Cluster 7 Biblissima+ projets en cours

Jean-Baptiste Camps, Matthias Gille Levenson, Sébastien Hamel, Lucence Ing, Alice Leflaëc, Laurence Mellerin, Luc Massip, Emmanuelle Morlock, Jules Nuguet, Philippe Pons

CLUSTER 7 – Interopérabilité et analyse des textes

Biblissima+

Jean-Baptiste Camps, Matthias Gille Levenson, Sébastien Hamel, Lucence Ing,
Alice Leflaëc, Luc Massip, Laurence Mellerin, Emmanuelle Morlock, Jules Nuguet, Philippe Pons

Premiers chantiers 2024

Le cluster 7 oriente ses travaux sur **l'interopérabilité, l'alignement et l'analyse des textes**. En plus du travail en cours sur DTS, cette année, l'équipe s'est donné comme objectif premier la constitution d'une base lexicale commune dans data.biblissima, dont la structure modulable permet le traitement **de langues et d'époques multiples**. Des **alignements de référentiels** sont en cours.
Un autre chantier concerne l'**alignement automatique de textes**. L'équipe vise à développer un outil permettant l'alignement de témoins de **différentes langues**, voire leur collation, à l'aide de modèles de langue. Elle s'intéresse, dans un premier temps, aux traditions médiévales.

Alignement multilingue

Approche originelle : alignement et collation (`collatex`) automatiques basés sur des *n-grams*, des **chaînes de caractères identiques**

Ao Ez	Li -	contes -	dit -	que -	qant Quant	li le	chevaliers chevalier	deseritez desherité	oï ouyt

Exemple simple de collation sous format tabulaire

→ nécessité de lemmatiser les textes, du fait de la forte variation graphique des langues vernaculaires médiévales

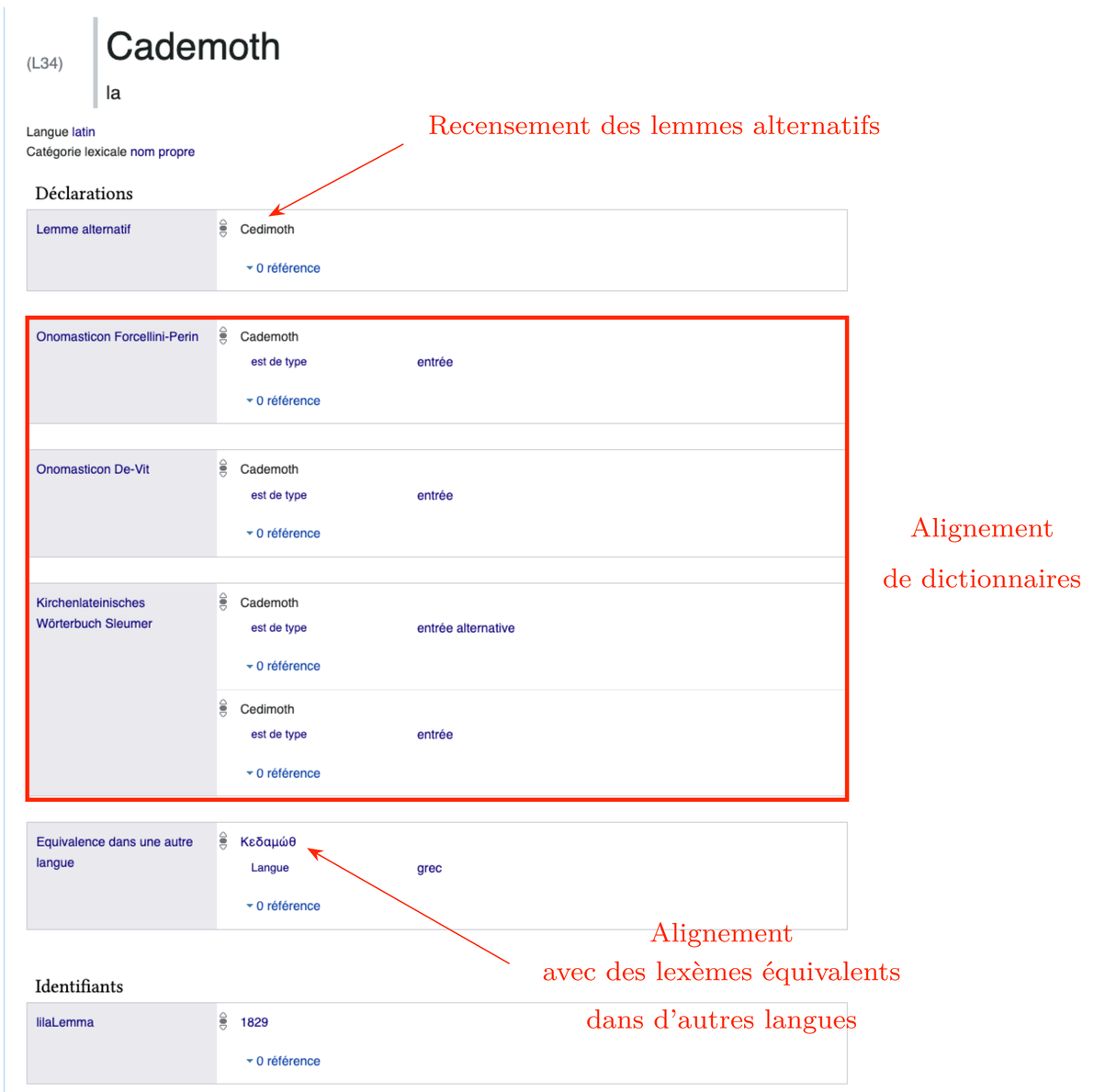
Nouvelle approche

— travail sur des **traditions multilingues** (*De Regimine Principum* en castillan méd. et en latin ; *Lancelot en prose* en français, castillan et italien médiévaux)
— une **approche sémantique**, qui utilise des modèles de langue multilingues (`bertalign` pour l'alignement, `awesome-align` pour la collation [à venir]), après un travail de tokénisation (entraînement de modèles pour la classification de tokens qui permettent de tokéniser les textes en propositions)

éd. Micha	éd. Sommer	*Lanzarote*	*Lancellotto*
et fet le signe de la **crois en mi son vis** puis enbrace l escu	et met le signe de la **vraie crois en mi son front & en mi son pis** puis enbrace son escu	e fiço la señal dela **cruz sobre si** e enbraço el escudo	e si fa el segno della **santa croce nel miluogo del suo viso** poscia imbraccia lo scudo
et **broche** le cheval des esperons	et **fiert** le cheual des esperons	e **firio** al cauallo de las espuelas	e **broca** il cavallo degli sproni

Extrait de la table d'alignement de certains des témoins

Création d'une base lexicographique pour les langues anciennes et le français médiéval



Lemmatisation du français médiéval

centre Jean Mabillon — École nationale des chartes

À l'École des chartes avait été entraîné un **modèle d'étiquetage de l'ancien français**. Un des enjeux actuels est de convertir les référentiels utilisés par ce premier modèle afin de permettre l'intéropérabilité des données en diachronie, ce qui nécessite notamment l'**alignement des lemmes du français médiéval**. Un autre enjeu est de poursuivre l'étiquetage de textes sur la période des XIVe-XVe siècles, afin de produire un modèle d'annotation permettant de couvrir l'ensemble des états de la langue médiévale.

Deux corpus bibliques et patristiques

HISOMA — Histoire et sources des mondes antiques

BiBlIndex rassemble les citations bibliques des textes chrétiens de l'Antiquité et du Moyen Âge, **Jerihna** les interprétations des noms hébreux issues de l'oeuvre de Jérôme de Stridon. Les alignements se font entre les référentiels de la Vulgate, de la Septante et de dictionnaires existants. Ils portent en premier lieu sur les noms propres des Bibles et leurs reprises dans les œuvres de Philon d'Alexandrie, Origène, Eusèbe de Césarée et Jérôme.

BiBLindex — JERIHNA

Un grand corpus de latin médiéval (700-1300)

cnrs — IRHT Institut de recherche et d'histoire des textes — Académie des Inscriptions et Belles-Lettres, Institut de France

La section de lexicographie de l'IRHT apporte un corpus lexical lemmatisé de 50 millions de mots de latin médiéval (800-1200), créé avec le soutien de l'ANR Velum. Biblissima+ permettra d'y ajouter 50 millions de mots supplémentaires (700 et 1300), pour l'étude de la diachronie à l'échelle Européenne et rendre compte du latin mérovingien et scolastique. Il sera interrogeable sur NoSketch-Engine et CQP-Web, consultable sous TEI-Publisher et téléchargeable au format XML.

Deux lemmatiseurs (latin et grec)

Collatinus est un lemmatiseur et analyseur morphologique de textes latins (Philippe Verkerk).

Eulexis est un lemmatiseur de textes en grec ancien (Yves Ouvrard, Philippe Verkerk).

Tous deux font partie de la Boîte à outils Biblissima (Baobab).