La production de l'espace dans l'imprimé d'Ancien Régime: le cas de la Gazette

François Dominic Laramée
fdl@francoisdominiclaramee.com
Université de Montréal, Canada

Quelle image mentale du monde un Français de l'époque moderne pouvait-il se tracer au contact des journaux et des livres? Comment caractériser le « message géographique » transmis par les ouvrages savants, les périodiques, les descriptions et les récits de voyage du XVIIIe siècle — et peut-être discerner leur influence sur les événements historiques ?

L'étude d'une telle problématique, qui constitue le cadre de ma thèse de doctorat, implique de faire appel à un corpus vaste et diversifié qui doit être examiné à la fois à l'aide de méthodes numériques de traitement de la langue naturelle et par une lecture intensive de documents ciblés. Mais comment appliquer la lexicométrie, la fouille de textes et l'apprentissage machine à des textes en français du XVIIIe siècle, pour lesquels il n'existe pas de modèle de langage approprié dans les logiciels d'analyse courants ? Et comment adapter des algorithmes conçus pour des sources numériques récentes et de très bonne qualité à des documents endommagés par les siècles, parfois océrisés à partir de microfilms alignés de façon imprécise, ou même handicapés dès leur création par une typographie irrégulière qui mystifie les outils d'océrisation ?

Cadre théorique

Depuis les années 1980, le tournant géographique inspiré par les travaux du philosophe Henri Lefebvre (Lefebvre, 2000) a démontré que l'espace est une construction sociale. Pour les Français du XVIIIe siècle, cette construction passait le plus souvent par la lecture, seule source de connaissance couramment disponible au sujet d'espaces lointains. Comment peut-on utiliser les concepts d'espace et de lieu (Tuan, 2006), de co-présence et de mobilité (Lévy, 1999; Lussault, 2007) pour reconstruire l'imaginaire spatial suscité au sein de communautés de lecteurs par l'accès simultané aux mêmes textes (Anderson, 2006), en particulier ceux des journaux et périodiques?

Corpus et méthodologie

Cette présentation examinera les résultats obtenus lors d'une analyse d'un corpus tiré de la Gazette (renommée Gazette de France en 1762), principal périodique de nouvelles sous l'Ancien Régime (Feyel, 2000). Une version numérisée de la Gazette est disponible sur Gallica, l'archive en ligne de la Bibliothèque nationale de France ; l'état de conservation des documents imprimés à partir desquels cette version a été constituée est cependant inégal, et par conséquent le taux de succès estimé à l'océrisation varie entre 99 % et 76 % ou même moins. Les fichiers .txt issus de l'océrisation ne peuvent donc pas être employés tels quels : non seulement les formes (chaînes de caractères représentant des mots) sont-elles fréquemment endommagées, mais des informations aussi importantes que les limites séparant deux articles et les lieux d'origine de ceux-ci sont également sujettes à un taux d'erreurs inacceptable. La retranscription manuelle du corpus, formé de dizaines de milliers de page en PDF, aurait quant à elle entraîné un coût d'acquisition déraisonnable. Quant aux méthodes usuelles de correction automatique des erreurs d'océrisation (Lopresti, 2009), elles se sont révélées peu efficaces dans ce contexte, ne permettant de réduire le taux d'erreur effectif que de moins de 0,1% — et le plus souvent dans des segments du corpus sans lien direct avec les questions de recherche étudiées.

Pour traiter ce corpus, il a donc fallu faire appel à une méthode itérative, où le choix de questions de recherche auxquelles répondre a déterminé les éléments du corpus qu'il fallait reconstruire et où les résultats de l'examen d'une version transitoire du corpus a guidé le choix des questions de recherche pour l'étape suivante. Cette méthode repose sur l'identification dans le corpus, à l'aide de l'algorithme de Levenshtein (Crump, 2014), de formes potentiellement produites par une reconnaissance incorrecte d'un certain nombre de mots-clés choisis en fonction d'une question de recherche; sur l'inspection visuelle de ces formes candidates pour éliminer de la liste celles qui correspondent manifestement à d'autres mots de la langue française que les mots-clés recherchés; et sur l'extraction semi-automatisée de métadonnées pertinentes compte tenu de la question de recherche étudiée, à partir du texte océrisé et d'une inspection visuelle du document PDF d'origine. En pratique, le taux de faux positifs obtenus en détectant toutes les formes dont la distance de Levenshtein par rapport aux mots-clés est de 3 ou moins dépasse les 95%, mais la sélection visuelle des candidates véritablement prometteuses permet d'augmenter le

nombre d'occurrences utilisables pour une analyse ultérieure de 25% à 30%, et ainsi d'assurer une meilleure couverture des éléments pertinents du corpus que ce qui aurait été possible autrement. (Notons que ces occurrences récupérées incluent non seulement les résultats d'erreurs d'océrisation mais aussi des orthographes inusitées des mots-clés recherchés.)

La présentation tracera les grandes lignes de ce processus, des résultats obtenus avec la Gazette, des éléments de la méthode qui se sont montrés généralisables à d'autres corpus bruités, et des limites que la prudence impose à la fois aux questions de recherche auxquelles il convient d'appliquer une telle méthode et aux conclusions que l'on peut en tirer. Afin d'augmenter le niveau de confiance envers les résultats, une multiplicité de méthodes numériques ont été appliquées aux textes et aux métadonnées, l'utilisation d'un seul algorithme, toujours problématique (Schmidt, 2013) étant particulièrement suspecte dans un contexte où la fiabilité des données laisse à désirer. Ces multiples méthodes incluent le partitionnement (Chen et al., 2004), la cartographie numérique, divers décomptes et l'étude des cooccurrences lexicales — soigneusement contrôlée par une inspection visuelle afin d'éliminer les effets de bord causés par l'absence d'un modèle de langage approprié pour le français du XVIIIe siècle — avec le logiciel de textométrie TXM (Heiden et al., 2010). Seuls les résultats à la fois cohérents entre les différentes méthodes et trop flagrants pour être expliqués par un accident de répartition du bruit dans les textes d'origine ont été conservés pour communication.

Résultats

Les premiers résultats portent sur la représentation de l'Amérique et du monde colonial au cours de la période entre 1740 et la fin de la Guerre de Sept ans. Il a notamment été possible de démontrer que l'immense majorité des articles de presse mentionnant les colonies provenaient de Londres ou de la péninsule ibérique plutôt que de la France elle-même et qu'ils présentaient le phénomène colonial d'un point de vue étranger ; que les colonies britanniques et le Brésil occupaient une place beaucoup plus importante que les colonies françaises dans l'imaginaire spatial construit par la Gazette ; et que la sous-représentation du monde colonial français dans les textes, et en particulier celle de la Nouvelle France continentale, était exacerbée en temps de paix, où le Canada devenait pratiquement invisible. À la lecture de ces résultats, il est permis de se demander si, du point de vue d'un lecteur de la Gazette, le moment colonial français en Amérique n'aurait pas semblé chose du passé, bien avant la signature du traité de Paris qui a consacré la cession du Canada à la Couronne britannique. Des recherches sur la co-présence de différents toponymes au sein des mêmes articles, des thèmes associés à ces toponymes et à la distance imaginaire représentée par les fréquences de mentions de lieux dans la Gazette sont en cours et leurs résultats pourront être intégrés à la présentation.

Bibliographie

Anderson, B. (2006). Imagined Communities: Reflections on the Origin and Spread of Nationalism, revised edition. London: Verso.

Chen, J., Ching, R. and Lin, Y. (2004). "An Extended Study of the K-Means Algorithm for Data Clustering and Its Applications." The Journal of the Operational Research Society, 55(9): 976-987.

Crump, J. (2014). "Generating an Ordered Data Set from an OCR Text File." Programming Historian, http://programminghistorian.org/lessons/generating-an-ordered-data-set-from-an-OCR-text-file

Feyel, G. (2000). L'annonce et la nouvelle: la presse d'information en France sous l'Ancien Régime, 1630-1788. Oxford: Voltaire Foundation.

Heiden, S., Magué, J-P. et Pincemin, B. (2010). "TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement." Proceedings of 10th International Conference on the Statistical Analysis of Textual Data — JADT 2010. Rome: Edizioni Universitarie di Lettere Economia Diritto, vol. 2, pp. 1021-1032.

Lefebvre, H. (2000). La production de l'espace, 4e édition. Paris: Anthropos.

Lévy, J. (1999). Le tournant géographique: penser l'espace pour lire le monde. Paris: Belin.

Lopresti, D. (2009). "Optical Character Recognition Errors and Their Effects on Natural Language Processing." International Journal on Document Analysis and Recognition, 12 (3): 141-51.

Lussault, M. (2007). L'homme spatial: la construction sociale de l'espace humain. Paris: Seuil.

Schmidt, B. (2013). "Words Alone: Dismantling Topic Models in the Humanities." Journal of Digital Humanities, 2(1): 49-65.

Tuan, Y. (2006). Espace et lieu: la perspective de l'expérience, Gollion: Infolio.