# Lire la science avec la linguistique de corpus : le cas du data-driven learning

Alex Boulton*[1]

[1]Analyse et Traitement Informatique de la Langue Française – Université de Lorraine, Centre National de la Recherche Scientifique : UMR7118 – France

**Résumé**

La recherche est une entreprise humaine, quelle que soit la discipline mais – peut-être – surtout en SHS (sciences humaines et sociales). Pour connaître son domaine de recherche, on lit les publications des autres avant de se lancer dans ses propres réflexions et expériences. Mais quelles publications, et comment les lire ? Le choix et l'interprétation que l'on en fait sont souvent le fruit d'une sérendipité peu souhaitable, mais il est possible d'apporter un complément de rigueur scientifique à l'entreprise. Pour le choix des publications, on peut mettre en place des protocoles de recherche et d'inclusion afin d'arriver à une collection (quasi-)exhaustive selon les critères établis. Quant à la lecture, la traditionnelle synthèse narrative offre une vision très riche mais avec une part importante de subjectivité ; à l'autre extrême, une méta-analyse réduit cette subjectivité (sans jamais l'éliminer complètement), aux dépens d'une perspective plus réduite qui passe à côté de nombreux éléments potentiellement essentiels. L'objectif de cette présentation est d'explorer une autre possibilité de " lecture " d'une collection d'écrits scientifiques, assistée par la linguistique de corpus.
Le domaine en question est celui que l'on appelle communément en anglais le *data-driven learning*, que l'on peut gloser par " apprentissage sur corpus " (ASC). Une méthodologie transparente a permis de rassembler une collection de 489 publications (principalement des articles de revue, des chapitres d'ouvrage, des communications écrites) comportant une évaluation empirique de différents aspects de l'ASC. Ces textes ont été convertis au format .txt compatible avec AntConc (Anthony, 2019), un total de 2,5 millions de mots. L'analyse porte non sur l'ensemble mais sur un sous-corpus des conclusions (250.000 mots), l'objectif étant de mieux comprendre les différentes recommandations qui ont été formulées à différentes périodes : une première, plus longue en raison du petit nombre de publications, de 1989-2003 ; les autres uniformément réparties en blocs de quatre années jusqu'en 2019. Les résultats, rapportés dans un article à paraître dans la revue *Langage Learning & Technologie* en octobre (Boulton & Vyatkina, 2021), sont tirés d'une analyse des listes de fréquences des mots et des clusters, et surtout une analyse des mots-clés et clusters-clés dans chaque période. Ces éléments sont groupés en thèmes pour mieux visualiser l'évolution de l'ASC à travers trois décennies, les suggestions retenues et celles, souvent récurrentes, qui se heurtent aux limites des avancées technologiques, pédagogiques, méthodologiques ou culturelles et, surtout, des contextes de recherche.

Anthony, L. (2019). *AntConc* (v.3.5.8m). Waseda University. https://www.laurenceanthony.net/software
Boulton A., & Vyatkina, N. (2021, à paraître). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3).

*Intervenant

**Mots-Clés:** synthèse de recherche, data, driven learning, apprentissage sur corpus