

**Cosme<sup>2</sup> (Consortium Sources Médiévales 2)**  
**Groupe de travail « Lemmes » - Atelier 2**

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

**Présents :** Renaud Alexandre, Mourad Aouini, Bruno Bon, Olivier Canteaut, Thibault Clérice, William Diakité, Margherita Fantoli, Simon Gabay, Christopher Geekie, Jean-Philippe Genet, Dominique Longrée, Eliana Magnani, Aude Mairey, Krzysztof Nowak, Yves Ouvrard, Nicolas Perreaux, Coraline Rey, Michał Rzepiela, Sergio Torres, Philippe Verkerk

**Excusés :** Paul Bertrand, Pierre Brochard, Sarah Casano-Skaghammar, Chris Fletcher, Laura Gili, Isabelle Guyot-Bachy, Estelle Ingrand-Varenne, 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.

Communications sur les travaux, dont aussi sur les entités nommées, par Mourad Aouini et Sergio Torres.

Présentation du projet « Opera latina » du LASLA par Dominique Longrée et Margherita Fantoli (Université de Liège).

2.

Vives discussions sur **l'évaluation des outils de lemmatisation :**

- constitution d'un corpus-test ou pas ? Un corpus différent des corpus d'apprentissage ? Pré-lemmatisé et corrigé, donc fiable pour l'évaluation ? Comparer les « faux-positifs » ?
- comment traiter les différents jeux d'étiquettes ? Re-étiquetage ou réduction au plus simple ?
- formats d'échange : TEI, .csv ?
- un corpus qui doit demeurer public.
- pour l'ancien français, utilisation possible de la BFM (<http://bfm.ens-lyon.fr/>)
- Treebank

P.Verkerk : Consensus autour de la nécessité d'un corpus lemmatisé (complètement, jusqu'au produit final) et vérifié (éventuellement par plusieurs personnes différentes), si on veut pouvoir évaluer sérieusement les différents outils.

Les jeux d'étiquettes et les formats des données ont été évoqués : chacun doit décrire ses données, pour que l'on puisse se mettre d'accord sur un format commun...

- Les 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), ...

>> sous-groupe latin : Collatinus, OMNIA, PALM...

- Différentes langues dans un même texte : penser à l'identification de type de texte, à la manière des détecteurs de langues existants.
- D'une manière générale, **pour commencer les tests**, il a été décidé que chaque projet concerné fabrique un échantillon de son corpus déjà lemmatisé pour qu'il soit testé sur les autres outils.
- Proposer aussi des échantillons de textes qui n'ont pas été utilisés pour des entraînements (certains sont déjà déposés ici : <http://bit.ly/2sz6Xwh>)
- les présentations et discussions des tests seront l'objet d'une demi-journée de travail lors de l'atelier 3 (novembre ou décembre 2018)

3.

### Actions de diffusion

Plusieurs formats et supports possibles et complémentaires : fiches, tableaux, livrets, tutoriel filmé...

- Exemple des « fiches d'outils » du Consortium CORLI : <http://explorationdecorpus.corpusecrits.huma-num.fr/txm/>
- Exemple du « guide méthodologique pour l'édition numérique de correspondances » du Consortium Cahier : [https://f-origin.hypotheses.org/wp-content/blogs.dir/1993/files/2018/03/Correspondance\\_CAHIER.pdf](https://f-origin.hypotheses.org/wp-content/blogs.dir/1993/files/2018/03/Correspondance_CAHIER.pdf)
- Tutoriel sur Youtube, accompagné la documentation écrite avec les lignes de commande.
- Elaboration collective d'un guide d'initiation à la lemmatisation en tant qu'opération préalable à l'exploitation statistique des corpus, du plus élémentaire au plus complexe (?)
- Opportunité d'élaborer un **tableau comparatif de la structure des outils et paramètres de lemmatisation**. Les éléments à renseigner, encore à compléter et à affiner, sont :
  - accessibilité : téléchargement ou plateforme web
  - interface
  - systèmes (Windows, Linux, MacOS X...)
  - tagueur (ou pas)
  - langue(s)
  - jeux d'étiquettes
  - corpus d'entraînement
  - réentraînement/personnalisation (ou pas)
  - licence
  - export/import - format(s)
  - dernière version/date
  - site web
  - documentation (manuel)
  - contacts/responsables

- Fiches/liste sur les **corpus lemmatisés librement disponibles** en ligne (à réfléchir)

4.

### Formation

- Une journée d'initiation pratique à la destination des étudiants et chercheurs intéressés - juin 2019
- Ensuite évaluer l'opportunité d'organiser une formation ou école d'été sur plusieurs jours.

5.

**Prochains ateliers**

- 3<sup>e</sup> atelier : novembre/décembre 2018 : évaluation des outils / actions de diffusion
- 4<sup>e</sup> atelier : juin 2019 : initiation pratique aux outils de lemmatisation