Le plain language en droit : réflexions méthodologiques sur l'utilisation des méthodes d'apprentissage-machine en linguistique de corpus

Manon Bouyé*[1] and Christopher Gledhill†[2]

[1]Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole (CLILLAC-ARP) – Universit de Paris – France

[2]Centre de Linguistique Inter-langues, de Lexicologie, de Linguistique Anglaise et de Corpus-Atelier de Recherche sur la Parole (CLILLAC-ARP) – Université de Paris – France

Résumé

Cette communication vise à explorer l'outillage de l'analyse de corpus, dans le cadre des langues contrôlées appliquées à la diffusion du discours juridique auprès du grand public. En particulier, nous nous intéressons à des textes de droit et de diffusion du discours juridique relevant du *Plain Language Movement* (ci-après PLM). Ce mouvement, à travers la prescription de termes et structures linguistiques à utiliser lorsqu'on s'adresse aux justiciables, vise à rendre la langue du droit " claire et simple".
Dans les guides rédactionnels pour un *plain language*, certains traits lexico-grammaticaux sont encouragés, comme l'utilisation de pronoms personnels de deuxième personne et de phrases d'une longueur maximale de 25 mots. D'autres à l'inverse sont découragés, typiquement l'emploi du passif, l'utilisation de certains mots jugés trop complexes par les institutions et organismes qui se réclament du *plain language*. Grâce à une méthodologie empruntant à la linguistique de corpus, à l'analyse de discours et au TAL, nous nous proposons d'interroger la pertinence de ces recommandations en tentant de répondre à la question suivante : les caractéristiques des textes se revendiquant clairs et simples et destinés aux justiciables, par comparaison à leur version dite complexe, correspondent-elles aux traits linguistiques se trouvant dans les recommandations du PLM ?

Notre étude porte donc sur deux corpus, l'un composé de textes juridiques anglophones, intitulé LEX (plus de 2 millions de mots), l'autre de leurs versions en *plain language* nommé PLAIN (929 000 mots). Les métriques de complexité linguistique mesurées sur ces corpus, issues du projet Ulysses 2019 (Université de Paris / Galway), incluent notamment des scores de lisibilité fondés sur la longueur des phrases et des mots, des métriques de densité et de variation lexicale, mais aussi de complexité syntaxique. Un algorithme de classification supervisé permet de prédire la catégorie des textes *(Lex* ou *Plain*) à partir de ces métriques ; l'algorithme indique ensuite les métriques qui sont les plus pertinentes pour classer les textes. Ainsi pouvons-nous les comparer aux consignes officielles du PLM, et voir si ces recommandations renvoient aux traits lexico-grammaticaux qui sont effectivement les plus adéquats pour différencier les textes dit complexes des textes simplifiés. Les résultats obtenus nourrissent une réflexion méthodologique sur l'utilisation d'outils statistiques et d'apprentissage-machine en analyse de discours, en particulier sur les biais que les métriques utilisées peuvent

*Intervenant

†Auteur correspondant:

entraîner.

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., & Zarrouk, M. (2019, September). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning* (pp. 308-320). Springer, Cham.

Cutts, M. (2013). *Oxford guide to plain English.* Oxford University Press, USA.

Gibbons, J. P. (Ed.). (2014). *Language and the Law.* Routledge, 11-49.

Gledhill, C., Martikainen, H., Mestivier, A., & Zimina-Poirot, M. (2019). Towards a Linguistic Definition of 'Simplified Medical English': Applying Textometric Analysis to Cochrane Medical Abstracts and Their Plain Language Versions. *LCM-La Collana/The Series*, 91-114.
Lu, X. (2014). *Computational methods for corpus annotation and analysis.* Springer.

Mots-Clés: apprentissage supervisé, guides rédactionnels, médiation juridique, métriques de complexité, plain language