Compte-rendu de la réunion PaLaFra
15 et 16 février 2016, ENS de Lyon

Participants : A. Carlier, S. Diwersy, L. Döhling, R. Eufe, C. Guillot, S. Heiden, N. Kanaoka, A. Lavrentiev, S. Ortner, E. Reichle, J. Vangaever, C. Wolff.

1. Corpus PaLaFra

1.1. Partie française

Une première sélection de textes a été faite, sur la base des critères suivants :

- Date de composition (textes les plus anciens)
- Domaines et genres discursifs : on privilégie les genres comparables à ceux des textes latins (vies de saints, quelques chartes, textes non littéraires)
- Textes en prose

Le corpus sélectionné comporte environ 1 million de mots. Il va du 9ème au début du 15ème s. (pour les vies de saints, on a des textes presque à chaque siècle). Il comporte quelques textes littéraires qui avaient déjà été étiquetés dans la BFM (*Chanson de Roland*, *Tristan* de Béroul, *Yvain* de Chrétien de Troyes, etc.).

Le corpus français ne comporte pas de lettres, contrairement au corpus latin. Il faudrait voir si on peut en ajouter.

1.2. Partie latine

Le corpus est issu des MGH et est constitué majoritairement de vies de saints (une trentaine), de chartes (environ 40), de quelques textes historiques (livre 6 de l'*Histoire* de Grégoire de Tours, *Chronique* de Frédégaire), d'une centaine de lettres et de quelques textes juridiques (leges). Il y a environ 200 textes au total. Ils sont exportables au format XML (ou HTML) avec, peut-être, une possibilité de récupérer le balisage du discours direct (L. Döhling). Les métadonnées des textes sont préparées par S. Ortner.

A. Carlier suggère d'inclure les continuations de Frédégaire, qui sont particulièrement intéressantes pour la latinité tardive, en les séparant bien du reste du texte (*Chronique des temps mérovingiens (Livre IV et Continuations)*, texte latin selon l'édition de J.M. Wallace-Hadrill, traduction, introduction et notes par O. Devillers et J. Meyers).

1.3. Corpus parallèle

Trois textes hagiographiques en cours de traitement : *Vie de saint Benoît, Vie de saint Lambert, Vie de saint Eustache*.

On décide de poursuivre les recherches pour voir si on peut y ajouter d'autres textes (auprès notamment de C. Pignatelli, du groupe de Gand et de C. Buridant).

On aborde la question de l'alignement et de la vérification des étiquettes sur ces textes. L. Döhling fait une démonstration de l'outil d'alignement Intertext. Un test a été réalisé à partir des versions Lasla et BFM de la *Vie de saint Benoît*.
L'équipe BFM doit envoyer la *Vie de saint Eustache* à R. Eufe pour qu'il regarde la partie latine du texte.
Il faudra trouver une solution pour la vérification manuelle de l'annotation des textes latins (trouver peut-être des étudiants pour faire le travail).

1.4. Diffusion du corpus

Les textes français sélectionnés sont (ou seront) intégrés à la BFM, dont les textes sont diffusés sous la licence libre CC BY-NC-SA. Les textes latins sélectionnés sont tous inclus dans la collection des MGH. Ils ont la même licence libre que les textes français de la BFM. Les MGH envisagent de créer un portail de téléchargement (Open Monumenta) qui comprendrait les textes latins avec les annotations PaLaFra. On peut envisager une triple diffusion pour le projet : dans les unités de recherche à Regensburg et Lyon et sur le site des MGH.
Les textes latins sont d'ores et déjà disponibles sur le site des MGH aux formats html et pdf. Sur le nouveau portail, ils le seront au format XML.

1.5. Métadonnées textuelles/Descripteurs

Les vies de saints latines ont été décrites à l'aide de la fiche PaLaFra. Une base de données doit héberger ces informations à Regensburg. Un tableau partagé permettra d'importer les métadonnées des textes latins dans le portail TXM de Lyon.
La fiche actuelle convient bien. Une indication de lieu/région de rédaction et de copie remplace l'information sur le dialecte pour la partie française. Il faudrait ajouter à la fiche le fait qu'il existe une traduction dans une langue moderne pour les textes de latin tardif. C'est une information importante pour l'exploitation et la diffusion du corpus (quelle que soit la langue de traduction), les textes étant souvent difficiles à comprendre. On ajoute autant de remarques que nécessaire à chaque champ de la fiche.
La question du type d'édition (+ ou - lachmanienne) pour les textes latins est discutée. On renseigne le mieux possible ce champ de la fiche pour les textes latins.

1.6. Plateforme d'analyse TXM

On propose d'intégrer les textes du corpus PaLaFra dans le Portail TXM de Lyon, mais on envisage une installation d'un portail en Allemagne. M. Selig doit se renseigner auprès de l'Académie de Bavière et du Leibniz-Rechenzentrum à Munich, L. Döhling propose d'installer un portail à Regensburg. S. Heiden démarche par ailleurs des collègues pour l'hébergement de portails TXM en Allemagne :
- (à l'occasion de discussions lors de la conférence TEI MM à Lyon en octobre 2015) Franz Fischer du centre Cologne Center for eHumanities (CCeH) de l'Université de Cologne a indiqué qu'il était possible d'envisager un hébergement de portail TXM dans cette Université pour un projet de SHS

- un contact est en cours avec le partenaire Université de Würzburg de DARIAH-DE pour que cette infrastructure puisse prendre en charge l'hébergement de portails TXM pour des projets de SHS en Allemagne
Il est suggéré que l'Université de Regensburg demande son affiliation au réseau DARIAH-DE pour pouvoir bénéficier de ses services.
Une traduction de la documentation TXM en anglais serait utile. Il existe une documentation anglaise pour l'installation de la version portail. Le manuel utilisateur (version pour poste) est en cours de traduction, il sera prêt dans un mois (mars 2016).
Une formation TXM sera organisée pour le projet. Les ateliers réguliers sont interrompus en ce moment. Une formation sera prochainement assurée lors de l'école thématique CNRS MISAT (Besançon). On prévoit un atelier lors de la prochaine réunion plénière PaLaFra.

2. Annotation

2.1. Morphosyntaxe

On ne modifie pas les jeux existants qui continuent d'être utilisés (par exemple, on n'ajoute pas dans Cattex la distinction entre verbe conjugué et auxiliaire).
Discussion sur l'intégration des deux langues dans les requêtes et sur l'utilité d'un jeu morphosyntaxique commun. Le jeu commun pourrait être utile pour le corpus parallèle. Il devrait donner aussi une meilleure visibilité au projet et à l'annotation. On décide de créer ce jeu sur la base des choix déjà discutés.
On évoque la question de la standardisation du jeu (voir notamment le projet Universal Dependencies – http://universaldependencies.org, les étiquettes de Leipzig glossing rules https://www.eva.mpg.de/lingua/resources/glossing-rules.php ainsi que le jeu Multext - http://nl.ijs.si/ME/V4). On décide de se renseigner sur les jeux actuels et d'essayer d'être le plus compatible possible.
Discussion sur la distinction entre morphologie et syntaxe : deux étiquettes (M et MS) dans Cattex qui permettent de traiter par exemple les infinitifs substantivés, les participes en emploi adjectival, etc. On discute de l'emploi des cas morphologiques latins dans les textes tardifs. Dans le projet CompHistSem, c'est la fonction syntaxique qui est annotée et on indique « non classique » lorsqu'il y a contradiction entre le cas et la fonction réelle. On décide d'adopter un système comparable dans PaLaFra en ajoutant des remarques chaque fois que nécessaire (statut de langue).
L'équipe de Regensburg/Tübingen envisage une annotation automatique syntaxique du latin à partir du modèle du Perseus Project.

2.2. Lemmes et lexiques

Un lexique est en cours de constitution à Lyon pour la lemmatisation du corpus français. Il est constitué à partir des ressources de la BFM et du Nouveau corpus d'Amsterdam (A. Stein). Il sera mis en ligne sous licence libre dès qu'A. Stein aura donné son accord.
Les ressources lexicales de CompHistSem disponibles sur le E-Humanities-Desktop de Francfort ne sont pas librement accessibles. Le projet PaLaFra vise à publier un corpus de

textes latins, et non des ressources lexicales pour cette langue. Pour le français, on envisage de distribuer à la fois le corpus et le lexique.
On note la création de superlemmes dans CompHistSem grâce à des liens entre lemmes. On se demande si ces superlemmes sont comparables aux étymons du FEW. Cela pourrait être une piste pour établir un lien entre les lemmes français et latins.

2.3. Outils d'annotation

L. Döhling a testé les outils LAPOS et Marmot (« conditional random fields ») sur les ressources linguistiques de Francfort. Ces tests ont montré que les outils basés sur des dictionnaires génèrent un taux d'ambiguïté beaucoup plus élevé mais beaucoup moins de formes inconnues. Il faut sans doute combiner les deux types d'outils. La prochaine étape est d'utiliser le corpus de Francfort (plus d'1 million de mots) comme corpus de test.
Il sera nécessaire d'avoir aussi des outils pour la correction manuelle des lemmes et des étiquettes. Jusqu'ici l'équipe CompHistSem utilisait Scratchit. M. Burghardt va évaluer différents outils. Un module d'annotation/correction des lemmes/étiquettes est en cours de développement dans TXM (version pour poste). Un premier développement permet l'annotation par le biais de concordances de séquences de mots avec des catégories de référentiels externes (comme le référentiel de sémantique historique SiMoGIH : http://www.symogih.org) (projet Bibliothèque Historique de l'Éducation en partenariat avec le laboratoire LARHRA).
N. Kanaoka a l'expérience de la correction des lemmes et des outils d'assistance. Un document fait état de ses suggestions pour une stratégie et des outils d'annotation/correction. Il a été envoyé aux participants à la mailing-list PaLaFra.

2.4. Précisions techniques (L Döhling)

- Formats des textes
  - Corpora should be available in different (XML-)formats, ready for import (LD)
    - TXM
    - TEI
    - Linguistic Annotation Framework
    - Corpus Workbench…
- Tokenisation
  - Latin: we will separate the clitics, i.e. –que (LD)
    - For visualization, these two tokens should appear connected, but with a hint, that these are in fact two tokens in terms of CQP queries for example
- Morphosyntaxe
  - Latin: we will use the CompHistSem tagset
    - Anno-guidelines will be prepared by ER
  - All(?) texts will be annotated with the new PaLaFra tagset too
  - We hope to create these tags automatically by mapping CompHistSem/Cattes09 to PaLaFra
- Lemmatisation
  - Latin: we will adopt the superlemma concept, introduced by CompHistSem
    - Superlemma: abstraction of different orthographic variants

3. Communication interne et externe

3.1. Partage de fichiers (repository)

On évite Dropbox et plus généralement tout ce qui est lié à une entreprise privée.
L'ENS de Lyon a une solution mais elle est peu flexible pour les personnes extérieures (système d'authentification lié aux institutions). Deux solutions sont envisagées :
(1) l'Académie de Bavière propose un système comparable à Dropbox (Sync + Share), un compte artificiel serait créé pour le projet
(2) l'université de Lille propose également un espace de disque partagé
(3) DARIAH.DE offre peut-être ce service (https://de.dariah.eu)

3.2. Site Web du projet PaLaFra

Présentation de L. Döhling du site actuel : Wordpress hébergé à l'U. de Regensburg : https://www-app.uni-regensburg.de/Fakultaeten/SLK/Medieninformatik/PaLaFra/wordpress-dev/
Toutes les pages devront être traduites en allemand, en anglais et en français :

- « News » : fil d'actualités, un blog
- « Project Description » : à partir du projet déposé
- « Texts » : description des textes du corpus
- « Teams » : partenaires du projet et les personnes impliquées
  - photos de groupe de Munich et de Lyon
- L. Döhling rassemblera les informations nécessaires pour chaque groupe
- À droite, adresses de contact, au moins l'université de Regensburg et l'ENS de Lyon et l'adresse mail info@palafra.org
- Une adresse de contact doit être créée
  - info@palafra.org réexpédié aux responsables de tous les partenaires
- Appel à propositions pour un logo.

3.3. Installation de portail TXM en Allemagne pour héberger le corpus latin

- Différentes options seront évaluée par L. Döhling
  - VM @ RZ Regensburg
  - VM @ DARIAH-DE

4. Calendrier

- Réunion restreinte 14-15 juin à Regensburg
- Réunion plénière 11-12 octobre à Lille

5. Planning

Explicitation des dépendances entre objectifs de recherche, corpus et outils

- Établir une liste de questions de recherche, par exemple 10 questions, associées à un chercheur ou à une équipe, dont le traitement s'appuie sur des annotations et sur des outils AVANT JUIN

Textes

- demander aux MGH les formats disponibles des textes latins, notamment le format XML, qui pourrait permettre de baliser les mises en forme éditoriales (discours direct, italiques, etc.) ; diagnostic sur le meilleur format de départ AVANT PAQUES
- collecte des textes des MGH annotés pour transfert à l'ENS de Lyon, si possible en version XML AVANT PAQUES
- transfert de la documentation correspondante
- C. Guillot contacte Sources chrétiennes et les Editions du Cerf pour régler la question des droits sur la *Vie de saint Benoît*
- import des deux corpus (latin/français) dans TXM AVANT OCTOBRE

Métadonnées/descripteurs

- documentation : décrire la façon dont on remplit les fiches pour les textes latins et les pratiques qui diffèrent de celles du manuel_descripteurs de la BFM
- ajouter un champ traduction moderne dans la fiche
- voir si l'on a besoin du champ « Type d'édition » pour créer un sous-corpus et/ou trier les textes
- dépôt sur le site de partage du tableau partagé de métadonnées pour les textes latins

Communication

- demande de M. Selig auprès de https://syncandshare.lrz.de/login pour obtenir un compte projet non personnel; voir aussi https://de.dariah.eu

TXM

- annonce de l'école thématique CNRS MISAT auprès de romanistik.de
- proposer un tutoriel BFM en allemand ?
- préparation de la formation PaLaFra AVANT OCTOBRE

Annotation

- Actualiser le jeu PaLaFra AVANT DEBUT MAI
    - Pbe du passé composé : analyser les composants de manière séparée
    - Distinguer pour les verbes les formes finies vs non finies, ou personnelles vs non personnelles (regrouper le subjonctif avec l'indicatif, à distinguer de l'infinitif et du participe)
    - Ledict/supradictus etc. : voir avec M. Selig la façon de les annoter
    - Ibi/inde : les classer dans les adverbes
    - Ajouter un champ supplémentaire correspondant à non-classical Latin dans CompHistSem

    - Regarder les jeux d'étiquettes normalisés
    - Guidelines en anglais pour CompHistSem (E. Reichle)
- Tables de conversion entre le jeu PaLaFra et les jeux existants AVANT DEBUT MAI
- Annotation pragmatico-discursive
    - Voir avec M. Selig quel outil d'annotation utiliser : Oxygen ou Word/Writer + Odette ?
- Annoter automatiquement les textes latins et français avec leur propre jeu (CompHistSem et Cattex) et avec le jeu PaLaFra AVANT OCTOBRE
- Introduire plus tard les distinctions nécessaires qui pourraient être transférées d'une langue à l'autre
- Corriger une partie des textes manuellement AVANT OCTOBRE