**Introduction à la lemmatisation pour médiévistes**

**Contributeurs:**

**Eliana Magnani**

**Nicolas Perreaux**

**Ariane Pinche**

**Sommaire**

[Qu’est-ce la lemmatisation ? 1](#_Toc50027227)

[Pour quoi faire ? 1](#_Toc50027228)

[Quels outils ? 2](#_Toc50027229)

[Par où commencer ? 2](#_Toc50027230)

[Bibliographie sélective 2](#_Toc50027231)

[Glossaire du traitement automatique de la langue - Yves Ouvrard, mars 2019. 3](#_Toc50027232)

[Outils. Fiches techniques 8](#_Toc50027233)

[Collatinus 11.1 8](#_Toc50027234)

[PALM 10](#_Toc50027235)

[Pie (tagueur) + Pyrrha (interface de correction et de préannotation) 12](#_Toc50027236)

## Qu’est-ce la lemmatisation ?

La lemmatisation est le processus de déflexionalisation des mots pour déterminer leur lemme (comme dans une entrée de dictionnaire). Les mots flexionnés sont appelés lexèmes.

Dans le cas des langues flexionnelles et à forte variation graphique, comme celles employées dans l’Occident médiéval – le latin et les langues vernaculaires –, le développement de procédures de recherche formalisées et assistées informatiquement implique la lemmatisation des textes utilisés. Ce procédé – manuel, semi-automatisé ou automatisé –, n’est pas chose nouvelle, mais au cours des dernières années, plusieurs lemmatiseurs ou de paramètres pour la lemmatisation des langues médiévales ont été créés.

## Pour quoi faire ?

* + Exemples concrets d’utilisation pour la philologie et/ou l’histoire :

Lemmatiser et annoter la morphosyntaxe un corpus permet par la suite de l’interroger pour le trier ou en vue d’analyses statistiques, opérations, souvent, aussi bien nécessaires au linguiste, qu’à l’éditeur scientifique qu’à l’historien. Par exemple à partir d’un corpus lemmatisé, il est facile de constituer un glossaire qui répertorie sous une même entrée toutes les variations graphiques d’un mot, en opérant un tri à partir du lemme de l’annotation qui, lui, est unique pour chaque mot du lexique.

L’analyse syntaxique d’un corpus étendu devient également possible pour détecter certains traits dialectaux. Par exemple, en ancien Picard, il arrive que le pronom personnel objet féminin se graphie *le* comme le pronom masculin, au lieu de *la*. L’annotation morphosyntaxique permet alors de calculer rapidement les statistiques d’apparition du phénomène dans les textes pour le dater ou déterminer les territoires où il est le plus fréquent. À l’échelle d’un texte, ces calculs permettent à l’éditeur de savoir si le phénomène est à la marge de son corpus ou pas, pour décider de corriger ou de conserver dans son édition la graphie en déterminant si elle relève du trait dialectal ou de l’erreur de copie.

Lemmatisation et analyse sémantique

## Quels outils ?

D’un point de vue technique, la lemmatisation automatique regroupe plusieurs opérations : l’acquisition du ou des textes dans un format numérisé élémentaire ; la tokenisation, c’est-à-dire le découpage du texte par lexème, au cours de laquelle le texte peut être pré-formaté selon les choix et besoins des concepteurs des outils (par exemple, nettoyage des caractères spéciaux, séparation d’enclitiques) ; l’étiquetage morphosyntaxique des formes (*POS tagging = part-of-speech tagging*) avec un jeux d’étiquettes très variable d’un outil à autre ; et le regroupement des formes sous le lemme correspondant. Plusieurs difficultés doivent être surmontées dans ce processus, notamment la désambiguïsation des homographes et les variations graphiques des formes.

Les outils actuels pour associer aux formes une étiquette (POS) correcte, se partagent en deux groupes principaux.

Les premiers utilisent des tagueurs probabilistes basés sur un lexique et des règles prédéfinies associés ou pas à des entraînements successifs qui améliorent la reconnaissance des formes et des lemmes (p. ex. TreeTagger avec les paramètres pour le latin médiéval OMNIA).

Les seconds sont basés sur les technologies plus récentes dites des « réseaux de neurones » (ou *deep learning*), des algorithmes qui, à partir d’un corpus pré-annoté (ou corpus d’entraînement) visent à apprendre l’application à créer les lemmes qui ne figurent pas dans le corpus d’entraînement, en fonction de leur représentation sémantique (les mots coocurrents) (p. ex. Pie, Hydra).

Dans les deux cas, un travail manuel en amont est nécessaire, l’établissement d’un lexique et des règles, ou l’annotation d’un corpus. En aval aussi l’intervention manuelle est nécessaire dans les deux cas pour corriger les corpus annotés par les lemmatiseurs.

## Par où commencer ?

étape par étape depuis le texte numérisé (avec note sur état critique de l’édition utilisée) (exemple de la formation d’octobre 2019 ?)

* + Acquérir d’un texte numérisé, récupéré sur internet, océrisé ou transcrit par le chercheur
  + Selon la qualité du texte numérisé, un nettoyage est nécessaire (par exemple, remplacer « & » par « et » ou « ae »/ « oe » selon les éditions)
  + Enregistrer le texte avec l’encodage UTF-8
  + Tokenisation : découpage du texte par lexème, un mot ou un signe de ponctuation par ligne (par exemple, tokeniseur OMNIA : <https://glossaria.eu/lemmatisation/#page-content> )
  + L’étiquetage morphosyntaxique des formes (*POS tagging* = part-of-speech *tagging*) (par exemple, le jeux d’étiquettes OMNIA, l’un des plus simples : ADV CON INT NAM NUM PON PRE PRO QLF SUB VBE SENT
  + vérification/correction ??

## Bibliographie sélective

* Roberto Busa, « The Annals of Humanities Computing : the Index thomisticus », *Computers and the Humanities,* 14, 1980, p. 83-90.
* Yves Ouvrard et Philippe Verkerk, « Collatinus, un outil polymorphe pour l’étude du latin », *Archivum Latinitatis Medii Aevi*, 72, 2014, p. 305-311.
* Steffen Eger, Rüdiger Gleim, Alexander Mehler, « Lemmatization and Morphological Tagging in German and Latin : A comparison and a survey of the state-of-the-art ». *Proceedings of the 10th International Conference on Language Resources and Evaluation*, 2016 (<http://www.lrec-conf.org/proceedings/lrec2016/pdf/656_Paper.pdf>)
* Steffen Eger, Tim vor der Brück, Alexander Mehler, « Lexicon-assisted tagging and lemmatization in Latin : A comparison of six taggers and two lemmatization methods », *Proceedings of the 9th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH),* Beijing, 2015, p. 105–113.
* Łukasz Gągała, « Authorship Attribution with Neural Networks and Multiple Features : Notebook for PAN at CLEF 2018 », In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, Laure Soulier (éd.) *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. Avignon, France, September 10-14, 2018, (<http://ceur-ws.org/Vol-2125/paper_146.pdf>).
* Louis Delatte, Étienne Évrard, « Un laboratoire d'analyse statistique des langues anciennes à l'Université de Liège », *L’Antiquité classique*, 30-2, 1961, p. 429-444.
* Bruno Bon, « OMNIA – Outils et Méthodes Numériques pour l’Interrogation et l’Analyse des textes médiolatins », BUCEMA - Bulletin du centre d’études médiévales d’Auxerre, 13, 2009, p. 291-292 (<http://journals.openedition.org/cem/11086>).
* Bruno Bon,, « OMNIA: outils et méthodes numériques pour l’interrogation et l’analyse des textes médiolatins (2) », BUCEMA - Bulletin du centre d’études médiévales d’Auxerre, 14, 2010, p. 251-252 (<http://journals.openedition.org/cem/11566>).
* Bruno Bon,, « OMNIA: outils et méthodes numériques pour l’interrogation et l’analyse des textes médiolatins (3) », BUCEMA - Bulletin du centre d’études médiévales d’Auxerre, 15, 2011, p. 333-334 (<http://journals.openedition.org/cem/12015>).
* Mourad Aouini, « A NooJ Module for Named Entity Recognition in Middle French Texts », In Johanna Monti, Max Silberztein, Mario Monteleone, Maria Pia di Buono (éd.), *Formalising Natural Languages with Nooj*  *2014*, Cambridge, 2015, p. 99-112.
* Mike Kestemont, Guy de Pauw, Renske van Nie, Walter Daelemans, « Lemmatization for variation-rich languages using deep learning », *Digital Scholarship in the Humanities*, 32-4, 2017, p. 797-815, (<https://doi.org/10.1093/llc/fqw034>).
* Enrique Manjavacas, Akos Kadar, Mike Kestemont, « Improving Lemmatization of Non-Standard Languages with Joint Learning », arXiv:1903.06939v1, 2019 (<https://arxiv.org/format/1903.06939v1>).

## Glossaire du traitement automatique de la langue - Yves Ouvrard, mars 2019.

Ce document peut être reproduit par n'importe quel moyen que ce soit, pourvu qu'aucune modification ne soit effectuée et que cette notice soit préservée.

* **Analyse morphologique** — L'analyse morphologique consiste à donner tous les traits morphologiques d'une forme. Par exemple, en français, le mot *aviez* : deuxième personne, pluriel, indicatif, actif. L'analyse morphologique est habituellement accompagnée de la lemmatisation de la forme et de son [POS](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#pos) : lemme *avoir*, verbe, 2ème personne du pluriel de l'indicatif actif. Souvent, un mot a plusieurs analyses morphologiques possibles. C'est le cas, par exemple, de la forme *avions* pour laquelle on donnera :
  + lemme *avoir*, verbe, 1ère personne du pluriel de l'indicatif actif.
  + lemme *avion*, nom, pluriel.
* **Code source** — L'ordinateur est incapable d'exécuter des ordres exprimés en [langage naturel](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#naturel) Il faut lui fournir une suite de uns et de zéros que l'humain a le plus grand mal à lire et à écrire. Pour obtenir un programme utilisable sur un ordinateur, le programmeur s'exprime donc dans un [langage іnformatique](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#langage) assez ressemblant au langage naturel, mais dont le lexique est très pauvre, et la syntaxe d'une rigueur absolue. Ce qu'il écrit alors est le **code source** du programme. Un programme appelé *compilateur* transforme ensuite le code source en code exécutable, cette suite de uns et zéros. Une autre possibilité est d'utiliser un *langage interprété*. Ce sera alors un *interpréteur* qui transforme directement le code source en instructions exécutables par la machine.
* **compilateur, interpréteur** — Opération qui consiste à transformer le [code source](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#code) en code exécutable par la machine. Le programme chargé de la compilation est un compilateur. L'interpréteur, au lieu de générer un exécutable, lit le code source et le transforme immédiatement en instructions exécutables.
* **encodage** — Aux débuts de l'informatique, la mémoire était limitée et coûteuse, et l'on s'est contenté de 127 caractères, qui tenaient dans sept bits (zéros et uns) : les chiffres, les lettres de l'alphabet latin, majuscules et minuscules, et quelques ponctuations. Ce premier encodage, nommé ASCII (American Standard Code for Information Interchange), a vite été remplacé par des jeux de caractères plus riches (255 caractères), mais qui ont le défaut d'être “localisés” (Europe Occidentale, Europe Centrale etc…). On a fini par aboutir au standard Unicode, qui possède 137 929 caractères (il a la possibilité d'en gérer plus d'un million). Avec les jeux de caractères “primitifs” (moins de 256 caractères), l'encodage se faisait systématiquement sur un octet qui est, de fait, l'élément de base pour le stockage des données. Pour que les caractères accentués apparaissent correctement, il fallait être capable d'indiquer quel jeu de caractères avait été utilisé. Avec l'adoption de l'Unicode, est apparue le problème de l'encodage : comment stocker une information qui aurait besoin, a priori, de 16 voire 32 bits dans des systèmes où la brique élémentaire en compte huit ? Plusieurs solutions, aux noms barbares (UTF-16LE, UTF-8, UTF-7), ont été proposées. Aujourd'hui, l'UTF-8 est en passe de devenir **LE** standard. Avant tout traitement automatique, il est prudent de connaître l'encodage du texte à analyser. Tout en étant conscient que le problème se pose surtout quand interviennent des caractères accentués ou des caractères non-latins (grecs ou cyrilliques).
* **entrée (dictionnaire)** — Dans le dictionnaire, il y a en général une entrée par lemme. L'entrée débute par la [forme canonique](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#canon) du mot, suivie par des indications morphologiques (POS, génitif, $ temps primitifs), l'étymologie, les différentes traductions, des exemples, et les formes irrégulières.
* **épenthèse** — L'épenthèse est l'apparition, à l'intérieur d'un mot, d'une consonne destinée à faciliter sa prononciation. En latin médiéval, le classique *damnum* est devenu *dampnum* ; *solemnitas* est devenu *solempnitas*.
* **flexion, paradigmes, modèles** — La flexion d'un lemme est sa capacité à être écrit et prononcé de plusieurs manières. La flexion obéit le plus souvent à un ensemble de règles qu'on appelle [modèle](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#modele). Le modèle d'un lemme est donné par le dictionnaire, en tête d'entrée, d'une manière implicite.
* **formats électroniques, encodage** — Il est devenu très facile de se procurer des textes, de toute époque et dans toute langue. Mais pour travailler sur un texte, il faut d'abord savoir quel est son *format*, c'est à dire quelle méthode a été utilisée pour l'enregistrer sur une machine. Voici une liste simplifiée des formats de texte :
  + Le texte pur, dont l'[encodage](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#encodage) peut varier. De plus en plus, les textes sont encodés en UTF-8 ;
  + Le texte balisé (html, LaTeX, markdown) permet de changer de la couleur, la disposition, la taille, la police, la graisse, le style, etc., à n'importe quel endroit du texte. Le plus souvent, ces indications non textuelles sont insérées dans des balises ouvrantes et fermantes. Dans le format html, les balises sont les caractères < et >.
  + Les fichiers issus des traitements de texte (odt LibreOffice, docx Word), dont les règles sont souvent très complexes, et la plupart du temps inutilisables tels quels par les logiciels d'analyse automatique. Pour remédier à cet inconvénient, il faut convertir ces fichiers en fichiers texte, soit grâce aux convertisseurs internes (enregistrer sous, exporter), soit grâce à des convertisseurs externes spécialisés.
* **forme** — Une forme est l'un des éléments de la [flexion](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#flexion) d'un lemme. Lorsqu'on conjugue un verbe, on cite l'une après l'autre toutes les formes du verbe. Chaque forme peut être étiquetée par une série de [traits morphologiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#trait) : genre, nombre, cas, personne, mode, temps, voix, etc. Certaines [POS](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#pos) n'ont qu'une forme, comme la préposition latine *ad*. D'autres en ont plus d'une centaine, comme les verbes.
* **Forme canonique** (ou mot-vedette, en anglais catchword) — Un [lemme](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#lemme) peut avoir un très grand nombre de [formes](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#forme), mais aussi parfois quelques [variantes graphiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#vargraph), par exemple *negligo* à côté de *neglego*. Pour désigner ce lemme, on utilise l'une de ses formes, qui correspond à une [morphologie](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#morphologie) précise, choisie par consensus entre les grammairiens. Pour le français, on a choisi le singulier des noms, le masculin singulier des adjectifs, l'infinitif présent actif des verbes. ex. *aimer*. Pour le latin, on donne la première personne du singulier du présent de l'indicatif actif ex. *amo*.
* **langage informatique** — Un langage informatique est un langage conçu spécialement pour être transformé en code exécutable, le seul auquel la machine puisse obéir. Très peu nombreux au début de l'ère informatique, ces langages se sont multipliés. Presque tous empruntent leur lexique à l'anglais. Leur syntaxe est extrêmement rigide, et la moindre erreur les rend inutilisables. Lorsque le programmeur a écrit un [code source](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#code), il utilise un [compilateur](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#compilation) ou un interpréteur, programmes spécialisés qui permettent de convertir le code source en code exécutable.
* **langage naturel** — Le langage naturel est celui que les humains ont toujours utilisé pour communiquer, alors que les [langages informatiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#langage) ont été créés de toutes pièces par les ingénieurs.
* **Lemmatiser, lemmatiseur** — Lemmatiser un mot, c'est trouver quel [lemme](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#lemme) l'a produit. Assez souvent, il y a plusieurs solutions, comme pour le mot *avions*, donné en exemple à l'article [analyse morphologique](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#analyse_morpho). Il y a donc deux sens à *lemmatiser* : soit trouver **tous** les lemmes pouvant produire une forme donnée, soit trouver **le** lemme que l'auteur du texte a utilisé pour produire cette forme.   
  La lemmatisation d'un texte se fait en plusieurs étapes :
  + La [tokenisation](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#tokenisation) consiste à transformer le texte en une liste de formes. Cette opération n'est pas très difficile, mais l'encodage en UTF-8 réserve parfois des surprises, en utilisants deux caractères pour n'en afficher qu'un.
  + La recherche des suffixes étrangers au lemme (en latin, *-que, -ue, -ne*) ;
  + La prise en compte de la graphie [ramiste](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#ramus), des [variantes graphiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#vargraph)
  + La lemmatisation proprement dite : quels lemmes peuvent donner cette forme ?
  + En cas de réponse multiple, le classement des résultats, en commençant par le plus probable.
* **lemme** — Un **lemme** est l'unité constituante d'un lexique. En latin, un lemme peut se rencontrer sous diverses [formes](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#forme), parfois très différentes les unes des autres. Par exemple, les quatre formes *fers, ferre, latos, tulerunt* appartiennent au même lemme *fero*. Des [règles syntaxiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#syntaxe) permettent de décider quelle forme du lemme employer. Pour saisir le sens d'un énoncé, il est indispensable de savoir identifier la [morphologie](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#morphologie) d'une forme, opération très rapide et inconsciente la plupart du temps.
* **lexicométrie** — La lexicométrie est l'étude de la quantification des lemmes dans un corpus, et de leur répartition. Elle est utilisée dans de nombreux buts : identifier l'auteur d'un texte, le dater, le situer, extraire des connaissances, comparer, etc.
* **Modèle** — Un modèle est un ensemble de règles qui permet de fléchir un lemme. On ne pourra donc ni conjuguer ni décliner avant de savoir quel modèle appliquer. Parmi ces règles, certaines donnent la méthode pour calculer un [radical](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#radical), d'autres disent quelle [désinence](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#desinence) ajouter à quel [radical](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#radical) pour obtenir la forme recherchée. Un modèle reçoit le nom d'un lemme très employé appliquant ce modèle. Par exemple : *rosa, templum, amo* Il vaut mieux que ce lemme n'ait aucune ambiguïté. Par exemple, le nom *amicus* est aussi un adjectif. Mieux vaut choisir *lupus*, qui est toujours un nom. Les dictionnaires papier ne donnent le modèle que de manière implicite. Par exemple, au lieu de dire que le lemme *cubitum* suit le modèle *templum*, le dictionnaire donne le génitif (i) et le genre (n.) : **cubitum**, i, n. : coude. La grammaire académique propose un nombre réduit de modèles. Les lemmatiseurs latins en utilisent plus d'une centaine. Dans la flexion des lemmes les plus courants, on trouve la plupart du temps des formes irrégulières, qui désobéissent aux règles de leur modèle.
* **morphologie** — Liste des [traits morphologiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#trait) caractéristiques d'une forme. Voici un tableau simplifié des [traits morphologiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#trait) en fonction des différents [POS](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#pos), pour la langue latine :
  + **nom** : cas, nombre
  + **pronom** : cas, genre, nombre
  + **adjectif** : cas, genre, nombre, degré
  + **adverbe** : degré
  + **verbe** : personne, nombre, temps, mode, voix.
  + **verbe, formes adjectivales** : cas, genre, nombre, temps, mode, voix   
    Exemples :
    1. *sustulisti* : *tollo*, 2ème personne du pluriel, parfait indicatif actif ;
    2. *gestas* : *gero*, accusatif féminin pluriel participe parfait passif.
    3. *cupidissimorum* : *cupidus*, génitif masculin (ou neutre) pluriel, superlatif
* **OCR** — Optical Character Recognition, fr. ROC (peu utilisé) L'OCR est un procédé qui consiste à photographier un texte, et à transformer l'image obtenue en texte, qu'on peut àlors corriger, et envoyer dans un lemmatiseur. L'OCR est loin d'être infaillible, et des relectures humaines sont nécessaires pour corriger un texte *océrisé*.
* **POS** — acronyme pour Part Of Speech. En français, il n'y a pas de consensus. On trouve *catégorie grammaticale*, *classe grammaticale*, *nature*, et même *partie du discours*. On peut définir de deux manières le concept de classe grammaticale :
  + par les [traits morphologiques](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#trait) de sa [flexion](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#flexion) ;
  + par le lien du lemme avec le monde réel : chose ou être (nom), propriété (adjectif), prédicat (verbe).
* **préfixe, suffixe** — Éléments constitutifs du lemme qui ѕe collent au début ou à la fin d'un mot. Par exemple, le verbe *clamo* peut recevoir
  + un préfixe : *declamo*
  + un suffixe : *clamito*
  + les deux : *conclamito*   
    Le latin connaît un type particulier de suffixe, qui, au lieu de modifier le sens du mot, lui ajoute un second mot :
  + le suffixe *-que*, qui équivaut à *et* + le mot : *rogandisque* est une autre manière d'écrire *et rogandis*.
  + le suffixe *-ne*, rend la phrase interrogative. *Ad amicos confugiam* Je me réfugierai auprès de mes amis. *ad amicosne confugiam ?* Me réfugierai-je auprès de mes amis ?
  + le suffixe *-ue*, pour indiquer une alternative : *bis terue* deux ou trois fois.
* **prosodie, métrique, quantité, scansion, accentuation** — Le latin possède des syllabes longues et des syllabes brèves. Les mots sont accentués en fonction de la disposition de ces longues et brèves. L'étude de cette prosodie peut être intéressante pour étudier la poésie et les clausules. Il est possible de scander automatiquement un texte, de l'accentuer, et de repérer ses clausules (suite caractéristique de longues et de brèves à la fin d'une phrase).
* **radical, désinence** — Le radical est la partie du [lemme](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#lemme) qui ne change pas lorsqu'on le [fléchit](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#flexion). La désinence est ce qu'il faut ajouter au radical pour obtenir une forme. Un lemme peut avoir plusieurs radicaux. Les verbes latins en ont habituellement trois : le radical d'infectum, le radical de perfectum, et le radical de supin. Ces radicaux sont donnés par le dictionnaire sous forme de *temps primitifs*, et dans le cas du modèle *amo*, on peut les calculer en suivant des règles ѕimples.
* **Ramus** — Pierre de la Ramée (1515 – 1572), outre une œuvre philosophique considérable, est connu pour avoir systématiquement différencié les deux prononciations des lettres *u* et *i*. Par *iuuenis*, en graphie ramiste, devient *juvenis*. Les manuels du secondaire et les dictionnaires sont en graphie ramiste. La plupart des éditions critiques sont restées en graphie ancienne.
* **syntaxe, analyse syntaxique** — La syntaxe d'une lange est l'ensemble des règles qui permettent de générer un énoncé grammatical dans cette langue. L'analyse syntaxique automatique est possible, mais obtient de moins bons résultats que [l'analyse morphologique](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#analyse_morpho).
* **tagger, tagueur** — Ou étiqueteur morpho-syntaxique. Les lemmatiseurs classiques sont incapables, en présence de plusieurs solutions de lemmatisation, de choisir laquelle est celle que l'auteur a voulu employer, et que le lecteur intelligent reconnaît sans difficultés. Aussi le lemmatiseur est souvent aidé par un tagueur, qui utilise des statistiques obtenues à partir d'un corpus d'entraînement. C'est un lecteur humain qui procède à un premire étiquetage, et l'ordinateur prend la suite.
* **tokenisation, token** — La tokenisation consiste à transformer le texte en une liste de formes, ou *tokens*. Cette opération n'est pas très difficile, mais l'encodage en UTF-8 réserve parfois des surprises, en utilisant deux caractères pour n'en afficher qu'un.
* **trait morphologique** — Quelle forme choisir lorsqu'on désire employer un lemme dans un énoncé ? Par exemple, parmi les formes du lemme *beau* [belles, belle, beaux, beau], laquelle choisir dans la phrase *Que la campagne est b…* ? Pour cela, on va utiliser une série de traits morphologiques qui essaient de correspondre à la réalité :
  + le genre : *campagne* est-il masculin ou féminin ?
  + le nombre : parle-t-on d'une ou de plusieurs campagnes ?   
    Ce reflet de la réalité est très imparfait. Voici la liste des traits morphologiques utilisés par le latin. Intentionnellement, les listes sont en ordre alphabétique :
  + cas : *ablatif, accusatif, datif, génitif, locatif, nominatif, vocatif*
  + genre : *féminin, masculin, neutre*
  + nombre : *pluriel, singulier*
  + personne : *deuxième, première, troisième*
  + degré : *comparatif, positif, superlatif*
  + temps : *futur, futur antérieur, imparfait, parfait, plus-que-parfait, présent*
  + mode : *adjectif verbal, indicatif, gérondif, impératif, infinitif, participe, subjonctif, supin*
  + voix : *actif, passif*   
    Chaque [catégorie grammaticale](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#pos) utilise un ou plusieurs de ces traits morphologiques. L'article *POS* en donne la liste pour chaque catégorie.
* **variante graphique** — Le latin a une très longue histoire, au cours de laquelle les principes de notation ont beaucoup évolué, principalement pour refléter l'évolution phonétique. On peut distinguer :
  + L'**assimilation**, qui transforme deux consonnes successives en une consonne double : *adcedo* devient *accedo*, *conminus* devient *comminus*.
  + La **contraction** est la perte d'une partie du mot qui ne se prononce plus : *amaueram* devient *amaram*, *adscendo* devient *ascendo*, *periculum periclum*.
  + une **abréviation** est la notation d'une partie du mot, le début ou la fin. *consules* s'écrit souvent *COSS*, *februarius* devient *febr*. La plupart des prénoms sont abrégés.
  + L'**agglutination** consiste à coller deux mots successifs pour n'en former qu'un : *et enim* devient *etenim*, *sic ut* devient *sicut*.
  + L'[**épenthèse**](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#epenthese). En outre, certaines voyelles se sont fermées ou ouvertes, les diphtongues se sont simplifiées en une seule voyelle : *ameicus* devient *amicus*, *auorsor, auersor*. Les consonnes aussi ont évolué : *colos* devient *color*, *caelebs* s'écrit *caeleps*, etc. Un lemmatiseur doit pouvoir identifier et traiter ces variantes graphiques. À partir du XVIe siècle, sous l'impulsion de [Ramus](file:///Users/elianamagnani/Documents/COSME/Cosme2/Groupe%20LEMMES/glossaire/glossaire.html#ramus), on a ajouté deux lettres à l'alphabet latin, *j* et *v*, afin de distinguer les deux prononciations du *i* et du *u*. Dans les noms propres germaniques, la ligature de deux *v* successifs, *vv*, est devenue la lettre *w*.

## Outils. Fiches techniques

### Collatinus 11.1

Outil multifonctionnel d'analyse de la langue latine

* *accessibilité* : téléchargement sur https://outils.biblissima.fr/fr/collatinus
* *interface* : graphique ; un cadre avec le texte et des onglets pour les différentes fonctions.
* *systèmes* : Windows, MacOS X ; fin octobre 2018 disponible aussi pour Linux/Debian
* *licence* : GNU GPL3
* *langage* : le code en C++ (avec les bibliothèques Qt5) a été découpé en "modules" qui peuvent être réutilisés dans une application spécialisée. Une documentation en a été faite avec Doxygen.
* *langue* : Latin
* *fonctions* : Lemmatisation, flexion, scansion et consultation de dictionnaires. Tous les lemmes ont une traduction en Français et/ou en Anglais. Six autres langues sont disponibles pour une partie des lemmes.
* *tagueur* : second order Hidden Markov Model ; export direct d'un texte tagué sous forme d'un fichier csv.
* *jeux d’étiquettes* : 91 étiquettes sur trois caractères ; 10 PoS (premier caractère), dont un associé au mode et 5 associés aux cas et nombre.
* *corpus d’entrainement* : textes lemmatisés au LASLA (1,7 millions de formes)
* *réentrainement/personnalisation* : non prévus, mais possibles (extérieurs à Collatinus, ce dernier tirant toutes les informations d'un seul fichier tags.la qui contient le nombre d'occurrences de chaque étiquette et de chaque séquence de trois étiquettes).
* *import* : format texte ou copier-coller dans la fenêtre (possibilité de nettoyer des signes diacritiques non-voulus.
* *export* : format texte ou copier-coller depuis la fenêtre ou export au format csv du résultat de la lemmatisation ou celui du tagueur.
* *serveur interne* : permet d'interroger Collatinus depuis un autre programme.
* *dernière version* : 11.1 (octobre 2018)
* *site web* : https://outils.biblissima.fr/fr/collatinus
* *documentation* : sous forme de pages html incluses dans le paquet d'installation.
* *contacts* : Yves Ouvrard et Philippe Verkerk (collatinus@biblissima-condorcet.fr)

**Détails sur le tagueur**

* *Le jeu d'étiquettes* : # , a , a11, a12, a21, a22, a31, a32, a41, a42, a51, a52, a61, a62, a71, a8 , c , ce , d , de , i , m , m11, m12, m21, m22, m31, m32, m41, m42, m51, m52, m61, m62, m8 , n , n11, n12, n21, n22, n31, n32, n41, n42, n51, n52, n61, n62, n71, n8 , p , p11, p12, p21, p22, p31, p32, p41, p42, p51, p52, p61, p62, p71, p8 , r , re , snt, v0 , v1 , v11, v2 , v21, v3 , v31, v4 , v41, w , w11, w12, w21, w22, w31, w32, w41, w42, w51, w52, w61, w62, x ; certaines sont des reliquats des codes du LASLA et ne sont pas utilisées par Collatinus.
* Le *PoS* est donné par la première lettre.

a : Adjectif

c : Conjonction ("ce " = conjonction enclitique = -que ou -ve)

d : aDverve ("de " = adverbe enclitique = -ne)

i : Interjection

m : noMbre

n : Nom

p : Pronom

r : pRéposition ("re " = préposition enclitique = -cum)

v : Verbe

w : forme verbale déclinée (participe, adjectif verbal etc...)

* Les *cas* sont numérotés dans l'ordre habituel (en Français) 1 : nominatif ... 6 : ablatif ; 7 : locatif ; 8 : indéclinable.
* Les *modes* sont

1 : indicatif

2 : subjonctif

3 : impératif

4 : tous les autres (sauf formes déclinées)

* *Commentaires* :

- Le 1 qui suit éventuellement le mode singularise le présent. Sans ce 1, cela regroupe tous les autres temps. L'idée était d'essayer de désambiguiser les formes du présent et du parfait à la 3e personne du singulier (et 1ère pl.) pour les verbes comme lego (lĕgĭt/lēgĭt). Pas sûr que ça marche.

- Le nombre a été conservé pour désambiguiser sing./pl. identiques dans quelques cas (en particulier la 4e déclinaison dŏmŭs/dŏmūs). Pas sûr que ça marche.

- Conclusion, on pourrait peut-être réduire le nombre d'étiquettes à une quarantaine.

### PALM

Accessibilité: Une plateforme web

Interface: Basé sur interfaces hommes-machines (front-end) permettant la consultation d’une bibliothèque, la gestion d’un corpus de travail, l’annotation semi-automatique et l’export d’un corpus.

Systèmes: Des navigateurs supportés par les système d’exploitation Windows, Linux et MacOS X essentiellement firefox et google-chrome?

Tagueur: Utilisation des modèles Treetagger pour trois langues à savoir moyen français, moyen anglais et latin.

Remarque : Un système à base de règles NooJ pour le moyen français et un système à base de réseaux de neurones pour les trois langues est en cours d’implémentation.

Langue(s): Moyen français, moyen anglais et latin médiéval

Jeux d’étiquettes: Jeu d’étiquettes définit selon la langue.

PALM propose les parties du discours suivantes pour le latin:

Verbe

Adverbe

Nom commun

Nom propre

Pronom

Adjectif

Préposition

Conjonction de subordination

Conjonction of coordination

Nombre

Ponctuation

Interjection

PALM propose les parties du discours suivantes pour le moyen français:

Verbe

Adverbe

Nom commun

Nom propre

Pronom

Adjectif

Déterminant

Préposition

Conjonction of subordination

Conjonction of coordination

Nombre cardinal

Nombre ordinal

Interjection

Ponctuation

Les parties du discours suivantes ont été utilisées pour le moyen anglais:

Verbe

Adverbe

Nom commun

Nom propre

Pronom

Nom verbal

Adjectif

Déterminant

Préposition

Conjonction

Nombre cardinal

Nombre ordinal

Infinitive Marker

Verbe+Pronom

Interjection

Ponctuation

Corpus d’entrainement: Corpus MEDITEXT

MEDITEXT est un corpus textuel d’abord rassemblé sous la direction de Jean-Philippe Genet et Claude Gauvard. Il a été corrigé et augmenté dans le cadre du projet du European Research

Council « Signs and States » de 2010 à 2014 (voir ci-dessous, paragraphe 1.8). Il est la base de la bibliothèque interne de PALM.

MEDITEXT, et par conséquent la bibliothèque interne de PALM, rassemble essentiellement des textes « politiques » : c’est-à-dire, soit des textes ayant trait à des événements politiques identifiés (discours ; lettres ; traités ; poèmes ; sermons ; chroniques) ; soit des textes consacrés de manière générale au bon et au mauvais gouvernement. Il rassemble également des textes gouvernementaux (proclamations, ordonnances) et des textes adressés au roi par ses sujets (cahiers de doléances ; requêtes ; lettres de rémission).

Pour le moment, la bibliothèque interne de PALM regroupe des textes d’origine anglaise (en anglais, français et latin) et d’origine française (en français et en latin). Mais nous souhaitons, à l’avenir, ajouter des textes en provenance d’autres pays européens.

Réentrainement/personnalisation: Il est possible que l’utilisateur constitue un corpus de travail personnalisé selon ces besoins et l’annoté semi-automatiquement. Cependant, il n’est pas encore envisageable de réentrainer le taggeur. En effet, notre démarche consiste à mettre en place une annotation automatique via plusieurs outils afin de rendre notre analyse fiable et efficace quel que soit le corpus de travail de l’utilisateur.

Licence: La licence CeCILL (http://www.cecill.info/)

Export/import - format(s): Vers plusieurs logiciels de textométrie comme Lexico3, Le Trameur, TXM et Hyperbase et vers des formats standard comme le .TXT et le XML-TEI.

Dernière version/date: Les dernières modifications de la version actuelle date de Juin 2017 et il y a une nouvelle version en cours de développement qui verra le jour cette année.

Site web: http://palm.huma-num.fr/

Documentation (manuel): En Français ( http://palm.huma-num.fr/PALM/faces/images/ManuelPALMFrMai2014.pdf ) et en Anglais ( <http://palm.huma-num.fr/PALM/faces/images/PALMManualEngMay2014.pdf> )

Contacts/responsables:

[aude.mairey@gmail.com](mailto:aude.mairey@gmail.com) & aouini.mourad@gmail.com

### Pie (tagueur) + Pyrrha (interface de correction et de préannotation)

Accessibilité

- Téléchargement + Installation pour Pie ou Pyrrha

- Interface web pour Pyrrha avec possibilité d’exploiter un webservice Pie pour prétaguer :

http://dev.chartes.psl.eu/ppa/

Interface

- Graphique pour Pyrrha (hors installation sur ordinateur personnel si souhaité)

- CLI (Command Line Interface) pour Pie ou API Web

Systèmes

- Linux et Mac pour Pie/Pyrrha en installation locale

- Tous pour Pyrrha installé sur un serveur web

Langage d’écriture du programme

Python

Langue(s)

- Disponible en ancien français, français classique, latin fin décembre, dépendant des jeux

d’entraînement et des fichiers de configuration.

Jeux d’étiquettes

Dépend des jeux d’entraînement / Pour l’AF : Tobler-Lommatzsch – POS +Morph = Cattex 2009,

pour Latin : LASLA

Cf. https://github.com/hipster-philology/pyrrha/tree/dev/app/configurations/langs

Corpus d’entraînement

*Ancien Français*

Geste 100 000 mots ; J.B. Camps, Lucence Ing, Alice Cochet, Elena Albarran (et des M2) ;

Corpus Juris Civilis env. 100 000 mots ; Frédéric Duval, Lucence Ing ;

Chrestien 200 000 mots, fournis par Pierre Kunstmann, et alignés avec nos référentiels (JBC, LI)

Wauchier 30 000mots ; Ariane Pinche ;

Lancelot 15 000 mots ; Lucence Ing.

*Occitan*

Montferrand env. 30k mots ; Comptes des consuls de Montferrand, éd. Lodge ; JBC, Gilles

Couffignal (Paris-Sorbonne), Marine Mazars ;

Flamenca env. 20k mots, éd. P. Meyer, fournis par Olga Scrivner.

*Latin*

Cf. http://web.philo.ulg.ac.be/lasla/textes-latins-traites/

Réentraînement/personnalisation (ou pas)

- Possible

Licence

- MIT pour Pyrrha

- Inconnue pour pie

Export/import – format(s)

- Plein texte vers TSV pour Pie

- Plein Texte ou TSV vers TSV (et bientôt TEI) pour Pyrrha

Dernière version/date

- Pie comme Pyrrha : Décembre 2018 (encore en développement actif, nouvelles fonctionnalités

très souvent)

Site web

- https://github.com/hipster-philology/pyrrha

- https://github.com/emanjavacas/pie

Documentation (manuel)

- Pie : Non

- Pyrrha : Pas pour l’instant

Contacts/responsables

- Pyrrha : Thibault Clérice , Julien Pilla

- Pie : Enrique Manjavacas, Mike Kestemont

- Modèles préentraînés : JB Camps ou Thibault Clérice

On obtient jusqu’à 90% de réussite sur les textes en ancien français avec le modèle lemmes, POS et morphosyntaxe, voir les résultats ci-dessous :

Français moderne littéraire

- POS : 0.9769

- Lemma : 0.9499

Ancien Français (au 31 octobre) avec Pie

Corpus : Wauchier + Geste (~130.000 mots) uniquement

- POS: 0.9335

- Lemma : 0.9003

- Morph : 0.8468

*Note* : Un problème de configuration important sur les lemma doit être corrigé. Les résultats

devraient augmenter suite à cette correction.