Synthèse de la séance ModOAP – Projet Manuels Scolaires du 02/12/2020

1. Présentation des corpus

a) Guides touristiques :

Le corpus est composé d'une sélection des guides de la BNF libres de droits, datant du XVIIe au XXe siècles. Les guides sont en français et déjà numérisés, beaucoup sont océrisés.
Les cartes et plans contenus dans les guides ont été numérisés à part.
La qualité de la numérisation est variable.
Le corpus compte environs 350 guides pour l'instant. Ils sont divisés en 3 types d'entrées :

- par *sélection*
- par *destination* : notamment Afrique/Maghreb et Asie (colonies principalement), ils sont classés par pays ou par région pour ce qui concerne la France métropolitaine.
- par *thématique* : expositions universelles, stations thermales/balnéaires, champs de bataille de la première guerre mondiale …

Les problématiques abordées pour ce corpus sont l'évolution du discours touristique, la massification et l'encadrement du tourisme, l'évolution du regard sur l'étranger, le regard du colonisateur sur le colonisé, ou le regard condescendant qui peut être porté sur d'autres pays.

b) Manuels à destination des sourds:

Pour ce corpus, 40 ouvrages sont déjà numérisés sur Gallica, et il en existe autant à la BNF non numérisés.
La question dominante concerne le décalage potentiel entre ces manuels et ceux destinés à l'éducation non-spécialisée :
- comparaison de la langue et la grammaire utilisées
- fréquences de termes et expressions, et leur évolution dans le temps
- clustering (regroupements thématiques de textes)

Il peut être intéressant de travailler sur les représentations du geste dans ces manuels.
- observation de l'évolution des signes (gestes graphiques). Il est possible de détecter ces signes.

c) Manuels scolaires :

La plupart des manuels sont déjà numérisés.

- ***Les manuels de sciences*** (Leçons de chose, sciences naturelles et physique) couvrent une période entre 1883 et 1914. Ils sont référencés dans les fiches fournies par Laurence Jung et Xavier Riondet. Il sera intéressant de retrouver les sources sur lesquelles s'appuient ces manuels, notamment dans la littérature scientifique. L'axe de recherche principal sur ce corpus concerne la représentation de la société française, notamment à travers les métiers représentés dans ces manuels destinés au peuple. Une liste de ces métiers a été envoyée par Xavier Riondet.
L'étude de ce corpus peut-être augmentée par un corpus de ***romans scolaires*** et et de ***livres de lectures*** de la IIIe République, qui véhiculent également une certaine représentation de la société de l'époque.

- ***Les manuels d'histoire*** numérisés ou en cours de numérisation ont également été listés.
L'étude des manuels d'histoire portera sur les divergences potentielles dans le traitement du roman national entre diverses sources, à travers des figures et des évènements historiques communs, mais aussi sur la création d'un roman consensuel et sa structure dans les manuels républicains.
Les propositions de problématiques de recherche concenant les manuels d'histoire et les Leçons de chose ont été précisées par Xavier Riondet.

d) Liens entre les corpus

Le projet ModOAP réunit ces trois ensembles de corpus car ils peuvent être rapprochés dans leurs natures et leurs objets d'étude, mais aussi car ils peuvent être confrontés à certains égards : il sera intéressant de comparer les manuels destinés aux sourds aux manuels non-spécialisés, ou bien les guides de voyage aux manuels, à travers la question du régionalisme par exemple.

2. Présentation d'outils en lien avec les problématiques

a) Plongements lexicaux
Intérêt des plongements de mots dans l'étude de documents historiques : la vectorisation du vocabulaire d'un corpus permet d'étudier les concepts utilisés au delà des occurrences de termes. Il est alors possible de calculer l'évolution dans le temps de la sémantique associée à un terme, et de repérer des textes sémantiquement proches de certains concepts sans que ces concepts y soient présents.
Ces techniques demandent des corpus importants pour fournir de bons résultats, ainsi que la création de sous-corpus par séries temporelles, qui doivent être vectorisés et alignés afin d'être directement comparables.

Référence :
https://www.tandfonline.com/doi/full/10.1080/01615440.2020.1760157
Melvin Wevers & Marijn Koolen (2020) Digital begriffsgeschichte: Tracingsemantic change using word embeddings, Historical Methods: A Journal of Quantitative andInterdisciplinary History, 53:4, 226-243, DOI: 10.1080/01615440.2020.1760157

b) Influence des idées dans le temps
Outil pour mesurer l'influence de certaines idées ou figures dans un corpus de discours politiques, reposant sur une approche par modélisation thématique. Cet outil permet d'observer la pérennité de certaines idées dans le temps, l'évolution du traitement d'un personnage, une religion, ou une idéologie, la capacité de certaines idées à devenir virales, quand et comment elles apparaissent ou disparaissent.
Cet outil propose une étude quantitative plutôt que qualitative, indépendante du sens des mots.
Il permettrait peut-être de dresser un arbre généalogique des textes étudiés, et de nous renseigner sur la provenance de cetaines idées.

Problèmes posés :
Une remarque a été formulée sur le caractère fermé de nos corpus : comment retracer l'influence (en amont comme en aval) des idées présentes dans nos corpus sur des textes extérieurs ?
Le même problème se pose pour la recherche des sources des manuels de science. Si certains de ces textes sont probablement disponibles sur Gallica, comment les retrouver au delà de quelques références connues ? Une première approche peut consister à chercher ces références en fonction des auteurs et des éditeurs connus des documents de nos corpus.

Référence :
https://www.pnas.org/content/115/18/4607
Barron, A. T., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French Revolution. Proceedings of the National Academy of Sciences, 115(18), 4607-4612.

c) Mesure de similarité structurelle
Outil pour mesurer la similarité structurelle de textes, utilisé sur un corpus de textes religieux : l'outil consiste à regrouper des schemes d'éléments thématiques présents dans les textes, afin de retracer leur agencement structurel.
Cet outil permet de relier des textes proches dans leur structure, même s'ils ne partagent pas un lexique commun.

Référence :
http://ceur-ws.org/Vol-2723/long47.pdf
Stine, Z. K., Deitrick, J. E., & Agarwala, N. Comparative Religion, Topic Models, and Conceptualization: Towards the Characterization of Structural Relationships between Online Religious Discourses. Proceedings http://ceur-ws. org ISSN, 1613, 0073.

d) Définir l'âge du lecteur visé par un texte
Outil MoDyCo pouvant potentiellement calculer à quelle tranche d'âge de lecteur un texte est destiné. Cet outil repose sur des critères syntaxiques, sémantiques et discursifs, et a été développé pour des textes contemporains.
Il a été précisé qu'il y a ambiguité avant 1870 sur la destination des manuels : on ne sait pas bien s'ils sont destinés au maître ou distribués aux élèves.
Les manuels sont aussi parfois hétérogènes, et présentent des segments de textes avec typographies distinctes, qui semblent être adressés à des publics différents.

La question du repérage de la complexité du français utilisé, différente de la question des tranches d'âges ciblées, a également été posée. D'un point de vue linguistique, il semble qu'il n'y ait pas de définition théorique de la complexité de la langue. Il existe cependant des mesures anciennes de lisibilité d'un texte, issues de la psycholinguistique, et implémentées dans des programmes.

3. Segmentation des pages des documents

Un travail est mené en parallèle sur la segmentation de zones au sein des pages des documents du corpus. Les outils à l'étude pour cette tâche sont Tesseract, DH Segment, et Doc Extractor.
La segmentation des pages en zones d'intérêt vise plusieurs objectifs :

- Associer les zones de légendes aux zones d'images correspondantes : on peut se demander notamment si les légendes sont systématiquement réutilisées avec les images.

- Etude de la mise en page : y a t-il des évolutions dans l'organisation des pages au fil du temps, dans la manière d'associer images et textes, ou est-ce que la maquette de ces pages reste stable au fil du temps ?
Il a été rappelé que la mise en page et le nombre d'images dans les documents sont liés à la question économique, ainsi qu'à l'évolution des techniques d'impression.

- Repérer les collocations entre éléments graphiques et textuels.
Dans les manuels scolaires, le lien entre texte et image est particulièrement intéressant, car l'image n'est pas seulement décorative mais aussi pédagogique.

- Il est possible de distinguer automatiquement images décoratives et images pédagogiques avec un classifieur entraîné à cette tâche.

- Les manuels contiennent plusieurs illustrations d'animaux, souvent utiles ou nuisibles, qu'il est également possible de classer.

- Distinguer des zones de textes descriptives de zones de textes suggérant une action du lecteur, autant dans les manuels (questions, exercices) que dans les guides (présence d'impératif, informations pratiques).

4. Résumé des axes de travail énumérés lors de la séance

- Requêtes de comparaisons textuelles
- Généalogie des textes et des notions
- Traitement des Entités Nommées
- Géolocalisation : est-ce qu'il y a des parcours repérables dans les guides et dans les textes
- Sémantisation de la mise en page, lien texte / image, etc...
- Question de la complexité du texte

Parmi ces axes, les **comparaisons textuelles et l'évolution des notions** semblent être prioritaires pour l'ensemble des corpus.

5. Prochaine séance

Pour la prochaine séance, nous proposerons sûrement des corpus à annoter en fonction des éléments qui ont été discutés, ainsi que des essais de comparaisons de textes à l'aide d'outils sous forme de Notebooks Jupyter, facilement utilisables. Cette séance permettra alors de discuter des essais réalisés et d'éventuels premiers résultats.

Il est envisagé enfin de faire appel à des étudiants intéressés pour des tâches d'annotation ou de numérisation à grande échelle.