RÉUNION DE TRAVAIL
OPEN FRENCH CORPUS / Métadonnées
3 FEV. 2023

Document commun
Thesaurus 'Typologie'

Agenda

Prochaine réunion envisagée la semaine du 6 mars, un doodle sera envoyé.

Présents: Christophe Benzitoun, Carole Etienne, Iona Galleron, Mathilde Guernut, Fatiha Idmhand, Marie-Paule Jacques, Alexey Lavrentiev, Christophe Parisse, Céline Poudat, Geoffrey Williams.

Documents partagés:
Dossier commun
221125_Doc. collaboratif

Avancement du travail

Du côté de CORLI, mise à jour d'un fichier commun avec description précise des corpus qu'on a puis harmonisation (md, formats)
Fichiers CORLI:
Fichier excel + ppt

Il faudra faire le point sur les corpus Cahier (sur Nakala) et Cocoon.
NB: CAHIER continue sa vie de consortium sous le nom d'ARIANE!

Du côté d'ARIANE/CAHIER: travail sur intégration automatique du thesaurus - ça aide pour
Mynkl → dépots de fichiers, ajout concept
ex: 50 romans, add concepts venant d'Opentheso (att! même concept pour l'ensemble)
"Modif avec CSV" → enrichissement entre données NAKALA & MD des corpus

Actions à venir

- réfléchir à la présentation de l'OFC, et aux corpus qui y seront rattachés
    - problème des dépôts multiples de corpus
    - il faudra un système dynamique, qui sera capable de prendre en compte les modifications éventuelles apportées aux corpus déposés; utilisation des outils de versionnage pour modif/enrichissement des corpus
    - → dans ORTOLANG, versions différentes ; NAKALA idem
    - si une ressource/corpus = autonome ou ressource/corpus = partie d'un autre travail ; → plusieurs sources
    - Les personnes cherchant dans les corpus → confusion qd W sur partie d'un corpus
        - chaque projet démarre en prenant une/des partie.s d'un corpus existant, d'un enregistrement avec des annotations différentes (bref, reprise de données venant d'ailleurs)
        - lien entre différentes données = traçable & visible en TEI
    - NAKALA: pas d'interface utilisateur pour récupérer un ensemble de données dans un seul zip pour l'instant ; pour cela il faut passer par l'API et programmer soit même la récupération (y compris être identifié dans NAKALA)
    - Quelle granularité pour la déclaration des ressources dans OFC? Au niveau du corpus, ou au niveau de l'unité composant un corpus?

- Christophe Parisse nous informe qu'E. Petitjean & C. Pestel ont recruté un stagiaire pour interroger txt dans corpus sur ortolang avec nv moteur de frantext Allegro (W sur txt & xml). Cela va nous permettre d'avancer plus rapidement que prévu sur cette question.

Comment avancer?

- nécessité d'un recrutement, stagiaire ou prestation
- Marie-Paule suggère un prestataire qui a travaillé de manière satisfaisante sur le projet Écrits scolaires
- On en avait déjà parlé, mais il faut vraiment qu'on mette en place un carnet d'adresse des prestataire qui pourra déjà prendre la forme d'un doc commun avec contacts (prestataires, stages etc) → doc. 'carnet d'adresse' (à compléter)

Etape 1: construction d'un jeu minimal de métadonnées

Deux niveaux de métadonnées: (i) au niveau du corpus / de la collection et (ii) au niveau de l'unité textuelle (e.g. texte, interaction).

On a travaillé sur le premier niveau de métadonnées, au niveau du corpus.

L'objectif est de créer une dizaine de métadonnées communes à CORLI/ARIANE.

MD communes (Sheet: Jeu métadonnées corpus minimal niveau 1)

- **Titre du corpus ou du projet**
- **Mode**: oral, écrit, multimodal
- **Domaine**: littéraire, juridique, journalistique, etc.: la colonne remplie dans le tableau ne renvoie à rien du tout
- **Genre**: conversation, roman…, il peut y en avoir plusieurs
- **Thèmes**: thèmes de conversation, par exemple cuisine; enfants…
- **Date de production**: correspond à un empan temporel (extraction automatique à partir des unités textuelles du corpus?)
- **Langue.s (norme ISO)**: ISO intègre dialectes, états de langue…
- **Responsable du corpus** (avec affiliation institutionnelle)
- **Scripteur-locuteur**: peut être anonyme (cas des écrits scolaires), dans certains cas, une extraction automatique des auteurs? Pour l'oral, si le locuteur est connu son nom aura évidemment plus d'importance, e.g. Macron.
- **Lieu**: grands champs, catégories générales
- **URL du corpus**: extraction automatique

Premier doc à soumettre (CoPil) pour relecture

Qualité des corpus et procédure de curation?
Pour Fatiha, plus-value de notre travail, label qualité de corpus = dans le projet Ariane. Cahier des charges et critères à définir. On pourrait ainsi vérifier les métadonnées et mettre un label.