PROGRAMME DOCTORAL MÉTHODES NUMÉRIQUES EN SHS

Contrats doctoraux Sorbonne Université de 36 mois débutant le 1er octobre 2021

Intitulé du projet de recherche doctoral (PRD)

Outils d'analyse textuelle pour le discours de presse sur des corpus historicisés

Le projet de recherche doctoral vise au repérage, à la mise en place, et au développement d'outils d'analyse textuelle simples d'usage et libres d'accès, applicable sur une grande variété de corpus, mais en particulier sur des corpus historicisés constitués de nombreux éléments liés par une chronologie forte.

• **Contexte.** — Les outils d'analyse de données textuelles sont de plus en plus nombreux, au point que des Journées internationales (JADT) soient organisées tous les deux ans depuis 1992. On peut donc s'étonner de vouloir rajouter une pierre à cet édifice déjà bien monté, d'autant que certains poids lourds émergent et s'imposent lentement.

Le premier d'entre eux, sans doute le moins légitime sur le plan scientifique mais le plus utilisé néanmoins, est le Ngram Viewer de Google : créé en 2010, il profite de l'immense bibliothèque numérique de la multinationale, lancée en 2004 et réunissant plus de 25 millions de volumes (Mallonee, 2019). Les recherches sont possibles sur plusieurs siècles (du XVIe au XXIe), en plusieurs langues. Mais la qualité de la numérisation, le choix des ouvrages, l'absence de liste exhaustive des ouvrages choisis, et l'impossibilité d'avoir accès rapidement aux textes originaux rend l'usage de cet outil problématique.

L'un des plus anciens, le Frantext de l'UMR « Analyse et traitement informatique de la langue française » (ATILF, CNRS) repose sur les travaux du Centre de recherche pour un Trésor de la langue française (CRTLF) créé en 1960, aux débuts de l'informatique. L'outil est pensé par des linguistes et repose sur un échantillon soigneusement constitué de 5 500 ouvrages du IXe au XXIe siècle. Les qualités scientifiques sont évidentes, mais l'outil souffre du manque de moyens financiers.

C'est sans oublier les nombreux couples outils/corpus créés par les spécialistes du domaine, telles les analyses de discours politiques présidentiels ou ministériels constitués dans le temps par Pascal Marchand et Pierre Ratinaud, par exemple.

• **Objectif scientifique.** — La création d'un outil en ligne et libre d'accès pour faire des analyses de données textuelles n'est pas hors de portée, à condition de pouvoir nouer des partenariats avec des acteurs qui ont déjà constitué des corpus. La base Gallica de la BNF offre plus de 6 millions de documents, et représente un point de départ idéal. En s'autorisant à créer un « Ngram Viewer » pour Gallica, on pourrait ainsi obtenir un concurrent sérieux (quoique francophone) sans avoir à

céder à la scientificité de l'approche. Il faudrait donc ouvrir la possibilité de requêtes historicisées, en définissant des bornes, mais aussi offrir des recherches de lemmes et de cooccurrences (on peut peut-être se rapprocher de l'ATILF pour utiliser leur interface).

• **Justification de l'approche.** — Le principe d'une analyse scientifique des données textuelles est fondé sur une délimitation claire du corpus (« Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », écrivait Patrick Charaudeau en 2009). L'exhaustivité du dépôt légal grâce auquel fonctionne de Gallica donne une garantie scientifique première. La possibilité de laisser l'outil en ligne rendra les études reproductibles et « falsifiables », rendant la *disputatio* possible.

• **Difficultés techniques et méthodologiques.** — Les difficultés techniques sont nombreuses, dont la première est d'abord juridique : le contenu de Gallica n'est pas ouvert, et il faudra obtenir les accords d'exploitation de la base textuelle et iconique. Les difficultés de développement sont nombreuses également, mais sans aucun doute davantage sur un plan ergonomique et conceptuel (« sciences fines ») que informatique ou logistique (« sciences dures »).

• **Recherche d'un·e doctorant· e-chercheur·se.** — Pour faire avancer le projet, l'offre d'un contrat doctoral à un·e jeune chercheur·se en sciences humaines (en sciences de l'information et de la communication, en sciences du langage, etc.) est l'assurance d'avoir une force vive capable à la fois de rechercher des solutions techniques et à la fois de les appliquer sur un corpus et une problématique de recherche précis.

• **Première bibliographie.** — Charaudeau, Patrick, 2009. « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique ». *Corpus.* Nº 8, pp. 37-66 — Froissart, Pascal & Laurence Corroy (2018). « L'éducation aux médias dans les discours des ministres de l'éducation, 2005-2017 ». *Questions de communication.* Nº 34 (décembre) — Mallonee, Laura (2019). « Is That a Hand? Glitches Reveal Google Books' Human Scanners », *Wired,* 7 février — Marchand, Pascal (2007). *Le grand oral. Les discours de politique générale de la Ve République.* Paris : INA & De Boeck — Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden (2010). « Quantitative Analysis of Culture Using Millions of Digitized Books ». *Science,* nº décembre — Roth, S. (2014), "Fashionable functions. A Google ngram view of trends in functional differentiation (1800-2000)", *International Journal of Technology and Human Interaction,* Vol. 10, nº 2, S. 34-58