Comité de pilotage CORLI
Compte-rendu
7 octobre 2022

Présents: Flora Badin, Antonio Balvet, Franck Cinato, Céline Dugua, Sarra El Ayari, Carole Etienne, Achille Falaise, Julie Glikman, Mathilde Guernut, Marie-Paule Jacques, Loïc Liégeois, Christophe Parisse, Céline Poudat, Marie-Anne Sallandre, Frédérique Sitri, Amalia Todirascu

Point métadonnées

À ce stade, une coopération avec Cahier est en cours. Cahier a en effet développé un thésaurus disponible ici https://opentheso.huma-num.fr/opentheso/?idt=43
A noter qu'à la base, c'est un thesaurus pour les corpus littéraires donc il est nécessaire d'évaluer sa pertinence, la pertinence de certaines branches (et pas d'autres), les catégories absentes, etc.
But: créer un jeu de métadonnées pour les corpus écrits / articuler corpus écrits et corpus oraux

- niveau minimal sur lequel on peut travailler
- définir niv 0 et niv 00 de MD nécessaire et utile
- rappel: pour corpus oraux: niveau 0 de MD (ni trop contraignant ni trop pauvre) ; à coder dans les corpora avec poss d'éditer et uniformité entre les différentes MD
- une fois les MD de base créées, il sera possible de créer (cf. Christophe) un formulaire de saisie de métadonnée avec l'outil TEI meta
- voir les corpus déjà déposés et voir si conversion possible puis mise en place d'outils d'interrogation

Description de MD terminée pour 2022

Suite à réunion avec Cahier (mai 2022, Nice) → répartition en fonction des types de txt
[→ Sascha Diwersy pour les textes parlementaires;
→ Christophe P, Christophe B et Carole Etienne pour les interactions orales;
→ Agnès TUTIN pour les textes scientifiques;
→ Céline pour la CMC (à construire à partir du projet CoMeRe);
→ Marie Chandelier pour la presse;
→ Marie-Paule Jacques pour les écrits scolaires]
corpus CR université ; géré par un IE - corpus en cours de constitution ; peut être intéressant pour partie "txt administratifs" (corpus avec image ; voir comment gérer)

Alexei → envoi lien vers Open Theso pour tester la sandbox
Céline → doc collaboratif prévu pour mise en commun
Carole Etienne: projet Orfeo (?) → panel de corpus varié permettant de constituer niv0 ; risque de travailler par groupe ; importance de se mettre du point de vue de l'utilisateur
Céline 3 angles

- chaque personne sur genre particulier
- point sur MD existantes
- paramètres de la situation énonciative

JGlikman: BFM: 2 cat : domaine (histq, litt etc) & genre - palier en dessous (roman etc)
pb : chq projet a sa propre def de genre. Recoupement des paramètres intéressant (ex corpus de SMS écrits & oraux)
But: opérationnel même si pas parfait

Point annotation de corpus

3 axes cf slide
1 - projet Palamède: MSH Lorraine ; recrutement IE sur le projet
2 - inception - adapter le logiciel aux pratiques des linguistes + faciliter import/export ; module stanza ; import dans TXM
3- en cours de test - ressource qui sera construite à partir d'inception

Point GUM

GUM: développement d'une ressource d'annotations en classe (resp. Mai Ho-Dac).

Mise en place d'un groupe de travail (master TAL Grenoble ; Mai Ho Dac ; Paris etc) en cours ; dvpt de cette ress d'annotation
Possibilité de travailler avec Lille (2 masters LGO + LTTAC), Strasbourg (2 masters: sciences du langage et Technologie des langues), Paris Sorbonne (cours Karën Fort), éventuellement Orléans (F.Badin).

Idéalement, on aimerait à terme pouvoir mettre en relation des chercheurs ayant un projet d'annotation et qui auraient besoin d'annotateurs (étudiants en master) → mettre tlm en relation à partir de la plateforme

Journée scientifique 2023 - un invité par projet → idées de personnes à inviter?

Point CITATION

Le projet CITATION a mené à une réunion qui a permis de confirmer et mieux connaître l'intérêt des membres de CORLI à ce projet et de faire des critiques constructives. Une prestation a été lancée pour réaliser l'état de l'art et créer des exemples réels de citation et d'extraits. Une demi-journée sera organisée pour faire un retour sur la prestation et planifier la suite du travail.

Journée scientifique CORLI - janvier 2023

On pense organiser en janvier 2023 (5, 9, 10, 11, 12 ou 16 janvier dans l'idéal) une journée scientifique internationale qui nous permettra à la fois d'inviter des experts reconnus, d'avoir un retour sur nos travaux et de diffuser nos avancées.

Pour **annotation**, on pense inviter Amir Zeldes (Georgetown University, Gum project).
Pour **citation,** Nicolas Larrousse
Pour **OFC:**
collègues allemands (IDS Mannheim) ; présentation sur mise en commun de corpus écrits
CNC Corpora https://korpus.cz/
Frédérique Sitri Archivu
Lou Burnard
Demander à Eva Soroli, ambassadrice Clarin, qui pourrait être une bonne personne ressource sur la création d'un corpus de référence pour une langue (format, métadonnées, sources, origines etc.).
Si vous avez des idées, n'hésitez pas à les écrire ici.

ANNOTATION:
Sketchengine (centralise corpus en différentes langues ;) → gestion multicouches ; volet annotation

Bilan annuel CORLI
document collaboratif à relire pour le 15/10
publications à envoyer en lien avec CORLI

ACTU
MA Sallandre - participation groupe de travail (nov21 sept22) ; recensement des corpus existants en LDS en différentes langues pour chaque pays. Responsable de la collecte pour la France ; site ouvert le 06/10/22
→ https://corli.huma-num.fr/interactions-avec-clarin/
La ressource CLARIN Resource Family for Sign Languages est maintenant en ligne:
https://www.clarin.eu/resource-families/sign-language-resources

F Badin financement stage pour le groupe QuECJ: 2x4 mois avec 2 stagiaires