

**Cosme<sup>2</sup> (Consortium Sources Médiévales 2)**  
**Groupe de travail « Lemmes » - Atelier 1**

**Paris - IRHT - Salle Jeanne Vielliard - 40 avenue d'Iéna (métro Iéna)**  
**6 novembre 2017 - 10h-18h**

**Participants :** Renaud Alexandre, Mourad Aouini, Pierre Brochard, Jean-Baptiste Camps, Chris Fletcher, Simon Gabay, Łukasz Gagala, Tim Geelhaar, Jean-Philippe Genet, Marlène Helias-Baron, Estelle Ingrand-Varenne, Fabrice Jecic, Anne-Françoise Leurquin, Eliana Magnani, Aude Mairey, Yves Ouvrad, Nicolas Perreaux, Coraline Rey, Evgeniya Shelina, Philippe Verkerk.

**Excusés :** Paul Bertrand, François Bougard, Bruno Bon, Olivier Canteaut, Pierre Chastang.

Compte-rendu par Eliana Magnani

Nous remercions l'équipe administrative de l'IRHT pour l'accueil et l'organisation matérielle de la journée.

*\* Un espace de partage de fichiers a été créé dans sharedocs de Huma-Num pour Cosme2 dont le groupe « Lemmes ». On y trouve les documents relatifs à cet atelier (programme, liste des participants, articles de référence, diaporamas...) :*

<https://goo.gl/wZbZSD>

1.

Présentation et discussion de différents « lemmatiseurs » : Collatinus (Y. Ouvrad, Ph. Verkerk), Pandora (J.-B. Camps), CompHistSem (T. Geelhaar), OMNIA (R. Alexandre), PALM (M. Aouini, C. Fletcher, A. Mairey)<sup>1</sup>.

*Voir les diaporamas et/ou articles de référence en suivant le lien ci-dessus.*

D'une manière générale, quels que soient leurs objectifs ou leurs structures (lexique + entraînement ; réseau de neurones), tous les outils (tagueurs et/ou paramètres) présentés sont estimés performants à environ 90% ( $\pm 5$ ). Ce sont donc les 5-15% d'erreurs qui demandent réflexion. En général, les applications achoppent là où se trouvent les problèmes historiques (S. Torres).

La **reconnaissance des noms propres** (personnes et lieux) figure parmi les erreurs récurrents d'étiquetage et donc le groupe poursuivra ses travaux autour de ce point.

---

<sup>1</sup> <http://outils.biblissima.fr/fr/collatinus/>  
<http://www.comphistsem.org/home.html>  
<http://www.glossaria.eu/treetagger/>  
<http://palm.huma-num.fr/PALM/>  
<https://github.com/hipster-philology/pandora>

L'autre question soulevée est l'absence d'évaluation systématique et comparative des tagueurs (N. Perreaux). Il a été décidé de créer un **corpus de référence**, structuré, avec un échantillon de différents types de texte (en latin, français et anglais médiévaux) pour tester les outils avec le même corpus afin d'obtenir une **évaluation** raisonnée.

L'objectif serait de concevoir, à partir de ces expériences, un « méta-tagueur » (T. Geelhaar) combinant les avantages des solutions proposées par chaque outil.

Pour ce faire les membres du groupe partageront les différents paramètres et ressources déjà existants.

**Un deuxième atelier** (coût d'environ 2000 €) sera organisé en juin 2018 autour de ces deux points, le matin, la présentation des recherches sur les **entités nommées** dans le cadre de doctorats en cours par Mourad Aouini et Sergio Torres ; l'après-midi, travail sur le **corpus-test et l'évaluation des outils de lemmatisation**.

Cet atelier intéressera sans doute d'autres groupes de Cosme2 (notamment « noms de lieux » et « prosopographie et identification des personnes »). Les collègues de la « Société française d'onomastique » (<https://www.sfo-onomastique.fr/>) (ind. F. Jejcic), ainsi que du réseau Heloise (<https://heloise.hypotheses.org/>) (ind. J.-Ph. Genet), pourraient également collaborer.

2.

Le groupe s'est également mis d'accord sur la mise en place d'autres **actions de diffusion et d'information scientifique**.

En s'inspirant du Consortium CORLI - Corpus, Langues, Interactions<sup>2</sup>, il tâchera de créer des **fiches descriptives** des outils de lemmatisation (à commencer par ceux présentés lors de l'atelier). Le profil de la fiche-type sera discuté par email à partir des propositions des membres du groupe.

Parallèlement aux outils, seront également recensés et décrits les **corpus historiques lemmatisés** librement disponibles (p. ex. Corpus Corporum (Zurich) - <http://mlat.uzh.ch/MLS/index.php?lang=0> ; CompHistSem - Computational Historical Semantics – for Analysing Latin Texts Semantically (Francfort) - <http://www.comphistsem.org/home.html> ; CBMA - (<http://www.cbma-project.eu/bdds2/la-base-sous-txm.html> ; etc...)

Ces fiches ainsi que les produits des travaux du groupe seront mis en ligne dans le site de **Menestrel** (<http://www.menestrel.fr/>).

### 3. Financements

**Les projets communs** développés entre différentes équipes partenaires de Cosme2.

Les équipes des CBMA (Lamop-Paris - E. Magnani) et du CIFM (CESCM-Poitiers - E. Ingrand-Varenne) sont en train de préparer la lemmatisation inédite d'inscriptions épigraphiques, à commencer par un corpus témoin, les 3 volumes bourguignons (VIII<sup>e</sup>-au XIII<sup>e</sup> s.) du CIFM n° 19, 20, 21 (déjà extraits et pré-traités par P. Brochard, Lamop) auxquels seront ajoutées les inscriptions des XIV<sup>e</sup> et XV<sup>e</sup> s. en latin et en langue vernaculaire (recensées à Poitiers, mais encore inédites). Pour aider à réaliser ce travail, il est demandé pour 2018, 3 mois de vacations (environ **10000 €**) (pour la relecture et

---

<sup>2</sup> <https://corli.huma-num.fr/>, voir en particulier « l'exploration de corpus : outils et pratiques » : <http://explorationdecorpus.corpusecrits.huma-num.fr/> avec le recensement et la fiche descriptive et les fiches « d'usage », avec la possibilité de « noter » l'outil avec des étoiles, par exemple, la fiche-page de TXM : <http://explorationdecorpus.corpusecrits.huma-num.fr/txm/>

l'uniformisation des textes, l'incorporation des *corrigenda* et *addenda*, ainsi que des textes inédits).

*Après l'atelier, d'autres demandes de financement pour 2018 ont été formulées.*

« Le projet TITULUS (CESCM - Poitiers), sollicite à nouveau l'aide de Cosme2 pour 2018 à hauteur de **7000 euros** : 2000 euros pour le recrutement d'un stagiaire pendant 3-4 mois pour le développement de l'édition électronique des inscriptions médiévales ; 5000 euros pour prolonger le recrutement d'un IE afin de terminer l'encodage du volume sur les inscriptions carolingiennes » (E. Ingrand-Varenne).

1 ou 2 ateliers en 2018 : **2000 ou 4000 €**

**TOTAL (provisoire) des crédits sollicités : 19000 ou 21000 €**

*En vue du rapport qui doit être remis au TGIR Huma-Num sur les **activités 2017**, nous profitons également de ce compte-rendu pour indiquer **les réalisations** des projets financés rattachés au groupe « Lemmes ».*

- « PALM a recruté une vacataire pour aider M. Aouini à formaliser les règles linguistiques concernant le moyen-français afin d'enrichir le module Nooj qui vient compléter la palette d'outils de PALM (les tagueurs et les dictionnaires) » (**6000 €**) (A. Mairey).
- « Le projet TITULUS, épigraphie médiévale numérique, a bénéficié d'un soutien de **7000 €** de COSME<sup>2</sup> en 2017, qui ont été utilisés pour deux actions. La première a permis le développement d'une interface de recherche dans la base XML eXist-db du projet, déclinée en recherche simple et recherche avancée (2500 €) <http://titulus.huma-num.fr/recherche/recherche.php> ; la seconde – en cours – a permis le recrutement d'un ingénieur d'études pour l'encodage en XML-TEI d'une partie d'un nouveau volume du *Corpus des inscriptions de la France médiévale*, portant sur les textes funéraires carolingiens (4500 €) » (E. Ingrand-Varenne).
- « Atelier 1 - 6 novembre 2018 : ± **2000 €** ».

#### 4. Le fonctionnement du groupe :

- Echanges par email (fiches, corpus-test...) ;
- Un ou deux ateliers de travail par an ;
- À envisager, à moyen terme, l'opportunité d'organiser une formation à l'utilisation des outils de lemmatisation.