Balisage et annotation d'un corpus diachronique de comptes rendus universitaires (1964 à nos jours)

Résumé

Cette communication propose une réflexion méthodologique sur les principes de balisage d'un corpus de comptes rendus universitaires. Elle s'inscrit dans le cadre d'un projet interdisciplinaire visant à analyser des archives universitaires numérisées en mettant en relation les régularités formelles observées avec différentes déterminations (évolutions socio-historiques et législatives, les spécificités disciplinaires, ...) saisies sous l'angle des genres de discours, de leur évolution et de leur institutionnalisation.
On considère le compte rendu (CR) comme un genre " tenant lieu " d'un autre discours : le CR est un texte écrit qui doit, au sein de l'institution où il est produit, " tenir lieu " ou " représenter " un événement de parole. La question de la représentation du discours autre (RDA) y est donc cruciale.

Sur le plan matériel, le corpus est constitué à partir de boîtes d'archives qui comprennent, outre les CR, des documents annexes tels que convocations, feuilles d'émargement, textes discutés au cours de la réunion. On trouve également dans certaines boites une première version corrigée (brouillon).

Nous exposerons tout d'abord la structure des métadonnées retenue pour ce projet : basée sur l'adaptation du modèle TEI META élaboré pour la description des données orales, notre démarche s'inscrit globalement dans le travail d'harmonisation des métadonnées de l'écrit, initiée au sein de CORLI[2] dans le but de favoriser l'interopérabilité et la mise à disposition.

Les balises propres au corps du texte, dont nous expliciterons la définition et les attributs, portent sur les champs suivants :

1) structure endogène du texte : paratexte (en-tête, nom du document, signature, présents et des excusés, sommaire et ordre du jour, pagination), texte (titres de section, paragraphes, textes insérés dans le CR)...

2) informations liées au contenu du texte : noms propres et statuts des intervenants, unités thématiques...

3) catégories identifiées via une analyse linguistique : discours direct, verbes et noms de parole, embrayeurs de personne...

Le choix des balises et de leur structuration doit permettre de mettre au jour des observables permettant des explorations étroitement articulées aux questions de recherche. Ainsi la caractérisation générique du CR comme "tenant lieu" conduit-elle à se pencher sur : les formes de RDA, les séquences thématiques métadiscursives consacrées au genre même, la relation entre le CR et les textes qui y sont intégrés...

On montrera comment ces choix recoupent partiellement le modèle proposé pour les textes de

représentation. On montrera également comment l'observation des modifications apportées sur les brouillons a fait apparaître certains observables (par exemple les nombreuses modifications affectant les changements de paragraphe).

---

https://github.com/christopheparisse/teimeta

[2]https://corli.huma-num.fr/

Bibliographie

Authier-Revuz J. et Lefebvre J., 2015, " L'entretien de presse : un genre discursif de représentation de discours autre ", Revista Investigações, 28, http://www.repositorios.ufpe.br/revistas/index.php/I

De Angelis R., 2015, " La linguistique de corpus à l'épreuve du numérique : textes, textures, documents " pp.81-95, http://htl.linguist.univ-paris-diderot.fr/hel/dossiers/numero11 < https://hal.archives-ouvertes.fr/hal-01511280>

Liégeois L. et al., 2015, "Using the TEI as a pivot format for oral and multimodal language corpora". Text Encoding Initiative Conference and Member's meeting 2015, Oct 2015, Lyon, France, http://tei2015.huma-num.fr/fr/. https://halshs.archives-ouvertes.fr/halshs-01345777

Mellet C. et Sitri F., 2013, " Les formes interprétatives de représentation du discours autre dans le genre du compte rendu : analyse de différents types d'indices ", in C. Desoutter et C. Mellet (éds.), Le discours rapporté : approches linguistiques et perspectives didactiques, Peter Lang, Linguistic Insights, pp.137-158.

Sitri F., 2016, " RDA et genres du 'tenant lieu' : le cas du 'compte-rendu' ", Revista Investigações, http://www.repositorios.ufpe.br/revistas/index.php/INV/article/view/1842.

Mots-Clés: analyse de discours, genre de texte, document structuré, linguistique de corpus, archives