Cosme2 (Consortium Sources Médiévales 2)
Groupe de travail « Lemmes » - Atelier 3

Paris - IRHT - Salle Jeanne Vielliard - 40 avenue d'Iéna (métro Iéna)
10 décembre 2018 - 10h-18h

Compte-rendu

Présents : Mathieu Beaud, Sébastien Brisbois, Pierre Brochard, Hélène Caillaud, Łukasz Gagala, Jean-Philippe Genet, Eliana Magnani, Yves Ouvrard, Nathalie Picque, Ariane Pinche, Coraline Rey, Sergio Torres, Laura Vangone, Philippe Verkerk

Excusés : Bruno Bon, Olivier Canteaut, Sarah Casano-Skaghammar, Thibault Clérice, William Diakité, Simon Gabay, Tim Geehlaar, Laura Gili-Thébaudeau, Estelle Ingrand-Varenne, Dominique Longrée, Aude Mairey, Aurore Menudier, Krzysztof Nowak, Nicolas Perreaux, Evgeniya Shelina

Compte-rendu par Eliana Magnani

Les documents du Groupe de Travail Lemmes sont déposés ici : https://goo.gl/wZbZSD
Vous pouvez transmettre pour dépôt et partage les résumés et/ou les présentations des communications, des articles, des corpus...

1. Diffusion

Fiches sur les outils et les paramètres de lemmatisation (diffusion sur la plateforme Ménestrel)

Ces fiches visent à donner un aperçu synthétique de l'existant, les dénominateurs communs et les spécificités de chaque outil.
Outils concernés (présentés dans les différents ateliers précédents) :
> prêt : **Collatinus, Pandora - Pyrrha**
> en attente : CompHistSem, Hydra, LASLA, Omnia, Palm.
> à contacter : LemLat 3.0 (Marco Passarotti - Projet européen LiLa (Linking Latin - relier toutes les ressources existantes pour le latin)
- **Liste d'entrées** (inspirée du consortium CORLI + des ajouts discutés en atelier)

- accessibilité : téléchargement ou plateforme web
- interface
- systèmes (Windows, Linux, MacOS X…)
- tagueur (ou pas)
- langue(s)
- jeux d'étiquettes
- corpus d'entrainement
- réentrainement/personnalisation (ou pas)
- licence
- export/import - format(s)
- dernière version/date
- site web
- documentation (manuel)
- contacts/responsables
- autres annotations

Ces fiches seront précédées d'une « ***Introduction à la lemmatisation pour médiévistes*** », à rédiger de forme collaborative, en pensant au public de néophytes.

- Qu'est-ce la lemmatisation ?
- Pour quoi faire ?
  - avec exemples concrets d'utilisation pour la philologie et/ou l'histoire :
    - Ariane Pinche - réunir les variantes orthographiques d'un terme à partir du lemme ;
    - Jean-Philippe Genet - recherche sur le mot « gouvernement » ;
    - Łukasz Gagala - les noms propres ;
    - Yves Ouvrad, Philippe Verkerk - persistance de la prosodie antique à l'époque médiévale
    - …
- Quels outils ?
  - réunir les outils par « grandes familles »
  - 2 principes des tagueurs
    - sur la base d'un lexique, ou d'un entrainement ou les deux
    - « réseau de neurones » (*deep learning*) - ni lexique, ni règles, mais représentation sémantique
- Par où commencer ?
  - étape par étape depuis le texte numérisé (avec note sur état critique de l'édition utilisée)
  - acquisition et nettoyage des textes
  - tokenisation
  - lemmatisation et l'analyse sémantique (Nicolas Perreaux)
  - …
- Glossaire des termes techniques (Yves Ouvrard)
- Bibliographie sélective

Des volontaires pour participer à la rédaction de ce petit guide ?
Eliana Magnani propose de réunir et mettre en forme les contributions

- Commencer une liste de corpus lemmatisés en libre accès

- CBMA pour TXM (27 094 documents, essentiellement des actes diplomatiques en latin, répartis du VI^e^ au XIV^e^ siècle, soit plus de 6,1 millions de mots - lemmatisation automatisée avec les paramètres Omnia - http://www.cbma-project.eu/bdds2/la-base-sous-txm.html)
- Corpus corporum - lemmatisation automatisée - http://mlat.uzh.ch/MLS/index.php?lang=0
- CompHistSem - http://www.comphistsem.org/texts.html
- …

2. Formation sous la forme d'une « journée découverte » des outils de lemmatisation

- **date : lundi 17 juin 2019**
- matinée : conférences - présentations générales de la lemmatisation et ses usages
- après-midi : formation à l'utilisation pratique des outils (Collatinus, Omnia, Palm, Pyrrha)
- **toutes les idées et propositions d'intervention sont les bienvenues**

3. Outils de lemmatisation : développements, tests, comparaisons

Présentations :

- les développements de Collatinus pour l'analyse syntaxique, le latin médiéval et les fonctionnalités particulières du tagueur-LASLA (Y. Ouvrard, Ph. Verkerk) - https://outils.biblissima.fr/fr/collatinus/
    - ajout des variantes graphiques pour le latin médiéval ; augmentation du lexique (Du Cange, Dictionnaires topographiques…) ; fichiers .dif pour ajouter un lexique spécifique
    - le tagueur pour le LASLA : philologues qui visent la lemmatisation parfaite ; pour les nouvelles formes, double lemmatisation par Collatinus et la liste du LASLA, comparaison, puis choix de la forme la plus précise.
    - PRAELECTOR : aide à la lecture ; liens syntaxiques possibles entre deux mots ; objectif : exploration des combinaisons possibles pour trouver le meilleur arbre.

- la constitution du corpus épigraphique bourguignon plurilinguistique et les premiers tests de lemmatisation - CBMA-CIFM (P. Brochard, E. Ingrand-Varenne, E. Magnani, A. Menudier, N. Perreaux) - https://journals.openedition.org/cem/15591
    - https://gitlab.huma-num.fr/lamop/cbma-epigraphie/commits/master

- le lemmatiseur-tagueur Hydra pour l'allemand (L. Gagala) - https://github.com/Lukasz-G/Hydra
    - réseau de neurones ;
    - *lemma-guesser* (construction lettre par lettre des lemmes) ;
    - nombreuses langues : moyen haut allemand, moyen bas allemand, ancien frison, latin…

- projet de corpus épigraphique en haut allemand provenant de l'est de la France - Bas-Rhin, Haut-Rhin et l'est de la Moselle (S. Brisbois) (par visio-conférence)
    - région pas encore pris en compte par les différents corpus épigraphiques (CIFM, Epigraphica Europea, Deutsche Inschriften Online…)

Bien qu'elle n'ait pas encore porté des fruits, on continue sur l'idée de réaliser des **essais comparatifs** entre les différents outils et paramètres de lemmatisation

- chaque projet concerné fabrique un **échantillon de son corpus déjà lemmatisé,** avec la **description de ses données** (jeux d'étiquettes, formats des données...).
- proposer des échantillons de textes qui n'ont **pas été utilisés pour des entrainements** (certains sont déjà déposés ici : http://bit.ly/2sz6Xwh)

- cette phase préliminaire vise à préparer un corpus lemmatisé et vérifié pour évaluer sérieusement les différents outils
- différentes langues - travailler par sous-groupes mais continuer à mener les discussions ensemble.
    - sous-groupe ancien français : Simon Gabay (coordination), Mourad Aouini (PALM), Thibault Clérice/Jean-Baptiste Camps (Pandora- Pyrrha), …
        - *demandent 2 mois de vacations pour établir un corpus raisonné du français (discussions entre Simon Gabay et Jean-Baptiste Camps)*

- sous-groupe latin : Collatinus, OMNIA, PALM… (je me permets de suggérer : Philippe Verkerk (Collatinus), Bruno Bon (OMNIA), Jean-Philippe Genet (PALM)...
- sous-groupe anglais : PALM (Aude Mairey, Chris Fletcher)