Projet ANR- 16-CE38-0010

PROFITEROLE

Programme DS08072016

A IDENTIFICATION

Acronyme du projet	Profiterole
Titre du projet	PRocessing Old French Instrumented TExts for the Representation Of Language Evolution
Coordinateur du projet (société/organisme)	Sophie Prévost (CNRS)
Période du projet (date de début – date de fin)	01/03/2017-28/02/2022
Site web du projet, le cas échéant	En construction

Rédacteur de ce rapport	
Civilité, prénom, nom	Sophie Prévost
Téléphone	0158076622/0621810150
Adresse électronique	sophie.prevost@ens.psl.eu
Date de rédaction	05/09/2022

Si différent du rédacteur, indiquer un contact pour le projet	
Civilité, prénom, nom	
Téléphone	
Adresse électronique	

Liste des partenaires présents à la fin du projet (société/organisme et responsable scientifique)	-Sophie Prévost (CNRS/ENS-PSL/Sorbonne Nouvelle) ; responsable scientifique. - Mathieu Dehouck(CNRS/ENS-PSL/Sorbonne Nouvelle) ; -Benoît Crabbé (Université Paris Cité) ; responsable scientifique. -Eric Villemonte de la Clergerie (INRIA). -Benoît Sagot (INRIA). -Mathieu Constant (Université de Lorraine). -Serge Heiden (ENS de Lyon) ; responsable scientifique. -Céline Guillot (ENS de Lyon). -Alexei Lavrentiev (CNRS/ENS de Lyon). -Mathieu Decorde (ENS de Lyon). -Nicolas Mazziotta (Université de Liège, Belgique) -Kim Gerdes (Université Paris Saclay). -Achim Stein (Université de Stuttgart, Allemagne) - Tom Rainsford (Université de Stuttgart, Allemagne). -Mathilde Regnault (doctorante, Sorbonne nouvelle/Université de Stuttgart, Allemagne). - Loïc Grobol (post-doctorant ENS-PSL) -Cristina Garcia-Holgado (Université de Strasbourg)

B RESUME CONSOLIDE PUBLIC

B.1 RESUME CONSOLIDE PUBLIC EN FRANÇAIS

PRocessing Old French Instrumented TExts for the Representation Of Language Evolution

Fournir des outils et des ressources pour le français medieval: analyseurs syntaxiques, lexiques, corpus et plateforme pour l'analyse de corpus structures et annotés

Le projet Profiterole avait trois objectifs principaux, étroitement corrélés, relevant des domaines de la linguistique et du Traitement automatique des langues (TAL) : tout d'abord, viser le développement d'une méthodologie d'exploration et d'annotation de données linguistiques hétérogènes tout en fournissant des analyseurs automatiques pour différentes étapes de la langue française ; ensuite, étendre les ressources linguistiques du français, en construisant un gros corpus annoté syntaxiquement (1 million de mots) et des lexiques morphologiques pour le français médiéval (9e-15e siècles) ; enfin, esquisser la modélisation des aspects morphologiques et syntaxiques de l'évolution diachronique du français.

La période médiévale constitue une période décisive pour l'étude de l'évolution du français. C'est en effet au cours de cette période que la plupart des changements morphologiques et syntaxiques fondamentaux ont été initiés et ont commencé à se répandre dans la langue. Se concentrer sur cette période chronologique nous permet donc de mieux comprendre l'évolution du français et de mieux appréhender certains mécanismes de changement qui ont également eu lieu dans d'autres langues. Les contraintes matérielles ont jusqu'à présent limité l'exploration des données ou d'autres analyses approfondies des recueils de textes des états anciens du français, qui nécessitent une exploitation partiellement automatisée des données. Ceci est particulièrement vrai pour la période médiévale. La constitution de grandes ressources, jusqu'à présent inexistantes, ouvre des perspectives nouvelles et prometteuses en ce qui concerne l'amélioration de notre connaissance de l'évolution du français d'un point de vue morphologique et syntaxique.

Une alliance entre Linguistique diachronique, Philologie, Traitement automatique des langues et Informatique

L'alliance de technologies et compétences variées et complémentaires a permis de mener au mieux les objectifs à atteindre :
- Compétences philologiques: elles ont permis de choisir au sein de la Base de Français Médiéval les textes retenus, tous libres de droit, sur des critères croisant caractéristiques linguistiques et métadonnées, afin de constituer un corpus diversifié et équilibré, aussi représentatif que possible de la période médiévale.
- Technologies et compétences relevant du Traitement automatique des langues: elles ont permis la constitution de lexiques (agrégation semi-manuelle des formes et variantes à partir des ressources existantes et calcul de pseudo-synonymes par apprentissage automatique) et développement d'analyseurs syntaxiques (symbolique, neuronaux et hybride).
- Technologies et compétences informatiques: elles ont permis de developer au sein de l'outil TXM (https://txm.gitpages.huma-num.fr/textometrie/) de nouveaux modules de haut niveau dédiés à l'importation d'annotations syntaxiques (en des formats différents) et à l'exploration syntaxique en vue d'une exploitation textométrique optimisée.
- Compétences en morpho-syntaxe et syntaxe historique: elles ont permis l'affinement des règles de l'analyseur syntaxique symbolique, et la fiabilité des corrections des annotations morpho-syntaxiques et syntaxiques.

Résultats majeurs du projets :

Les résultats majeurs du projet consistent en une série de livrables, en accès ouvert selon leurs licences respectives précisées sur les différents sites d'accès aux ressources:

a) Le corpus Profiterole annoté morpho-syntaxiquement et syntaxiquement est diffusé sous plusieurs formes :

- .conll-u : fichiers annotés syntaxiquement en UD au format CoNLL-U : https://gitlab.huma-num.fr/profiterole/corpus-profiterole ;
  A venir : https://universaldependencies.org/
- .txm : fichier corpus binaire à charger dans le logiciel TXM avec des annotations syntaxiques interrogeables par les moteurs CQP et TIGERSearch (portail BFM http://txm-bfm.huma-num.fr) ;
- en ligne : corpus interrogeable directement sur le portail de la Base de français médiéval http://txm-bfm.huma-num.fr (moteur CQP uniquement).

b) les analyseurs syntaxiques

- HOPS : analyseur syntaxique neuronal:
  Code : https://github.com/hopsparser/hopsparser
  Modèles : https://zenodo.org/record/6542539
- MetaMOF: analyseur syntaxique symbolique :
  Métagrammaire : https://gitlab.inria.fr/mgkit/MetaMOF.
  Chaîne de traitement : https://gitlab.inria.fr/almanach/alpi.
- DYALOG-SRNN: analyseur syntaxique neuronal
  Code : https://gitlab.inria.fr/clergeri/dyalog-srnn
  Modèles : https://zenodo.org/record/7298545
- DYALOG-SRNN/MetaMOF: analyseur hybride
  A venir

c) Les lexiques :

- OFrLex:
  https://gitlab.inria.fr/almanach/alexina/ofrlex
- OFrLex/BFMGOLDLEM:
  https://github.com/CristinaGHolgado/oldfrench_lexicon

d) Extension TXM « Syntactic Annotation » sous licence GNU GPL V3 :

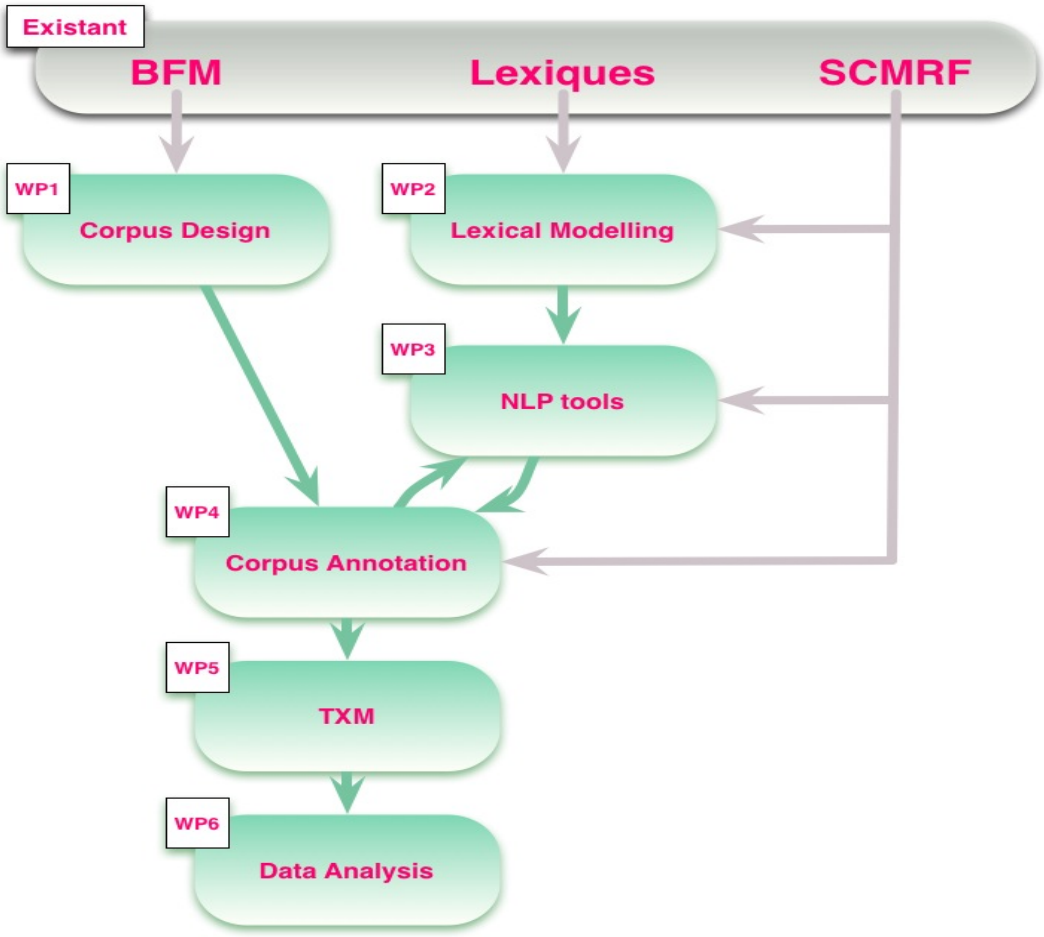
Les sources sont disponibles en ligne (dans six projets de composants) :

- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.rcp

Production scientifique depuis le début du projet

La production scientifique depuis le début du projet a consisté en communications dans des conférences nationales et internationales, qui ont permis de présenter les enjeux et objectifs du projet, ainsi que ses différents volets, en particulier: le développement des différents parsers et la constitution des lexiques, les modalités et procédures d'annotation du corpus et celles de l'exploitation des données annotées.

Illustration :



Informations factuelles :

Le projet Profiterole, qui relève à la fois de la recherche fondamentale et expérimentale, est dirigé par Sophie Prévost (CNRS/ENS/Université Sorbonne Nouvelle). Il associe Mathieu Dehouck

(CNRS/ENS/Université Sorbonne Nouvelle), Benoît Crabbé (Université Paris Cité), Eric Villemonte de la Clergerie (INRIA), Benoît Sagot (INRIA), Mathieu Constant (Université de Lorraine), Serge Heiden (ENS-Lyon), Céline Guillot (ENS-Lyon), Alexei Lavrentiev (CNRS/ENS-Lyon), Mathieu Decorde (ENS-Lyon), Nicolas Mazziotta (Université de Liège, Belgique), Kim Gerdes (Université Paris Saclay), Achim Stein (Université de Stuttgart, Allemagne), Tom Rainsford (Université de Stuttgart), Mathilde Regnault (Université de Stuttgart), Loïc Grobol (Université Paris-Nanterre) et Gaël Guibon (Université de Lorraine), Cristina Garcia-Holgado (Université de Strasbourg) ainsi que les laboratoires suivants : Lattice (CNRS/ENS/Université Sorbonne nouvelle), IRHIM (CNRS/ENS-Lyon), LLF (CNRS/Université Paris Cité) et Almanach (INRIA). Le projet a débuté en mars 2017 et il a duré 60 mois. Il a reçu un soutien financier de 372.000 euros de l'ANR.

B.2 RESUME CONSOLIDE PUBLIC EN ANGLAIS

PRocessing Old French Instrumented TExts for the Representation Of Language Evolution

Providing tools and resources for Medieval French: parsers, lexicons, corpus and platform for the analysis of structured and annotated corpora

The Profiterole project had three main goals, closely correlated, that fall within the fields of linguistics and Natural Language Processing (NLP): First, targeting the development of a methodology to explore and annotate heterogeneous linguistic data while providing automatic analysers for various stages of the French language; second, expanding linguistic resources for French, by building a large syntactically annotated corpus (1 million words) and morphological lexicons for Medieval French (9^{th}-15^{th} centuries); finally, starting modeling morphological and syntactic aspects of the diachronic evolution of French.
The Medieval period constitutes a critical period for the study of the evolution of French. It is indeed during this period that most core morphological and syntactic changes were initiated and began to spread throughout the language. Focusing on this chronological span therefore allows us to achieve a better insight into the evolution of French and a better understanding of certain mechanisms of change that have also taken place in other languages. Material constraints have so far limited data mining or other extensive analyses of French diachronic text collections, which call for a partially automated utilization of the data. This holds true especially for the Medieval period. The building of so-far lacking large resources has opened up new and promising perspectives as regards the improvement of our knowledge of how French has evolved from a morphological and syntactic point of view.

An alliance between diachronic linguistics, philology, Natural language processing (NLP) and computer science

The alliance of various and complementary technologies and skills has enabled us to achieve our objectives:
- Philological skills: they have allowed us to choose the selected texts from the *Base de Français Médiéval* (BFM), all free of copyright, relying on criteria that associate linguistic characteristics and metadata, in order to constitute a diversified and balanced corpus, as representative as possible of the Medieval period.
- Technologies and skills related to NLP: they allowed the constitution of lexicons (semi-manual aggregation of forms and variants from existing resources and calculation of pseudo-synonyms by automatic learning) and the development of syntactic parsers (symbolic, neural and hybrid).
- Technologies and computer skills: they allowed the development of new high-level modules within the TXM software (https://txm.gitpages.huma-num.fr/textometrie/) dedicated to the import of syntactic annotations (in different formats) and to syntactic exploration for an optimized textometric exploitation.
- Skills in historical morpho-syntax and syntax: they allowed the refinement of the rules of the symbolic syntactic parser, and the reliability of the corrections of morpho-syntactic and syntactic annotations.

Main results of the project:

The major results of the project consist in a series of deliverables, available in open access according to licences which are specified on the different websites:

a) The morpho-syntactically and syntactically annotated **Profiterole corpus**, is released in different formats:

- .conll-u : files annotated according to universal dependencies guidelines in CoNLL-U format: https://gitlab.huma-num.fr/profiterole/corpus-profiterole ;
  To appear : https://universaldependencies.org/
- .txm : binary corpus to be uploaded in the TXM desktop software, with syntactic annotations searchable with both CQP et TIGERSearch search engines (http://txm-bfm.huma-num.fr) ;
- online : the corpus can be directly searched in the TXM portal software of the *Base de français médiéval* http://txm-bfm.huma-num.fr (only CQP search engine).

b) Syntactic parsers:

- HOPS : neural syntactic parser:
  Code : https://github.com/hopsparser/hopsparser
  Models : https://zenodo.org/record/6542539
- MetaMOF: symbolic syntactic parser:
  Metagrammar : https://gitlab.inria.fr/mgkit/MetaMOF.
  Processing chain: https://gitlab.inria.fr/almanach/alpi.
- DYALOG-SRNN: neural syntactic parser
  Code : https://gitlab.inria.fr/clergeri/dyalog-srnn
  Models: https://zenodo.org/record/7298545
- DYALOG-SRNN/MetaMOF: hybrid syntactic parser
  To be released

c) Lexicons:

- OfrLex
  https://gitlab.inria.fr/almanach/alexina/ofrlex
- OFrLex/BFMGOLDLEM:
  https://github.com/CristinaGHolgado/oldfrench_lexicon

d) TXM extension « Syntactic Annotation » :

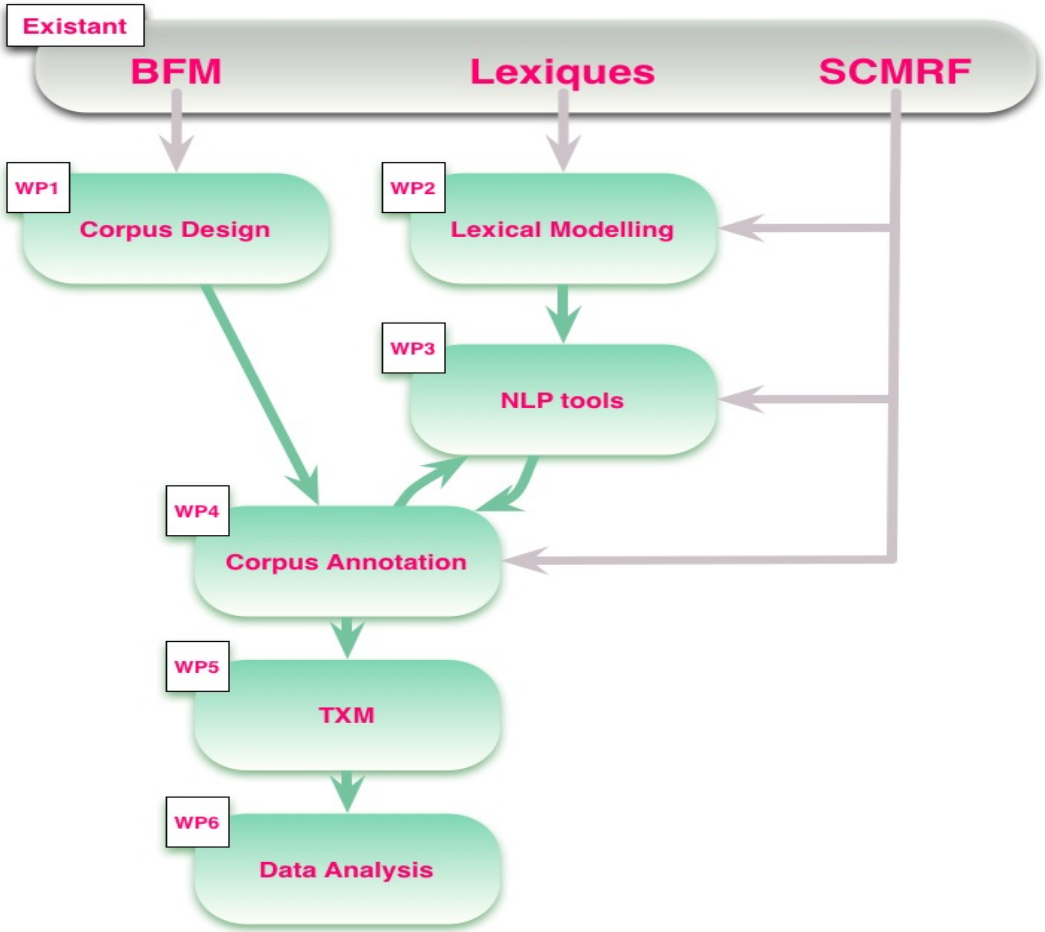
available under GNU GPL V3 licence, sources available online in six different plugin projects:

- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.rcp

Scientific production since the beginning of the project :

The scientific production since the beginning of the project has consisted of communications in national and international conferences, which have made it possible to present the stakes and objectives of the project, as well as its various aspects, in particular: the development of the different parsers and the constitution of the lexicons, the modalities and procedures of annotation of the corpus and those of the exploitation of the annotated data.

Illustration :



Factual Informations :

The Profiterole project, which pertains both to fundamental and experimental research, is led by Sophie Prévost (CNRS/ENS/University Sorbonne Nouvelle). It associates Mathieu Dehouck (CNRS/ENS/Université Sorbonne Nouvelle), Benoît Crabbé (University Paris Cité), Eric Villemonte de la Clergerie (INRIA), Benoît Sagot (INRIA), Mathieu Constant (University of Lorraine), Serge Heiden (ENS-Lyon), Céline Guillot (ENS-Lyon), Alexei Lavrentiev (CNRS/ENS-Lyon), Mathieu Decorde (ENS-Lyon), Nicolas Mazziotta (University of Liege, Belgium), Kim Gerdes (University Paris Saclay), Achim Stein (University of Stuttgart, Germany), Tom Rainsford (University of Stuttgart), Mathilde Regnault (University of Stuttgart), Loïc Grobol (University Paris-Nanterre) and Gaël Guibon (Université of Lorraine), Cristina Garcia-Holgado (University of Strasbourg) as well as the following labs: Lattice (CNRS/ENS/ University Sorbonne nouvelle), IRHIM (CNRS/ENS-Lyon), LLF (CNRS/University Paris Cité) and Almanach (INRIA). The project started in march 2017 and it lasted 60 months. It received a financial support of 372.000 euros from the ANR.The project started in march 2017 and it lasted 60 months. It received a financial support of 372.000 euros from the ANR.

C MEMOIRE SCIENTIFIQUE

Mémoire scientifique confidentiel : non

C.1 RESUME DU MEMOIRE

PRocessing Old French Instrumented TExts for the Representation Of Language Evolution

Fournir des outils et des ressources pour le français medieval: analyseurs syntaxiques, lexiques, corpus et plateforme pour l'analyse de corpus structures et annotés

Le projet Profiterole avait trois objectifs principaux, étroitement corrélés, relevant des domaines de la linguistique et du Traitement automatique des langues (TAL) : tout d'abord, viser le développement d'une méthodologie d'exploration et d'annotation de données linguistiques hétérogènes tout en fournissant des analyseurs automatiques pour différentes étapes de la langue française ; ensuite, étendre les ressources linguistiques du français, en construisant un gros corpus annoté syntaxiquement (1 million de mots) et des lexiques morphologiques du français médiéval (9e-15e siècles) ; enfin, esquisser la modélisation des aspects morphologiques et syntaxiques de l'évolution diachronique du français.
La période médiévale constitue une période décisive pour l'étude de l'évolution du français. C'est en effet au cours de cette période que la plupart des changements morphologiques et syntaxiques fondamentaux ont été initiés et ont commencé à se répandre dans la langue. Se concentrer sur cette période chronologique nous permet donc de mieux comprendre l'évolution du français et de mieux appréhender certains mécanismes de changement qui ont également eu lieu dans d'autres langues. Les contraintes matérielles ont jusqu'à présent limité l'exploration des données ou d'autres analyses approfondies des recueils de textes des états anciens du français, qui nécessitent une exploitation partiellement automatisée des données. Ceci est particulièrement vrai pour la période médiévale. La constitution de grandes ressources, jusqu'à présent inexistantes, a ouvert des perspectives nouvelles et prometteuses en ce qui concerne l'amélioration de notre connaissance de l'évolution du français d'un point de vue morphologique et syntaxique.

Une alliance entre Linguistique diachronique, Philologie, Traitement automatique des langues et Informatique

L'alliance de technologies et compétences variées et complémentaires a permis de mener au mieux les objectifs à atteindre :
- Compétences philologiques: elles ont permis de choisir au sein de la Base de Français Médiéval les textes retenus, tous libres de droit, sur des critères croisant caractéristiques linguistiques et métadonnées, afin de constituer un corpus diversifié et équilibré, aussi représentatif que possible de la période considérée.
- Technologies et compétences relevant du Traitement automatique des langues: elles ont permis la constitution de lexiques (agrégation semi-manuelle des formes et variantes à partir des ressources existantes et calcul de pseudo-synonymes par apprentissage automatique) et développement d'analyseurs syntaxiques (symbolique, neuronaux et hybride).
- Technologies et compétences informatiques: elles ont permis de developer au sein de l'outil TXM (https://txm.gitpages.huma-num.fr/textometrie/) de nouveaux modules de haut niveau dédiés à l'importation d'annotations syntaxiques (en des formats différents) et à l'exploration syntaxique en vue d'une exploitation textométrique optimisée.
- Compétences en morpho-syntaxe et syntaxe historique: elles ont permis l'affinement des règles de l'analyseur syntaxique symbolique, et la fiabilité des corrections des annotations morpho-syntaxiques et syntaxiques.

C.2 ENJEUX ET PROBLEMATIQUE, ETAT DE L'ART

La période médiévale (9^{e}-15^{e} s.) constitue une période décisive pour l'étude de l'évolution du français : c'est au cours de ces siècles que la plupart des changements morphologiques et syntaxiques fondamentaux se sont produits. L'étude de cette période nous permet donc de mieux comprendre l'évolution du français et de mieux appréhender certains mécanismes de changement qui ont également eu lieu dans d'autres langues. Pour cela il est nécessaire de pouvoir explorer une grande quantité de données. Or la communauté manquait jusqu'ici de vastes corpus enrichis morpho-syntaxiquement ET syntaxiquement. En effet, la *Base de Français Médiéval* (9-15^{e} s) contient près de 7 millions de mots, avec un enrichissement morpho-syntaxique partiellement vérifié, mais sans annotation syntaxique. Le corpus *Syntactic Reference Corpus of Medieval French* (9^{e}-13^{e}) <srcmf.org> contient 250 000 mots, enrichis en morpho-syntaxe et syntaxe et 170 000 de ce corpus ont été mots convertis au format UD: https://universaldependencies.org/. Ce corpus reste cependant

de taille limité et ne concerne que l'ancien français. Enfin le Corpus MCVF *Modéliser le changement: les Voies du Français* (http://www.voies.uottawa.ca/index.html) contient près de 1 200 000 mots, annotés selon modèle syntaxique en constituants, avec un parti-pris théorique (générativiste) très fort et ne contient que 8 textes pour le moyen français (760 000 mots), le choix ayant été fait de ne pas échantillonner les textes, ce qui a *de facto* limité le nombre de textes et par conséquent la diversité du corpus.
Il était donc nécessaire de constituer un vaste corpus annoté, représentatif de l'ensemble de la période médiévale, et annoté selon un modèle standard (Universal Dependencies), permettant une utilisation par un grand nombre de linguistes (et en outre une comparaison avec d'autres corpus annotés selon le même modèle).
Les contraintes matérielles ont jusqu'à présent limité l'exploration des données ou d'autres analyses approfondies des recueils de textes des états anciens du français, qui nécessitent une exploitation partiellement automatisée des données. La constitution d'un vaste corpus ouvre des perspectives nouvelles et prometteuses en ce qui concerne l'amélioration de notre connaissance de l'évolution du français d'un point de vue morphologique et syntaxique.

C.3 APPROCHE SCIENTIFIQUE ET TECHNIQUE

WP1 : CONSTITUTION DU CORPUS A ANNOTER

Le WP 1 « Corpus » a été chargé de sélectionner les textes du corpus à partir de la Base de français médiéval, de procéder à leur échantillonnage interne éventuel, d'effectuer les conversions entre les différentes représentations et formats numériques (XML TEI, XML TEI TXM, CoNLL-U, XML TIGER), d'assurer le suivi des versions et la synchronisation des textes et des annotations et enfin de publier le corpus pour son exploitation dans le cadre du projet et au-delà.
Le corpus PROFITEROLE est composé de 63 textes intégraux ou échantillonnés, soit 992 117 mots au total. Il est équilibré sur le plan diachronique du 12^{e} au 15^{e} siècle (près de 250 000 mots par siècle). Les textes en prose constituent 53 % du volume du corpus (40 % est en vers et 7 % est un mélange de prose et de vers). La représentation des dialectes est beaucoup plus inégale, ce qui reflète la situation en littérature française médiévale. Pour près d'un tiers du corpus, l'appartenance dialectale ne peut être établie. Dans le reste du corpus les dialectes les mieux représentés sont l'anglo-normand, le normand, le champenois et le picard. Le domaine littéraire est le mieux représenté (43 % des mots du corpus), les domaines didactique, historique et religieux couvrent chacun près de 18 % du corpus. Les métadonnées complètes des textes du corpus sont présentées dans le tableau en annexe.
Le sous-corpus « gold » est composé des textes dont l'annotation syntaxique a été réalisée ou vérifiée manuellement. Dans la version 1.0 du corpus, cette partie représente 183 821 mots répartis dans 14 textes. Le noyau du corpus vérifié provient du corpus SRCMF (http://srcmf.org). Dans le cadre de PROFITEROLE ses annotations ont été converties au format Universal Dependencies, les tokens de ponctuations ont été ré-intégrés et des échantillons de textes de moyen français ont été ajoutés. La vérification des annotations du corpus PROFITEROLE continuera au-delà de la fin du projet ANR et le sous-corpus « gold » augmentera en accord au fur et à mesure des mises à jour.
L'annotation automatique a été réalisée dans le cadre des WP3 et WP 4 et intégrée dans la version TXM du corpus dans le cadre du WP 5.
Le corpus PROFITEROLE (incluant le sous-corpus « gold ») est diffusé sous plusieurs formes :

- .conll-u : fichiers annotés syntaxiquement en UD au format CoNLL-U (https://gitlab.huma-num.fr/profiterole/corpus-profiterole) ;
  A venir : : https://universaldependencies.org/
- .txm : fichier corpus binaire à charger dans le logiciel TXM avec des annotations syntaxiques interrogeables par les moteurs CQP et TIGERSearch (portail BFM http://txm-bfm.huma-num.fr) ;
- en ligne : corpus interrogeable directement sur le portail de la Base de français médiéval http://txm-bfm.huma-num.fr (moteur CQP uniquement).

Le corpus interrogeable en ligne est accessible à tous les utilisateurs de la Base de français médiéval (sur simple inscription gratuite). Pour interroger les annotations syntaxiques en langage CQL, on dispose de la propriété de mot « ud-deprel » (étiquette syntaxique UD), en combinaison

éventuellement avec les propriétés « ud-id » (identifiant du mot dans la phrase), « ud-head » (identifiant du mot régissant) et « ud-head-deprel » (étiquette syntaxique du mot régissant). La documentation complète des étiquettes est fournie par le projet Universal Dependencies (https://universaldependencies.org), des exemples de mises en application et de requêtes sont fournis dans le diaporama et dans l'exemplier de l'atelier « Outils pour l'exploitation lexicale, morphosyntaxique et syntaxique de la Base de français médiéval » (https://halshs.archives-ouvertes.fr/halshs-03788509) du colloque Diachro X, Paris, 2022.

WP2 : CONSTRUCTION DES LEXIQUES

Deux démarches parallèles ont été menées.

WP2.1 OFrLex

Mise au format unifié du lexique et de sources

Le lexique OFrLex a été créé en amont du post-doc par l'agrégation semi-manuelle de différentes sources : le FROLEX, le Altfranzösisches Wörterbuch, le wiktionary, le lexique de l'ancien français et le dictionnaire électronique de Chrétien de Troyes.
LIVRABLE : lexique au format .lex

Enrichissement lexical automatique

Une fois le lexique constitué il a été convenu de procéder à sa complétion par deux approches principales : une seconde phase d'agrégation automatique d'informations et la détection de variantes possibles n'étant répertoriées dans aucune des sources disponibles.
(1) Obtention de variantes supplémentaires par agrégation. Des variantes supplémentaires ont été obtenues automatiquement par proposition d'agrégations. Cette étape est rendue possible grâce à la mise en place d'une représentation du lexique et de toutes les sources dans une base de données relationnelle commune.
(2) Calcul de pseudo synonymes. Nous avons obtenu de potentielles variantes ou termes connexes que nous regroupons sous la désignation "pseudo synonymes". Ils sont obtenus automatiquement par le biais d'apprentissage automatique et de de mise en relation des lexèmes dans un espace vectoriel commun.
Les enrichissements ont été intégrés en conservant une traçabilité. Les variantes fiables trouvées ont été intégrées tandis que les variantes proposées et les pseudo-synonymes ont été mis sous une couche supplémentaire.
LIVRABLES : (1) lexique et sources mises en commun en une base de données relationnelle ; (2) modèle de langue (ELMO) pré-entraîné sur le corpus PROFITEROLE ; (3) liste des pseudo synonymes et de leur entrée lexicale cible.

Interface de visualisation / mutualisation des modifications

Une interface de visualisation et de modification collaborative a été mise en place spécialement pour le lexique. Elle permet une identification rapide des différentes sources des variantes trouvées et des pseudo synonymes proposés. Elle permet également une modification manuelle conservant pour chaque modification effectué l'historique des différentes opérations. La validation des variantes ou pseudo synonymes fait l'objet d'une campagne de validation ciblée et personnalisable. L'exportation du lexique se fait également via l'interface.
LIVRABLE : application web collaborative https://gguibon.fr:8888/#/

Exploitation du lexique pour l'étude de l'évolution de la langue

Des travaux de recherche ont été effectués sur l'évolution lexicale et plus précisément sur la vérification de plusieurs hypothèses :
1) Est-il possible de capturer des informations ciblées dans des représentations vectorielles des entrées lexicales ? Cette hypothèse a été confirmée en exploitant le modèle de langue appris sur le corpus PROFITEROLE pour la création de méta plongements lexicaux diachroniques dédiés aux lexèmes. La rétention d'informations correctes dans ces vecteurs (prédiction du siècle et de l'UPOS) a été vérifiée empiriquement.
2) Peut-on prédire ces méta plongements lexicaux en diachronie ? L'hypothèse a été confirmée en prédisant la représentation vectorielle d'un lexème du siècle n à n+2. Ce faisant, une complexité croissante a été montrée selon la distance temporelle.

3) Peut-on prédire diachroniquement la graphie d'un lexème ? Cette hypothèse a été partiellement confirmée en réussissant à prédire les graphies courtes avec une performance correcte, mais totalement erronée pour des graphies plus longues.
LIVRABLES : (1) méta plongements lexicaux diachroniques ; (2) modèles d'apprentissage profond pour la prédiction diachronique de lexèmes (vecteurs d'informations et graphie)

Limites et perspectives

Ce WP est l'objet de plusieurs limitations et d'améliorations nécessaires. Tout d'abord, l'usage de l'application pour une campagne de validation nécessite une coordination à plus long terme ainsi que l'implication ponctuelle, mais régulière, de spécialistes de la langue cible. En l'état, la campagne n'a donné qu'à quelques validations de pseudo synonymes en phase de test. Enfin, le lexique est donc utilisable mais perfectible et pas encore complètement fiable sans ces phases de mise en commun de la vérification de chaque entrée lexicale.

WP2.2 Enrichissement et normalisation du lexique : OFrLex et BFMGOLDLEM

L'objectif était d'améliorer et d'enrichir un lexique dans le but de rendre cette ressource plus exploitable pour un parseur syntaxique, ainsi que d'augmenter le nombre de ressources TALN pour le corpus Profiterole.
Ces tâches ont compris, premièrement, la comparaison des entrées du corpus BFMGOLDLEM avec le lexique OFrLex, qui a permis de connaître l'état de ce dernier (entrées partagées). Deuxièmement, d'augmenter la couverture du lexique du parseur (OFrLex) à partir différentes stratégies, qui incluent tant l'utilisation de ressources lexicales externes, que l'utilisation de méthodes d'apprentissage automatique. Cela a eu pour but d'identifier les formes inconnues et de fournir leurs informations morphologiques respectives avec fiabilité. Cette fiabilité est basée sur l'accord entre les propositions issues des différentes ressources (approche non contextualisée), ainsi que par une lemmatisation contextuelle des textes du corpus Profiterole. Différentes opérations ont été réalisées dans ce contexte afin de renseigner les nouvelles formes et de pouvoir les rattacher à une entrée existante du lexique ou d'en ajouter des nouvelles. Ces ressources sont le corpus BFMGOLDLEM, le lexique FROLEX et l'outil LGeRM (http://stella.atilf.fr/LGeRM). Les opérations ont consisté principalement en: alignement d'entrées lexicales après une normalisation de ces ressources, mise en œuvre de règles de tri, génération de variants, requêtes automatiques sur un dictionnaire en ligne (https://anglo-norman.net/). Cependant, cela a soulevé un nouveau problème, à savoir la présence de variants graphiques provenant de l'utilisation de différents référentiels pour les lemmes et qui apportent du bruit dans le lexique. Des tâches secondaires ont été réalisées dans cette ligne afin d'estimer le nombre de variantes graphiques par catégorie et de déterminer des méthodes possibles pour traiter ce problème (e.g.: sélectionner le lemme le plus fréquent dans les doublons). Enfin, ce travail a donné lieu à un modèle amélioré (composé par des nouveaux textes) pour la lemmatisation du corpus Profiterole avec l'outil Pie (https://pypi.org/project/nlp-pie), utilisé précédemment (Holgado, Lavrentev, et Constant 2021).
Une description détaillée est disponible dans le document suivant:
https://docs.google.com/document/d/1wg8wy6GxxQc0TzKdzfELMev7QlvwN5Ak9bfZVZTliuk/edit?usp=sharing

WP3- WP4 DEVELOPPEMENTS DES ANALYSEURS, ANNOTATIONS ET CORRECTIONS

Deux analyseurs syntaxiques neuronaux, un analyseur syntaxique symbolique et un analyseur hybride ont été développés dans la cadre du projet

- **MetaMOF**

MetaMOF désigne tout à la fois (1) une description grammaticale de haut niveau sous forme de méta-grammaire, (2) une grammaire d'arbres adjoints (TAG) à large couverte engendrée par cette méta-grammaire et (3) un analyseur syntaxique résultant de la compilation de cette grammaire.
En tant que méta-grammaire, *MetaMOF* est une adaptation à l'ancien français de *FRMG*, une méta-grammaire à large couverture du français contemporain. *MetaMOF* hérite, sur l'ensemble de ses déclinaisons, de l'éco-système développé autour de *FRMG*.
En tant qu'analyseur, *MetaMOF* prend en entrée une phrase et calcule l'ensemble des analyses syntaxiques possibles pour cette phrase accessible sous forme d'une forêt partagée de dérivations.

Dans les cas où aucune analyse complète de la phrase n'est possible, un ensemble d'analyses partielles est retourné, couvrant au mieux l'ensemble de la phrase.
En pratique, un algorithme de désambiguïsation est utilisé pour sélectionner la meilleure analyse dans l'ensemble des analyses retournées. Cet algorithme s'appuie sur un ensemble de règles heuristiques pondérées portant essentiellement sur une ou plusieurs dépendances syntaxiques. Une telle règle va par exemple favoriser les dépendances d'un verbe avec son sujet. Règles et poids sont manuellement définis dans un premier temps, largement hérités de ceux de *FRMG* avec néanmoins des adaptations spécifiques. Il est ensuite possible d'utiliser un corpus arboré (en l'occurrence le treebank *SRCMF-UD*) pour apprendre une meilleure pondération des règles associées à des configurations plus complexes de traits linguistiques.
En bout de processus, après désambiguïsation, la meilleure analyse peut être convertie dans divers schémas et formats d'annotations syntaxiques, et en particulier UD (*Universal Dependencies*) dans le cadre du projet *PROFITEROLE*.
La nature de l'analyseur et de la méta-grammaire sous-jacente permet cependant d'explorer l'ensemble des analyses possibles pour une phrase, de se rendre compte que certaines phrases ne bénéficient pas d'analyse complète, et de quantifier l'importance de tel ou tel phénomène syntaxique (tel le passif ou les comparatives). Ceci ouvre la voie à des explorations linguistiques fines, faites à partir de requêtes des arbres générés par la grammaire ou d'éléments de la description syntaxique dans la méta-grammaire. On a ainsi accès à des niveaux de profondeur d'analyse qui sont absents du corpus arboré.

- **HOPS**

HOPS est un analyseur syntaxique neuronal, prévu comme suffisamment général pour être utilisé avec des langues variées, issues des données Universal Dependencies, et en particulier sur le cas du français médiéval. Il s'appuie sur les modèles de langues neuronaux état de l'art de la famille BERT.
HOPS est un analyseur à base de graphes, dont le cœur est un système de score de relations de dépendances syntaxiques et un algorithme de recherche d'arbre couvrant maximal. Il repose en particulier sur le modèle d'attention bi-affine sur lequel s'appuient la plupart des analyseurs à l'état de l'art. Il complémente ce cadre générique par une prédiction autonome d'étiquettes de parties du discours (*Part of Speech*) comme objectif auxiliaire pour un apprentissage multi-tâches.
HOPS permet par ailleurs de tenir compte d'informations obtenues de façon non-supervisées sur des corpus non-annotés, sous formes de représentations lexicales vectorielles contextuelles (de type BERT ou de type FastText. Ces ressources permettent d'améliorer significativement les performances de HOPS, comparé à ce qu'il est possible d'obtenir en exploitant uniquement des corpus arborées — de tailles nécessairement plus modestes. En particulier l'adaptation au français médiéval de modèles de type BERT pour français contemporain a permis à HOPS de dépasser nettement l'état de l'art précédent pour l'analyse syntaxique de l'ancien français.

- **DYALOG-SRNN**

DYALOG-SRNN est un analyseur syntaxique neuronal qui étend *DYALOG-SR* un analyseur statistique, et ce dans le cadre de l'environnement en programmation logique *DYALOG*.
DYALOG-SR est un analyseur par transition, de base de type *shift-reduce*, étendu pour couvrir des cas de dépendances croisées (dépendances non-projectives). *DYALOG-SRNN* ajoute une couche neuronale utilisant de l'attention bi-affine pour prédire un gouverneur et une relation syntaxique pour chaque mot, ces prédictions étant rendu plus cohérentes via un algorithme de type MST (*Maximum Spanning Tree*). Ces prédictions fournissent déjà une analyse de qualité mais elles sont en fait ensuite utilisées comme indices par *DYALOG-SR* pour fournir sa propre analyse.
DYALOG-SRNN autorise l'utilisation de plongements lexicaux (*embeddings*) non contextuels, en l'occurrence de type *FastText* obtenu sur le corpus *BFM* dans le cadre de Profiterole. Une évolution bienvenue serait l'utilisation de plongements contextuels fournis par des modèles de langues comme *BERT*, par exemple en poursuivant le pré-entraînement du modèle de langue du français contemporain *CAMEMBERT* sur le corpus *BFM*.
DYALOG-SRNN est entraîné sur le corpus *SRCMF-UD*. Le modèle résultant de cet apprentissage fournit de meilleures analyses que *MetaMOF* sur ce corpus mais comme la plupart des analyseurs statistiques et neuronaux, ses performances vont avoir tendance à se dégrader en dehors de son domaine d'entraînement.

- **DYALOG-SRNN/MetaMOF**

DYALOG-SRNN/MetaMOF utilise les sorties de *MetaMOF* comme indices permettant de guider les décisions d'un modèle appris avec *DYALOG-SRNN* toujours sur le treebank *SRCMF*. Ce couplage avec *MetaMOF* fournit plus de capacité de généralisation au modèle *DYALOG-SRNN* ainsi obtenu, le rendant théoriquement plus robuste sur l'ensemble du corpus Profiterole, comme montré en 2014 avec le couplage *DYALOG-SR/FRMG* sur le treebank hétérogène *SEQUOIA*.

Du vote entre parseurs

Pour accélérer les corrections des dépendances syntaxiques, on peut fournir à l'annotateur une proposition d'arbre générée par un modèle automatique : lorsque l'on a accès à plusieurs de ces modèles, et surtout lorsque ceux-ci sont basés sur des architectures différentes ou sont entraînés de manières différentes, au lieu de présenter les différentes propositions à l'annotateur, on pourra préférer les agréger pour ne proposer qu'un seul arbre.
Une manière simple pour y parvenir est de faire voter les modèles des 4 parseurs (voir supra) pour les relations de dépendances.
Étant donné un token dans la phrase, chaque modèle vote indépendamment pour le gouverneur de ce token (une voix pour le gouverneur prédit par son arbre) et pour le label de la relation (une voix pour la relation prédite dans son arbre). Une fois les votes établis, on utilise le l'algorithme de Chu-Liu-Edmonds pour trouver un arbre maximisant la somme des votes des gouverneurs.
Pour les labels de relations, on choisit le label majoritaire, en cas d'égalité on donne un biais à Hops car il a globalement un meilleur score que les autres modèles.
Dans un second temps, quand un certain nombre de phrases en moyen français ont été corrigées, nous les avons utilisées pour deux choses : s'assurer que les votes étaient meilleurs que les prédictions de chacun des modèles pris indépendamment et pour raffiner la méthode de vote.
De fait, les votes apportent un gain non négligeable, et ce d'autant plus que les textes traités sont en moyen français alors que les modèles ont été entraînés pour l'ancien français.
Pour améliorer la méthode de vote, nous nous sommes intéressés aux coalitions de parseurs.
En effet, les différents modèles ont tous des résultats décents mais ne s'accordent pas à 100%, ce qui implique que dans l'écrasante majorité des cas, au moins un parseur a prédit le bon gouverneur et/ou la bonne relation. Mais comme nous avons quatre modèles, il arrive qu'il n'y ait pas de majorité. Nous avons donc décidé d'associer pour chaque partie-du-discours et pour chaque type de coalition (SRNN et MetaMof peuvent s'accorder sur une solution et Hops et SRNN hybride en avoir choisi une différente chacun) le groupe de modèles qui a raison le plus souvent. Ainsi chaque modèle peut avoir plus de poids que les autres sur les phénomènes pour lesquels il est meilleur.

A l'issue de ces votes, a lieu la correction manuelle d'une partie des données : celles-ci sont ensuite versées dans le sous-corpus Gold, ainsi progressivement étoffé, et servent de données d'apprentissage pour ré-entraîner les parsers et annoter les données non encore vérifiées manuellement. L'hypothèse – avérée- est que les performances des parsers s'améliorent du fait de l'accroissement des données d'apprentissage vérifiées, et donc que les corrections sont moindres. S'ensuit alors le processus de vote, et la procédure est répétée jusqu'à ce que l'ensemble des données ait été vérifié manuellement.

WORKPACKAGE 5 : INTEGRATION ET EXPLOITATION DU CORPUS PROFITEROLE AVEC TXM

La plateforme TXM implémente les outils d'analyse de corpus textuels de la textométrie. L'objectif des travaux du WP5 étaient d'intégrer dans la plateforme tous les moyens d'exploiter les informations syntaxiques d'un corpus pour pouvoir exploiter textométriquement le corpus Profiterole. En plus de l'implication des partenaires du projet, ce WP a bénéficié du travail à plein temps de développement informatique de Matthieu Decorde, en tuilage avec les développements réalisés pour le projet ANR ANTRACT ANR-17-CE38-0010.
À la fin du projet on peut dire que tous les objectifs du WP ont été atteints.

WP5.1. IMPORT ET PARSAGE DU CORPUS

Tout d'abord la possibilité d'importer les informations syntaxiques d'un corpus a été implémentée sous deux formes :

- d'une part l'**import d'annotations syntaxiques** dans un corpus TXM existant depuis un autre corpus. Cette fonctionnalité a d'abord été développée pour aider la logistique du corpus Profiterole mais s'avère intéressante au final pour la logistique de corpus de nombreux autres projets (eg projet « 13 Novembre » ANR-16-EQPX-0003) car cela permet d'intégrer des annotations syntaxiques dans n'importe quel corpus TXM richement structuré et annoté.
- d'autre part l'**import de corpus annotés en syntaxe** pour construire un nouveaux corpus TXM.

Ces deux possibilités ont été développées pour deux représentations syntaxiques différentes :

- d'une part dans le format arborescent **XML-TIGER** propre au moteur de recherche TIGERSearch intégré à TXM ;
- d'autre part dans le format tabulaire **CoNLL-U** qui est exploitable par le moteur de recherche CQP utilisé par TXM.

La prise en compte du format CoNLL-U n'était pas prévue au début du projet mais CoNLL-U s'est avéré être le format de référence du projet. Donc nous avons dû l'adopter pour manipuler le corpus Profiterole.

Il en résulte que les outils développés dans le cadre du projet sont plus généraux que prévu car le format CoNLL-U est également utilisé par de nombreux outils de TAL et de nombreux corpus en langues contemporaines. De fait, les outils développés pour le corpus Profiterole sont utilisables également pour des corpus en langues contemporaines, notamment grâce au formalisme pivot en morpho-syntaxe et en syntaxe UD.

Nous avons réalisé des prototypes d'appel de parseurs syntaxiques (UDPipe et Stanford NLP notamment) - pour l'objectif "parsing integration" du WP, par contre nous n'avons pas intégré d'appel aux parseurs réalisés par les partenaires du projet car ils n'étaient pas disponibles sous cette forme.

WP5.2. INTEGRATION DE MOTEURS SYNTAXIQUES ET APPARIEMENT DE MOTEURS D'EXTRACTION

La stratégie de double requêtage - plein texte par CQP & syntaxique par TIGERSearch ou CQP - a été entièrement implémentée et validée dans les outils d'exploitation. L'utilisateur peut donc désormais combiner n'importe quelle configuration de corpus (sous-corpus ou partition) à des extractions par annotations syntaxiques par TIGERSearch ou CQP. La relation entre les configurations et les extractions est réalisée au niveau des mots/tokens : les moteurs partagent tous la même séquence de mots/tokens d'un corpus TXM donné.

WP5.3. INTEGRATION D'ANNOTATIONS SYNTAXIQUES DANS LES OUTILS D'UTILISATEURS FINAUX

Ce développement a permis d'augmenter les commandes intégrées de TXM Index et Concordances qui exploitent désormais directement les annotations syntaxiques sous deux formes (TIGER ou CoNLL-U/CQP) mais également de développer quelques utilitaires expérimentaux basés sur des extractions par le moteur TIGERSearch pour les besoins d'exploitation du projet Profiterole : outils « TIGER Summary », « TIGER Index », « TIGER Ratio », « TIGER SVO Summary ».

Nous avons également développé un nouvel outil de visualisation des arbres syntaxiques sous deux formes au choix : TIGER ou UD. Cette visualisation est reliée hypertextuellement aux différents outils pertinents de TXM : Concordances, Édition, Index. L'utilisateur peut donc désormais facilement accéder à une représentation visuelle de l'arbre syntaxique de n'importe quelle phrase d'un corpus.

Ces nouveaux outils TXM ont été validés par des travaux communiqués au colloque international Diachro X <https://diachro10.sciencesconf.org> (Guillot et Lavrentiev 2022), et diffusés par le biais d'un atelier de formation lors du même colloque.

LIVRABLES

Tous les outils développés pour la plateforme TXM sont disponibles en ligne sous licence open-source GNU GPL V3 dans l'extension « Syntactic Annotation » version 1.0.0.202209131616 à installer dans le logiciel TXM pour poste <https://www.textometrie.org>.

Les sources de l'extension TXM « Syntactic Annotation » sont disponibles en ligne (dans six projets de composants) :

- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.conllu.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.core

- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.tigersearch.rcp
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.core
- http://forge.cbp.ens-lyon.fr/redmine/projects/txm/repository/show/TXM/trunk/org.txm.treesearch.rcp

Le corpus PROFITEROLE compatible TXM est disponible en ligne à la fois dans un portail TXM : https://txm-bfm.huma-num.fr (ce portail permet également de télécharger le corpus Profiterole TXM sous forme binaire et de le charger dans TXM pour poste) et sous forme de fichiers source '.conllu' (https://gitlab.huma-num.fr/profiterole/corpus-profiterole).
Pour exploiter le corpus Profiterole sous forme de fichiers source avec TXM il faut télécharger les fichiers .conllu depuis ces entrepôts et les importer avec le module d'import « CoNLL-U+CSV » de TXM pour poste.
Tous les outils disponibles dans l'extension TXM « Syntactic Annotation' » sont documentés dans le « Tutoriel de l'extension Annotation Syntaxique » [en cours de rédaction].

C.4 RESULTATS OBTENUS

Tous les livrables annoncés ont été livrés, ou sont en cours de livraison (correction en cours du corpus annoté). L'objectif de fournir des ressources (corpus et outils) pour le français médiéval a été atteint.

C.5 EXPLOITATION DES RESULTATS

La correction des annotations du corpus étant encore partielle, seules quelques études ont été jusqu'ici menées (voir références). L'accroissement progressif du corpus Gold va permettre de multiplier les études et donc la production scientifique associée au projet.

C.6 DISCUSSION

Pas de commentaires

C.7 CONCLUSIONS

En dépit de différents obstacles rencontrés (signalés dans le rapport intermédiaire et dans les demandes de report de l'échéance finale), j'estime que le projet a atteint tous ses objectifs de manière très satisfaisante. Les ressources produites constituent une contribution majeure d'un point de vue linguistique et patrimonial (corpus), tout en participant au développement des outils du TAL. Ces ressources devraient permettre des progrès notables dans notre compréhension des états anciens de la langue et des changements qu'a connus le français, de même que dans les processus généraux de changement linguistique.

C.8 REFERENCES

D LISTE DES LIVRABLES

Date de livraison	N°	Titre	Nature (rapport, logiciel, prototype, données, …)	Partenaires (souligner le responsable)	Commentaires
En cours	1	Corpus Profiterole	Données	Alexei Lavrentiev, Sophie Prévost, Céline Guillot, Benoît Crabbé, Loïc Grobol, Mathilde Regnault, Eric de la Clergerie, Mathieu Dehouck	
Février 2022	2	Parser MetaMOF	Logiciel	Mathilde Regnault	

Date de livraison	N°	Titre	Nature (rapport, logiciel, prototype, données, …)	Partenaires (souligner le responsable)	Commentaires
Février 2022	3	Parser HOPS	Logiciel	Benoît Crabbé, Loïc Grobol	
Février 2022	4	Parser Dyalog-SRNN	Logiciel	Eric de la Clergerie	
Février 2022	5	Parser Dyalog-SRNN/MetaMOF	Logiciel	Eric de la Clergerie	
Février 2022	6	Lexique OFrLex	Données	Gaël Guibon, Benoît Sagot	
Février 2022		Lexique OFrLex-BFMLEMGOLD	Données	Cristina Holgado, Mathilde Regnault	
Février 2022		TXM Extension syntaxique	Logiciel	Serge Heiden, Matthieu Decorde	

E IMPACT DU PROJET

E.1 INDICATEURS D'IMPACT

Nombre de publications et de communications (à détailler en E.2)

		Publications multipartenaires	Publications monopartenaires
International	Revues à comité de lecture		
	Ouvrages ou chapitres d'ouvrage		
	Communications (conférence)		6
France	Revues à comité de lecture		
	Ouvrages ou chapitres d'ouvrage		
	Communications (conférence)	1	5
Actions de diffusion	Articles vulgarisation		
	Conférences vulgarisation		
	Autres		3

Autres valorisations scientifiques (à détailler en E.3)

	Nombre, années et commentaires (valorisations avérées ou probables)
Brevets internationaux obtenus	
Brevet internationaux en cours d'obtention	
Brevets nationaux obtenus	
Brevet nationaux en cours d'obtention	
Licences d'exploitation (obtention / cession)	
Créations d'entreprises ou essaimage	
Nouveaux projets collaboratifs	
Colloques scientifiques	
Autres (préciser)	Développement de logiciels et extension de logiciels existants

E.2 LISTE DES PUBLICATIONS ET COMMUNICATIONS

International

Publications et communications monopartenaires :

Grobol, Loïc, Sophie Prévost, et Benoît Crabbé (2022). Is Old French Tougher to Parse? In Proceedings of the 20th International Workshop on *Treebanks and Linguistic Theories*, (TLT) <hal-03506500v1>.

Grobol, Loïc, Mathilde Regnault, Pedro Javier Ortiz Suárez, Benoît Sagot, Laurent Romary, et Benoit Crabbé (2022). BERTrade: Using Contextual Embeddings to Parse Old French. In Proceedings of the 13th *International Conference on Language Resources and Evaluation* (LREC) <hal-03736840v1>

Guibon, Gaël et Benoît Sagot (2020). OFrLex: A Computational Morphological and Syntactical Lexicon for Old French. In Calzolari, Nicoletta & al. (éd.) *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC) < https://aclanthology.org/2020.lrec-1.393/ >

Regnault, Mathilde, Sophie Prévost et Eric Villemonte de la Clergerie (2019). Adapting an Existing French Metagrammar for Old and Middle French. Présentation orale à *HICOV- Historical Corpora and Variation*, Mai 2019, Cagliari, Italie.

Regnault, Mathilde, Sophie Prévost, et Eric Villemonte de la Clergerie (2019). Challenges of language change and variation: towards an extended treebank of Medieval French. In 18th International Workshop on *Treebanks and Linguistic Theories* (TLT), Août 2019, Paris, France.

Stein, Achim (2020). Preserving Semantic Information from Old Dictionaries: Linking Senses of the "Altfranzösisches Wörterbuch" to WordNet. In Calzolari, Nicoletta & al. (éd.), *Proceedings of the 12th International Conference on Language Resources and Evaluation* (LREC).

France

Publications et communications monopartenaires

Grobol, Loïc, et Benoît Crabbé (2021). Analyse en dépendances du français avec des plongements contextualisés. In Actes de la 28e *Conférence sur le Traitement Automatique des Langues Naturelles* (TALN) <hal-03265893v1>.

Guillot Barbance, Céline et Alexei Lavrentiev (2022). Le renforcement morphologique du démonstratif en français : le cas des formes préfixées en i-. Présentation orale au colloque *Diachro X*, Mai 2022, Paris, France.

Holgado, Cristina, Alexei Lavrentiev et Mathieu Constant (2021). Évaluation de méthodes et d'outils pour la lemmatisation automatique du français médiéval. In Actes de la 28e *Conférence sur le Traitement Automatique des Langues Naturelles* (TALN). <hal-03265897>

Regnault, Mathilde. (2019). Adaptation d'une métagrammaire du français contemporain au français médiéval. In Actes de la 26e édition de la *Conférence sur le Traitement Automatique des Langues Naturelles* (TALN) et 21e édition de la *Conférence jeunes chercheur·euse·s RECITAL* ⟨hal-02147686⟩

Sagot, Benoît (2019). Développement d'un lexique morphologique et syntaxique de l'ancien français. In Actes de la 26e *Conférence sur le Traitement Automatique des Langues Naturelles* (TALN) ⟨hal-02148701v2⟩

Publications et communications multipartenaires

Prévost, Sophie, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev et Serge Heiden (2022). Profiterole : un corpus morpho-syntaxique et syntaxique de français médiéval. Présentation orale au Colloque *La constitution de corpus en diachronie longue : méthodologies, objectifs et exploitations linguistiques et stylistiques*, oct. 2022, Grenoble France (publication prevue).

Autres

Grobol, Loïc (2021). La Queste del Analyseur: leaving no data untouched to parse Old French. Présentation invitée au *Forschungskolloquium* de l'Institut für Linguistik/Romanistik de Stuttgart, Juin 2021.

Regnault Mathilde, Sophie Prévost et Eric Villemonte de la Clergerie (2018). Poster de présentation du projet Profiterole au Salon de l'Innovation de la conférence TALN, mai 2018, Rennes, France.

Zeina Tmart, Alexei Lavrentiev, Céline Guillot, Sophie Prévost (2022). Atelier sur les Outils pour l'exploration lexicale, morphosyntaxique et syntaxique de la Base de français médiéval. *Diachro X*, Mai 2022, Paris, France ⟨halshs-03788509⟩

E.3 LISTE DES ELEMENTS DE VALORISATION

Le projet a produit des analyseurs syntaxiques neuronaux et symbolique et des extensions pour le logiciel et la plateforme TXM (voir plus haut Résultats majeurs du projet et Approche scientifique du projet (C3)).

E.4 BILAN ET SUIVI DES PERSONNELS RECRUTES EN CDD (HORS STAGIAIRES)

Tous les personnels non permanents recrutés ont trouvé un poste ou un contrat doctoral à l'issue du projet.

Identification				Avant le recrutement sur le projet			Recrutement sur le projet				Après le projet				
Nom et prénom	Sexe H/F	Adresse email (1)	Date des dernières nouvelles	Dernier diplôme obtenu au moment du recrutement	Lieu d'études (France, UE, hors UE)	Expérience prof. Antérieure, y compris post-docs (ans)	Partenaire ayant embauché la personne	Poste dans le projet (2)	Durée missions (mois) (3)	Date de fin de mission sur le projet	Devenir professionnel (4)	Type d'employeur (5)	Type d'emploi (6)	Lien au projet ANR (7)	Valorisation expérience (8)
REGNAULT Mathilde	F	regnaultm@orange.fr	5/09/2022	Master	France	0	ENS Paris	Doctorante	36	Février 2022 (août 2020 pour le financement	CDD 6 ans (01/09/20)	Enseignement et recherche publique (Université de Stuttgart, Allemagne)	Enseignant-chercheur (tenure track)	Non	-
GROBOL Loïc	H	loic.grobol@gmail.com	5/09/2022	Doctorat	France	0	ENS Paris	Post-doctorant	23	Aout 2021	CDI (01/09/21)	Enseignement et recherche publique (Université Paris-Nanterre)	Enseignant-chercheur	Non	-
GUIBON Gaël	H	gael.guibon@gmail.com	5/09/2022	Doctorat	France	0	INRIA	Post-doctorant	12	Juin 2020	CDI (01/09/22)	Enseignement et recherche publique (Université de Lorraine)	Enseignant-chercheur	Non	-
DECORDE Mathieu	H	matthieu.decorde@ens-lyon.fr	5/09/2022	Master	France	Ingénieur d'études	ENS Lyon	Ingénieur d'études	6	Mai 2019	CDI (01/12/20)	Enseignement et recherche publique (ENS de Lyon)	Ingénieur d'études	Oui	Oui
GARCIA-HOLGADO Cristina	F	cristina.garcia-holgado@etu.unistra.fr	30/09/2022	Master	France	0	ENS Paris	Ingénieure d'études	5,5	Février 2022	Inscription en doctorat en cours	Enseignement et recherche publique	Contrat doctoral	non	Oui