver devant le sujet d'une subordonnée sans qu'il s'agisse d'une instance d'ordre des mots Objet-Sujet. L'exemple suivant illustre une requête sur l'ordre des mots OVS au sein d'une proposition principale. Les limites de la phrase ne seront pas dépassées grâce à la condition ud-id!="1" ajoutée à chaque mot sauf le premier et les subordonnées sont exclues grâce à la condition ud-head-deprel="root" :

```
[ud-deprel="obj" & ud-head-deprel="root"] [ud-id!="1"]* [ud-
deprel="root" & ud-id!="1"][ud-id!="1"]* [ud-deprel="nsubj" & ud-
head-deprel="root"& ud-id!="1"].
```

38 La liste complète des propriétés de mots disponibles et des exemples de requêtes supplémentaires sont fournis dans le Tutoriel du corpus Profiterole accessible sur le portail BFM-TXM.

39 Les requêtes CQL sont un outil puissant, mais leur complexité devient rapidement gênante en cas de recherche de constructions longues impliquant des dépendances en cascade. Par ailleurs, à ce jour le portail BFM ne permet pas de visualiser des arbres syntaxiques. En revanche, le portail permet de télécharger en un simple clic le corpus Profiterole au format .txm utilisable avec l'application TXM pour poste qui offre de nombreuses fonctionnalités supplémentaires décrites dans la section suivante. Il suffit de sélectionner le corpus par son icône et de cliquer sur le bouton « Télécharger » de la barre d'outils ou bien de faire un clic droit et sélectionner cette commande dans le menu contextuel.

3.3. Exploitation avec TXM pour poste

3.3.1. Installation de l'extension « Annotation syntaxique »

40 TXM pour poste est un logiciel *open source* diffusé gratuitement pour Windows, Mac et Linux dont le développement est coordonné au sein du laboratoire IHRIM à l'ENS de Lyon (Heiden *et al.* 2010). Toutes les informations sur TXM et les méthodes d'analyse qu'il implémente sont réunies sur son site officiel[22]. Afin de profiter de l'ensemble des possibilités d'exploitation d'annotations syntaxiques dans TXM, il est nécessaire de disposer de la version 0.8.3 ou ultérieure du logiciel.

41 Une fois le logiciel TXM installé, il faut ajouter l'extension « Syntactic Annotation » depuis le menu « Fichier > Ajouter une extension ». Cette extension enrichit TXM de plusieurs fonctionnalités pour l'importation et l'exportation d'annotations syntaxiques, pour des requêtes syntaxiques par moteurs de recherche et pour la visualisation d'annotations syntaxiques.

3.3.2. Import et export d'annotations syntaxiques (CoNLL-U et TIGER-XML)

42 Les fonctionnalités d'import et d'export de corpus annotés en syntaxe ou d'annotations aux formats CoNLL-U et TIGER-XML ont été développées dans le cadre du projet Profiterole pour l'analyse et la diffusion de son corpus, mais elles ne sont pas strictement nécessaires pour son exploitation, puisque le corpus diffusé comporte déjà toutes les annotations et peut être chargé dans TXM directement. En revanche, le

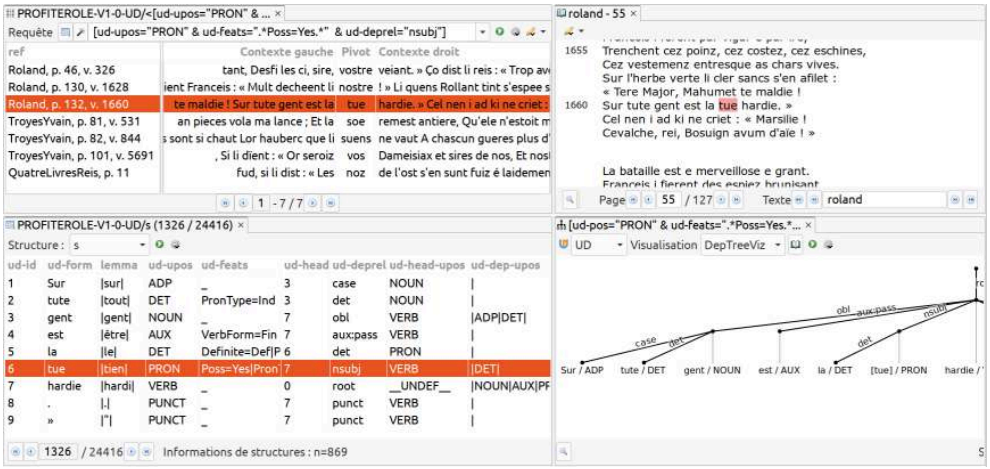
nouveau module d'import CoNLL-U permet d'importer dans TXM n'importe quel corpus du projet Universal Dependencies, ou produit par des parseurs utilisant ce format et le module d'import TIGER-XML permet d'importer n'importe quel corpus dans ce format.

3.3.3. Exploitation des annotations syntaxiques avec le moteur CQP

43 Les requêtes CQL présentées dans la section 3.2 sont bien entendu utilisables dans TXM pour poste. En plus de l'exploitation de ces requêtes sous forme d'index et de concordances, TXM permet de visualiser les annotations syntaxiques sous forme d'arbres avec la nouvelle commande « Arbre Syntaxique ». L'utilisateur peut choisir entre trois modes de visualisation (TIGERSearch ou bien DepTreeViz ou Brat pour Universal Dependencies) et sélectionner les propriétés de mots à afficher, la partie de la phrase correspondant à la requête étant mise en relief par des crochets (comme c'est le cas de « [tue] / PRON » sous l'arbre syntaxique dans la Figure 1).

44 Les différentes vues de résultats produites par les outils de TXM sont reliées par des liens hypertextes permettant d'obtenir des parcours d'analyse intégrés comme l'exemple de parcours sur les « pronoms possessifs sujets » illustré à la Figure 1.

Figure 1. Exemple de parcours d'analyse des « pronoms possessifs sujets » dans TXM avec l'extension « Annotation Syntaxique »



45 Le contexte de la troisième ligne de la concordance (située en haut à gauche) de la requête [ud-upos="PRON" & ud-feats=".*Poss=Yes.*" & ud-deprel="nsubj"] est élargi par le lien hypertexte « Afficher en plein texte » qui ouvre l'édition du texte Roland (située en haut à droite) à la page où se trouve l'occurrence mise en évidence. L'arbre syntaxique de la phrase contenant le mot est visualisé (en bas à droite) par le lien « Afficher l'arbre syntaxique » depuis la page d'édition. Enfin, le navigateur de corpus est ouvert (en bas à gauche) sur la phrase correspondante en affichant les colonnes de propriétés de mots correspondants à toutes les informations CoNLL-U disponibles (dont le nom est préfixé par « ud- » : id, form, upos, feats, head, deprel...).

3.3.4. Exploitation des annotations syntaxiques avec le moteur de recherche TIGERSearch

46 Le moteur de recherche TIGERSearch (König *et al.* 2003) spécialement conçu pour des corpus annotés en syntaxe permet de formuler des requêtes plus sophistiquées sur les

relations syntaxiques et notamment de travailler sur des dépendances profondes (lorsque plusieurs niveaux de nœuds intermédiaires se placent entre un mot et la racine d'une phrase). Les requêtes sont généralement plus longues qu'en CQP, mais l'écriture sur plusieurs lignes et l'utilisation de variables et de labels de dépendances les rendent relativement faciles à écrire et à interpréter. Les deux requêtes suivantes permettent, par exemple, d'extraire des phrases SOV dans des propositions principales et indépendantes d'une part :

```
#pivot:[pos="VERB"]
& #clause:[cat="root" & type="VFin"]
& #clause >L #pivot
& #clause >D #obj:[cat=("obj"|"ccomp"|"obj\:advneg"|"obj\:advmod")]
& #clause >D #suj:[cat=("nsubj"|"csubj")]
& #obj >L #objhead:[pos!=("PRON"|"VERB")]
& #suj >L #sujhead:[]
& #sujhead.* #objhead & #objhead.* #pivot
```

47 et dans des subordonnées d'autre part :

```
#pivot:[pos="VERB"]
& #clause:[cat!="root" & type="VFin"]
& #clause >L #pivot
& #clause >D #obj:[cat=("obj"|"ccomp"|"obj\:advneg"|"obj\:advmod")]
& #clause >D #suj:[cat=("nsubj"|"csubj")]
& #obj >L #objhead:[pos!= ("PRON"|"VERB")]
& #suj >L #sujhead:[]
#sujhead.* #objhead & #objhead.* #pivot
```

48 Elles permettent de quantifier le recul différencié de l'ordre SOV selon le type de proposition : alors que cette combinaison, prédominante en latin classique, se raréfie rapidement en principale/indépendante (13,5 % de l'ensemble des 6 combinaisons possibles aux 9e-12e s., 7 % au 13e s., 2,5 % au 14e s. puis 1,5 % au 15e s.), elle se maintient davantage en subordonnée (24,7 % aux 9-12e s., 15,3 % au 13e s., 10,4 % au 14e s., puis 4,2 % au 15e s.). Elles permettent, en faisant varier la dernière ligne, de dresser un tableau global de l'évolution de l'ordre des constituants majeurs en français médiéval (Prévost à paraître).

49 Le moteur de recherche TIGERSearch est disponible pour les commandes « Syntactic Tree », « Index » et « Concordances ».

3.3.5. Exploitation des annotations syntaxiques par combinaison des moteurs CQP et TIGERSearch

50 On peut combiner des contraintes d'extraction par TIGERSearch et par CQL en appliquant des requêtes TIGERSearch sur des corpus TXM construits par requête CQL, comme dans ceux produits par les outils Sous-Corpus et Partition. Un exemple type est d'abord de créer un sous-corpus de textes spécifiques tout en situant les mots par rapport à certaines structures textuelles (comme tous les débuts ou fins de sections, de paragraphes ou de discours direct). Puis d'appliquer une extraction TIGERSearch qui sera limitée aux mots de ce sous-corpus. Cela permet de croiser de façon souple des contraintes syntaxiques avec des contraintes structurelles.

3.3.6. Utilitaires spécifiques

51 Un certain nombre d'utilitaires visant à faciliter l'analyse syntaxique ont été élaborés dans le cadre du projet Profiterole. Ils sont accessibles par le menu « Utilitaires > tiger > exploit ».

52 « TIGER Summary » dénombre les résultats d'une requête TIGERSearch avec prise en compte ou non des sous-graphes (occurrences multiples au sein d'une phrase). « TIGER Index » produit un index hiérarchique des valeurs de propriétés de nœuds de matchs TIGERSearch. L'utilitaire « TIGER Ratio » calcule le rapport entre le nombre de matchs de deux requêtes TIGERSearch. Enfin, l'utilitaire « TIGER SVO Summary » produit un tableau de fréquences des différents ordres de constituants principaux en fonction de nombreux paramètres définis dans un tableau de paramètres fourni par l'utilisateur.

3.4. Documentation

53 La description complète des fonctionnalités d'exploitation d'annotations syntaxiques est présentée dans la section « Extension Annotation Syntaxique » du *Manuel de TXM* (Heiden 2023).

4. Conclusion et perspectives

54 La constitution et l'annotation syntaxique du corpus Profiterole, ressource inédite pour la période médiévale, et le développement de l'environnement qui permet son exploration, ouvre la voie à un renouvellement, tant quantitatif que qualitatif, des études de la syntaxe du français médiéval. Profiterole constitue en outre un précieux corpus d'apprentissage pour créer de nouveaux parseurs dans la perspective de l'enrichissement syntaxique automatique d'autres textes de cette période.

BIBLIOGRAPHY

Bojanowski P., Grave E., Joulin A. & Mikolov T. (2017). « Enriching Word Vectors with Subword Information », *Transactions of the Association for Computational Linguistics* 5 : 135-146.

Dozat T. & Manning C. (2016). « Deep Biaffine Attention for Neural Dependency Parsing », *in CoRR*, juin 2016. http://arxiv.org/abs/1611.01734.

Grobol L., Prévost S. & Crabbé B. (2021). « Is Old French tougher to parse? », *in Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories* (TLT, SyntaxFest 2021). Sofia, Bulgaria : Association for Computational Linguistics, 27-34.

Grobol L., Regnault M., Ortiz Suarez P., Sagot B., Romary L. & Crabbé B. (2022). « BERTrade : Using Contextual Embeddings to Parse Old French », *in Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France : European Language Resources Association, 1104-1113.

Heiden S., Magué J.-P. & Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *in JADT 2010 : 10th International Conference on the Statistical Analysis of Textual Data*. Rome, Italie. https://shs.hal.science/halshs-00549779.

Heiden S. (2023). *Manuel de TXM*, section « Extension Annotation Syntaxique version 1.0 ». https://pages.textometrie.org/txm-manual (à venir pour décembre 2023).

König E., Lezius W. & Voormann H. (2003). *TIGERSearch User's Manual*. Stuttgart : IMS, University of Stuttgart. https://www.ims.uni-stuttgart.de/documents/ressourcen/werkzeuge/tigersearch/manual.html.

Kroch A. & Santorini B. (éd.) (2021). *Penn-BFM Parsed Corpus of Historical French*, version 1.0. https://github.com/beatrice57/mcvf-plus-ppchf.

Martineau F., Hirschbühler P., Kroch A. & Morin Y.-C. (éd.) (2021). *MCVF Corpus*, parsed, version 2.0. https://github.com/beatrice57/mcvf-plus-ppchf.

McDonald R., Pereira F., Ribarov K. & Hajič J. (2005). « Non-projective dependency parsing using spanning tree algorithms », *in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Vancouver, British Columbia, Canada : Association for Computational Linguistics, 523-530. https://aclanthology.org/H05-1066.

Prévost S. (à paraître). « Grammaticalisation et changements constructionnels à l'épreuve de l'évolution de l'ordre des mots en français ».

Regnault M. (2019). « Adaptation d'une métagrammaire du français contemporain au français médiéval », *in TALN-RECITAL 2019 - 26e édition de la conférence TALN (Traitement Automatique des Langues Naturelles) et 21e édition de la conférence jeunes chercheur·euse·s RECITAL*, juillet 2019, Toulouse, France. https://inria.hal.science/hal-02147686.

Regnault M., Prévost S. & Villemonte de la Clergerie E. (2019). « Challenges of language change and variation: towards an extended treebank of Medieval French », *in* M. Candito, K. Evangn, S. Oepen et D. Seddah (éd.) *Proceedings of the 18th InternationalWorkshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 144-150. https://www.aclweb.org/anthology/events/ws-2019/#w19-78.

Smith J. C. (2002). « Middle French. When ? What ? Why ? », *Language Sciences* 24 : 423-445.

Stein A. & Prévost S. (2013). « Syntactic annotation of medieval texts : the Syntactic Reference Corpus of Medieval French (SRCMF) », *in* P. Bennett, M. Durrell, S. Scheible et R. Whitt (éds.), *New Methods in Historical Corpora Corpus Linguistics and International Perspectives on Language*, CLIP vol. 3, Tübingen : Narr, 275-282.

Villemonte de La Clergerie E. (2005). « From Metagrammars to Factorized TAG/TIG Parsers », *in Proceedings of the Ninth International Workshop on Parsing Technology* (IWPT). Vancouver, British Columbia, Canada : Association for Computational Linguistics, 190-191.

Villemonte de La Clergerie E., Sagot B. & Seddah D. (2017). « The ParisNLP entry at the ConLL UD Shared Task 2017 : A Tale of a #ParsingTragedy », *in Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada : Association for Computational Linguistics, 243-252.

NOTES

1. Le corpus Profiterole a été développé dans le cadre du projet ANR du même nom. Profiterole est l'acronyme de : PRocessing Old French Instrumented TExts for the Representation Of

Language Evolution. Le compte-rendu final du projet est disponible à : https://www.lattice.cnrs.fr/projets/projets-passes/projets-anr/projet-anr-profiterole.

2. Notons que les frontières du moyen français ont toujours été quelque peu mouvantes, la nature des critères retenus étant assez variable (voir Smith 2002). Il a de plus été proposé depuis une vingtaine d'années d'étendre la borne finale jusque 1550, en relation avec l'émergence de la notion de français préclassique.

3. Syntactic Reference Corpus of Medieval French, http://srcmf.org, voir Stein et Prévost (2013).

4. http://srcmf.org/#sec:documentation

5. https://universaldependencies.org

6. https://github.com/beatrice57/mcvf-plus-ppchf, voir Martineau, Hirschbühler, Kroch et Morin (2021) et Kroch et Santorini (2021).

7. La rareté des données avant la fin du 11e siècle nous a conduit à regrouper les textes de cette période avec ceux du 12e siècle.

8. Le classement des textes en domaines et/ou en genres est complexe et a donné lieu à de nombreuses propositions. Nous retenons ici le classement en domaines proposé par l'équipe de la Base de Français Médiéval, et qui fait désormais largement référence pour les textes médiévaux : http://corptef.ens-lyon.fr/spip.php?article62.

9. Le passage du latin au français : constitution et analyse d'un corpus numérique latino-français, http://palafra.org.

10. DEscription et MOdélisation des Chaînes de Référence : outils pour l'Annotation de corpus (en diachronie et en langues comparées) et le Traitement automatique, https://www.lattice.cnrs.fr/democrat.

11. La liste des textes, et leurs métadonnées, est disponible sur https://txm-bfm.huma-num.fr/txm/?command=metadata&path=/PROFITEROLE-V1-0.

12. http://bfm.ens-lyon.fr/spip.php?article176

13. https://universaldependencies.org/u/pos/all.html

14. https://universaldependencies.org/u/overview/syntax.html

15. L'analyseur attribue à chaque paire de mots un score estimant la vraisemblance qu'ils forment une relation de dépendance dans le contexte de la phrase. Un algorithme construit ensuite un arbre en dépendances en sélectionnant un gouverneur pour chaque mot tout en garantissant que les contraintes suivantes sont respectées : l'arbre en dépendances est bien formé (absence de boucle…), tous les mots de la phrase y ont une place et la somme des scores de vraisemblance des paires sélectionnées est supérieure à celle de n'importe quel autre arbre respectant les deux précédentes contraintes.

16. Après avoir calculé une représentation (un vecteur) pour chaque mot dans le contexte de la phrase, le score de vraisemblance de chaque paire est obtenu en multipliant les vecteurs correspondant aux deux mots de la paire par le truchement d'une matrice asymétrique. L'asymétrie de la matrice permet de produire un score différent en fonction duquel des deux mots est gouverneur et duquel est dépendant dans la relation. C'est du truchement par la matrice du produit des deux vecteurs que l'opération tient son nom d'attention bi-affine.

17. Un arbre est projectif si, lorsqu'on représente toutes les dépendances au-dessus de la phrase, il n'y a pas de croisement. C'est le cas quand les relations de dépendances sont imbriquées.

18. Dans FastText, la représentation d'un mot est calculée à partir du mot lui-même ainsi que des n-grammes de caractères qui le composent.

19. Voir le compte-rendu du projet pour la disponibilité des codes et modèles : https://www.lattice.cnrs.fr/projets/projets-passes/projets-anr/projet-anr-profiterole.

20. https://gitlab.huma-num.fr/profiterole/corpus-profiterole

21. Corpus Query Language, https://cwb.sourceforge.io/files/CQP_Manual.

22. https://www.textometrie.org

23. Contrainte qui vise à éliminer les objets pronominaux et propositionnels, dont la position, respectivement préverbale et postberbale, est quasiment fixe depuis les débuts du français.

ABSTRACTS

The Profiterole ANR project aimed at the constitution of new resources for Medieval French (from 9th to 15th century): a morpho-syntactically annotated corpus and lexicons, the creation of syntactic parsers for Medieval French, the development of tools for the dissemination and textometric analysis of syntactic annotation in the context of the TXM platform, and finally the analysis of some syntactic aspects of the evolution of the French language. First we describe the constitution of the Profiterole corpus in terms of texts, genres and dates and annotation schemes. Then we introduce the syntactic parsers that have been developed alongside the constitution of the corpus, and explain how their outputs have been combined in order to assist the manual correction of the annotation of the corpus. Eventually, we discuss the distribution modalities of the data and the parsers that have been produced during the project. Special emphasis is set on the tight integration of the corpus to the TXM software, both in the online version reachable from the website of the Base de Français Médiéval (BFM) and the offline version, with a number of examples of CQP and TIGER queries to help the exploration of the corpus.

Le projet ANR Profiterole avait pour objectifs la constitution de ressources pour le français médiéval (9e-15e s.) : un corpus annoté en (morpho-)syntaxe et des lexiques, la conception d'analyseurs syntaxiques pour le français médiéval, le développement d'outils de diffusion et d'analyse textométrique de l'annotation syntaxique dans le contexte de la plateforme TXM, et, enfin, la modélisation de certains aspects syntaxiques de l'évolution du français. Nous commençons par décrire la constitution du corpus Profiterole en termes de choix de textes, genres et périodes et de types d'annotation. Puis nous présentons les modèles d'analyse syntaxique développés conjointement à la constitution du corpus, ainsi que la manière dont leurs prédictions ont été combinées pour assister la correction manuelle de l'annotation du corpus. Enfin, nous abordons les modalités de diffusion des données et modèles produits dans le cadre du projet avec un accent particulier sur l'intégration du corpus annoté à TXM, tant dans sa version en ligne accessible depuis le portail de la Base de français médiéval (BFM) que dans sa version pour ordinateur personnel, avec des exemples de requêtes CQP et TIGER facilitant l'exploration et l'analyse du corpus.

INDEX

Keywords: Medieval French, Old French, Middle French, corpus, annotation, syntactic parsing, parser, textometry, TXM
Mots-clés: français médiéval, ancien français, moyen français, corpus, annotation, analyse syntaxique, parseur, textométrie, TXM

AUTHORS

SOPHIE PRÉVOST

Lattice (UMR 8094, ENS-PSL, Université Sorbonne Nouvelle)

LOÏC GROBOL

MoDyCo (UMR 7114, Université Paris Nanterre)

MATHIEU DEHOUCK

Lattice (UMR 8094, ENS-PSL, Université Sorbonne Nouvelle)

ALEXEI LAVRENTIEV

IRHIM (UMR 5317, CNRS, ENS de Lyon)

SERGE HEIDEN

IRHIM (UMR 5317, ENS de Lyon, CNRS)