RÉUNION DE TRAVAIL
OPEN FRENCH CORPUS
25 NOV. 2022

Document commun
Thesaurus 'Typologie'

Agenda

Réunion de travail 9 décembre 2022 à 9h30, en présentiel (salle de réunion du LATTICE (salle 512), 1, rue Maurice Arnoux, Montrouge) et sur Zoom
MaJ du thesaurus sur la sandbox par Alexey
Réunion de travail type atelier à prévoir (notam. ajouts MD)

Textes parlementaires - Sascha (slide 8)

Cas particulier des débats parlementaires, CR de séances publiques publiées dans JOs ; représentation différée de l'oral
Travail à partir du projet ParlaMint
La structuration interne relève des règles de l'institution et reflète l'organisation des débats à l'assemblée, idem pour MD
Importance de la prise en compte en amont des requêtes qu'on voudra utiliser

Métadonnées indispensables
Séance : ID, législature, type de session (ordinaire / extraordinaire [spécificité française]), numéro, date + heure, présidence
Sections correspondant aux points de l'ordre du jour : ID, numéro, titre, type (ouverture / clôture de séance, discussion des articles d'une loi, …)
Tours de parole retranscrits : ID, numéro, ID locuteur, type (interruption, prise de parole accordée)
Locuteur : ID, nom, groupe parlementaire, rôle dans le débat [membre d'un groupe parl., présidence, membre du gouvernement, rapporteur, …], mandat [député, ministre, …]])
Paragraphes : ID
Incidents / événements recensés dans les commentaires : ID, type (incident, kinesic, vocal), description

Métadonnées optionnelles
Au niveau du compte rendu : date de publication dans le JO
Au niveau des locuteurs :

- députés : circonscription, début du mandat, fin du mandat, parti politique, début de l'appartenance au groupe parlementaire actuel, …, fonctions / nominations au sein du parlement (membre du bureau hors présidebce, …) commission(s) + données biographiques
- membre du gouvernement : ministère, ordre protocolaire, début de la nomination

Adéquation Thesaurus Typologie
Discours institutionnel, juridique, politique
Types : argumentatif, descriptif

→ importance du rôle du locuteur: en fonction du rôle qu'occupe le locuteur dans le débat, codes différents (formules figées, voc spécifique)
Rapporteurs (députés nommés présentant loi) ont des rôles orchestrés & un langage basé sur des formules
Mandat = élément compliqué, se confond souvent avec rôle
→ incidents / événements recensés dans les commentaires : différent en fonction des pays - commentaires divers qui peuvent relever des incidents + description de l'incident
→ beaucoup de md pouvant être mise pour une personne: pb de l'étiquette 'non-inscrit' (couvre des bords politiques très différents), pourrait être précisé par le parti politique auquel la personne appartient
Données bio: genre, année de naissance
Voir slide 11: CR concernant le locuteur dans fichier root
Voir slide 12 pour exemples d'incidents

Remarques & échanges
Dans un corpus d'interactions (orales ou écrites), les données encodées se confondent ; pose la question des headers des corpus à chaque niveau.
Les MD seront différentes en fonction de la façon de considérer le document (structure vs niveau logique).
Question des CR comme étant une transcription de l'oral vs retranscription/représentation
Thesaurus typologie: développé pour le consortium CAHIER avec comme base les textes littéraires ; n'a pas été pensé pour les débats parlementaires ou les textes oraux
Thesaurus typologie: absence de domaine de texte 'politique' ; voir si administratif fonctionne ; création d'un grand domaine comprenant juridique/institutionnel avec des sous-domaines

Existence de deux niveaux distincts de représentation dans les CR: version JO & version débats

CR: tradition discursive dépendante des domaines et des institutions. Différence entre les verbatims de CR relevant de sous-domaines politiques (conseils municipaux ou assemblée nationale) et les CR universitaires.

Archives, CR, rapports - Frédérique, Virginie, Grigoriy (slide 41)

MD en cours de réflexion : distinguer TEI du doc numérique et du doc source
Travail de numérisation du doc source (cf slide 43)
Travail en diachronie
CR de CA & rapports de laboratoires
Pas de MD sur tours de parole ; split en paragraphes / tous les énoncés ne sont pas attribués à des locuteurs (travail en diachronie, certains txt datent des 70's)

Interactions orales - Carole, Christophe B, Christophe P (slide 14)

Pas de réflexion sur les MD indispensables et optionnelles
A quelle échelle faire porter les MD: texte complet ? parties?
Slide 15: organisation générale venant de TEICORPO - indications génériques 'core' ; info sur recueil de données ; info sur les locuteurs (nécessaires pour oral)

Métadonnées indispensables
Situation enregistrée (setting) : nature ; prof privé/en public ; modalité (présentiel/ tél/visio) ; milieu (académique, associatif…) ; consignes (+/- planifié/ dirigé) ; nombre de locuteurs ; lieu, date captation, responsable
Enregistrement : audio / vidéo (plusieurs vues) ; qualité (diffusion/HD… ; licence ; format ; anonymisation
Transcription : identifiant ; transcripteur ; durée ; licence ; langue ; format ; anonymisation
Locuteur : identifiant ; âge ou distinction adulte/enfant
- **Métadonnées optionnelles**

Situation enregistrée : lien avec d'autres situations (longitudinal/transversal) ; résumé ; nb transcriptions ; nb enregistrements
Enregistrement : information caméras/montage ; collecteur
Transcription : logiciels utilisés ; support du dialogue (livre, script…) ; transcripteur ; annotations (type+convention)
Locuteur : rôle dans l'interaction ; liens avec les autres locuteurs ; niveau d'expertise ; sexe/genre ; date de naissance ; niveau éducation ; lieu de naissance ; langue 1 ; autres langues ; locuteurs, nb tours de parole (? calculé) ; profession actuelle/antérieures ; lieu de résidence ; appartenance régionale dominante ; temps de parole (?calculé)
Adéquation Thesaurus Typologie
Nature + prof/privé/public + présentiel/tél/visio + milieu (académique, associatif…) + consignes (+/- planifié/ dirigé) + nb locuteurs

Remarques & échanges
Les MD pour l'oral sont peu réutilisées d'un travail à l'autre alors qu'il est très compliqué de créer des MD pour l'oral → jeu minimal de MD permettant d'être réutilisé. Besoin de MD peu contraignantes: info manquantes, besoins différents en fonction des domaines des chercheurs.
Problèmes spécifiques aux corpus oraux: anonymisation
Exemple de CR (Nanterre), voir 'weban_PV_CA_2012-12-17.pdf'

Le Thesaurus Typologie a été pensé pour les textes et pas la communication orale. Le thésaurus fonctionne par micro-thesaurus : micro-thesaurus à créer pour la communication orale avec ajout de catégories

Textes scientifiques - (slide 27)

liens avec autres référentiels très importants
Open theso: thématiques définies, peut-être + adapté à opentheso ; pas d'info sur auteurs (à ajouter ou autre référentiel pour 'créateur.s' des txt

Métadonnées indispensables
Titre, droit, emplacement, liens avec d'autres référentiels
Informations sur les auteurs
Informations sur les supports
Domaines scientifiques
Métadonnées optionnelles
Liens avec d'autres publications, avec une suite de textes, des thèmes ou projets de recherche
Adéquation Thesaurus Typologie
Genres scientifiques ?

Remarques & échanges
Problème des personnes qui parlent, des auteurs dans les textes oraux, les débats, les textes scientifiques, les textes institutionnels
Proposition de distinction entre auteur (responsabilité finale du texte) et scripteur
Auteur: responsabilités du document diffusé avec toutes les conséquences que ça implique
Scripteur: responsabilité interne dans l'institution mais pas de responsabilité 'morale'

CMC - Céline (slide 28)

Corpus se situant entre oral & écrit ; englobe technologie de communication, réseaux sociaux, sites type wiki, environnements 3d (gaming, 2nd life…)

Métadonnées indispensables
Genre(s): encodage (ex. encodage dans le <encodingDesc> pour Simuligne)
Locuteurs avec leur alias + éventuellement certaines caractéristiques démographiques / possiblement anonymisés

Remarques & échanges
MD indispensables
Mêmes problème que l'oral concernant les MD indispensables: anonymisation et locuteur.
Les MD relatives à la situation d'énonciation sont encodées à différents endroits
Question des événements: bodily activity vs onscreen
Genre du forum en plusieurs données: channel, domaine, factualité, interaction, prépa, but → Voir si adaptation possible du domaine 'factualité'

Presse - Marie (slide 35)

Corpus de presse: moins d'interaction (différée), pas de tours de parole (sauf cas spécifique de type interview), problème d'accès aux bases d'articles de presse et de diffusion, spécificité du type de presse (nationale, régionale…)

Métadonnées indispensables
id de l'article ; nom du journal ; date ; année ;

Métadonnées optionnelles
Echelle de diffusion (nat/reg) ; type de diffusion (en ligne/papier) ; langue ;
Sections/pages de publication ; titre ; auteur ;

Problème spécifique au corpus de presse: contrat d'autorisation de diffusion des données

Remarques & échanges
Les articles journalistiques sont déjà renseignés dans le thesaurus Typologie ; ajout du type de publication (périodique ou ponctuel)

Écrits scolaires - Claire (slide 36)

Spécificités des écrits scolaires (exemple illustration+texte): comprennent des écrits de l'élève, les commentaires des prof, les corrections de l'élève ; les écrits sont souvent réaliser en plusieurs étapes ; interrogation sur la notion de 'co-auteur' (rôle de l'enseignant)

Métadonnées indispensables
Transcription : identifiant, responsable (corpus, transcription, recueil), taille du document, nombre de versions, date et lieu du recueil, langue
Métadonnées concernant le scripteur : âge, genre, niveau de classe, CSP des parents, niveau d'études de la mère (pas toujours disponible), langues parlées dans la famille (pas toujours disponible)
Métadonnées concernant l'enseignant : ancienneté dans le métier, conceptions didactiques (recueillies si possible par entretien)
Métadonnées concernant le devoir : consigne, support (cahier de brouillon, cahier du jour, feuille volante), aides à l'écriture si connues

Métadonnées optionnelles
Métadonnées concernant l'établissement :
- école / collège / lycée / BTS / université
- secteur urbain / périphérie ville / rural
- Réseau d'Education Prioritaire (REP) / non REP
- secteur favorisé / mixte / défavorisé

Adéquation thésaurus Typologie
Manque
- catégorie "scripteur" (indications sur l'élève auteur du texte + précisions degré d'intervention de l'enseignant)
- catégorie "scolaire" dans "domaine du texte"
- catégorie "image" dans "origine"
- question de la catégorie "public cible"
- Introduction de sous-catégories dans "genre de texte" (peu adaptées aux textes non littéraires en l'état) celles qui sont présentes sont adaptées à la littérature mais pas aux autres types d'écrit
- Ajout d'une catégorie génétique : avant texte / texte terminé

Remarques & échanges
Dans thésaurus Typologie, manque de données scripteur ; la catégorie 'public cible' est mal adaptée aux écrits scolaires (public cible 'officiel' ou explicite vs public cible implicite (l'enseignant)

Textes juridiques - Amalia, Frédéric (slide 40)

Issu des travaux du projet DEMOCRAT

Métadonnées indispensables
- date, siècle
- organisation qui l'a émis (Commission européenne, Cour européenne de justice etc.)
- genre (arrêt, convention, décision, règlement, code civil, code pénal)
- source (numéro d'identification)
- langue de rédaction

Métadonnées optionnelles
- identifier les versions si évolution dans le temps (p.e. textes de loi)
- thème
- participants (p.e. arrêt)

Adéquation thésaurus Typologie
Typologie plutôt pour le français médiéval, à mettre à jour avec d'autres catégories

Paramètres situation d'énonciation (slide 48)

Pistes de paramètres à prendre en compte afin d'avoir des paramètres transversaux pour les corpus à disposition
Question des objets/entités communs (sur quoi portent les MD) → aide pour la réflexion en considérant les objets communs à décrire plutôt que construire sur objets très spécifiques (type cahier d'école) ; Liste d'objets saillants à déterminer
Importance des spécificités de l'écrit/oral (une dichotomie écrit/oral risque enfermement)

Thésaurus Typologie: travail de traduction, documentation accessible sur utilisation technique du thesaurus ; codes avec informations multilingues afin de faciliter prise en main ; système de synonymes