Opportunités et limites en stylistique computationnelle de la poésie : Détection automatique de l'enjambement en anglais

Eulalie Monget, Pablo Ruiz Fabo
Université de Strasbourg, LiLPa UR 1339 / F-67000 Strasbourg, France

Plusieurs initiatives internationales témoignent de l'intérêt actuel pour les analyses littéraires assistées par des moyens computationnels, comme le groupement *Digital Literary Stylistics (SIG-DLS)* de l'*Alliance for Digital Humanities Organizations*. Le colloque *Plotting Poetry* montre la variété de phénomènes poétiques abordés à l'aide d'outils informatiques (https://plottingpoetry.wordpress.com/). Un numéro thématique de la revue *Langages* (2015) en fournit une autre synthèse.

Les projets impliquant l'annotation linguistique automatique pour l'analyse littéraire partagent certains soucis : comment opérationnaliser des concepts d'analyse littéraire sur la base d'annotations issues des outils de Traitement automatique des langues (TAL), originalement conçues pour une analyse non-littéraire ? Comment évaluer nos annotations automatiques d'un trait stylistique, en termes de, et au-delà de, la comparaison avec des données de référence annotées manuellement ? Quel gain de connaissance spécifique à l'analyse littéraire est atteignable à travers l'annotation stylistique automatique, qui serait impossible sans traitement informatique ?

C'est des questions qui nous occupent également dans notre projet sur la détection automatique de l'enjambement dans la poésie en anglais. L'enjambement implique une discordance entre les pauses requises par la structure métrique (fins de vers ou hémistiche) et des pauses demandées par la syntaxe ou le sens (cf. Golomb, 1979, p. 269). On le rencontre souvent lorsqu'un un syntagme est éclaté sur deux vers successifs, contrariant l'attente d'une pause à la fin du premier vers. Hors cette caractérisation générale, la définition de l'enjambement ne fait pas consensus (cf. Quilis, 1964 ; Hollander, 1975 ; Golomb, 1979 ; Hussein et al., 2018 ; Delente, 2019). C'est une raison pour développer des logiciels qui implémentent les différentes définitions possibles : en détectant automatiquement leurs occurrences sur un grand corpus, les atouts et limites de chaque définition pourront être examinés au vue d'un exemplier large. De plus, il n'y a pas d'études publiées sur la détection automatique de l'enjambement en anglais, contrairement à l'allemand (Hussein et al., 2018) ou espagnol (Ruiz et al., 2017, http://prf1.org/anja/index/).

Concernant l'opérationnalisation, nous adoptons une définition à base largement syntaxique (Quilis, 1964) : l'enjambement se produit quand la fin de vers coupe certaines séquences à forte cohésion interne. Ses atouts : premièrement, la facilité d'opérationnalisation. La définition implique des séquences d'étiquettes grammaticales, des dépendances et constituants syntaxiques, fournis par les librairies de TAL. Deuxièmement, l'intérêt de vérifier si cette approche, déjà appliquée en espagnol (Ruiz et al., 2017), serait applicable à l'anglais. Nous avons constaté des limites, ayant modifié la typologie pour mieux gérer l'anglais (voir https://git.unistra.fr/enj/corpus-reference) ; plus généralement, Delente (2019) discute les limites des définitions syntaxiques.

La qualité des résultats du TAL décroît pour les textes littéraires (Bamman, 2017). Or, des gains de qualité dans multiples tâches de TAL ont été récemment obtenus par les modèles neuronaux, que nous exploitons, avec les librairies spaCy (Honnibal et Montani, 2017) et AllenNLP (Gardner et al., 2017). Notre étude est une opportunité pour tester leur robustesse sur un corpus exigeant.

Pour l'évaluation, nous avons annoté manuellement l'enjambement, selon notre définition, dans 60 poèmes de genres variés des 19e et 20e siècles (voir https://git.unistra.fr/enj/corpus-reference). Le corpus servira à comparer la détection automatique avec l'annotation humaine. Au-delà, on voudrait annoter automatiquement un corpus diachronique large pour déceler de possibles tendances dans la distribution de l'enjambement.

Le projet permet une réflexion sur plusieurs sujets : l'adoption et adaptation de technologies linguistiques pour l'opérationnalisation de concepts littéraires, les problèmes d'évaluation en annotation stylistique automatique, et le potentiel et limites des approches pour contribuer à des nouvelles connaissances en littérature.

Références

Bamman, D. (2017). Natural Language Processing for the Long Tail. In *Digital Humanities 2017*. Montréal. https://dh2017.adho.org/abstracts/408/408.pdf

Delente, É. (2019). Le traitement automatique du rythme régulier. Un cas particulier : L'enjambement. Présenté à *Plotting Poetry (and Poetics) 3 - Machiner la poésie (et la poétique) 3*. 26 septembre 2019, Nancy. plottingpoetry.wordpress.com/programme-3

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M., & Zettlemoyer, L. (2018). AllenNLP : A Deep Semantic Natural Language Processing Platform. *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 1–6. https://doi.org/10.18653/v1/W18-2501

Golomb, H. (1979). *Enjambment in poetry : Language and verse interaction*. The Porter Institute for Poetics and Semiotics. Tel Aviv University.

Hollander, J. (1975). « Sense variously drawn out » : On English enjambment. In *Vision and Resonance* (p. 91‑116). Oxford University Press.

Honnibal, M. & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://github.com/explosion/spaCy

Hussein, H., Meyer-Sickendiek, B., & Baumann, T. (2018). Automatic Detection of Enjambment in German Readout Poetry. *9th International Conference on Speech Prosody 2018*, 329‑333. https://doi.org/10.21437/SpeechProsody.2018-67

Quilis. A. (1964). *Estructura del encabalgamiento en la métrica española. [La structure de l'enjambement dans la métrique espagnole]*. Consejo Superior de Investigaciones Científicas, patronato Menéndez y Pelayo, Instituto Miguel de Cervantes.

Revue Langages (2015). *Traitement automatique des textes versifiés : Problématiques et pratiques.* Numéro coordonnée par Éliane Delente et Richard Renault. *Langages, 199.* https://www.cairn.info/revue-langages-2015-3.htm

Ruiz, P., Martínez Cantón, C., Poibeau, T., & González-Blanco, E. (2017). Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 27‑32. https://doi.org/10.18653/v1/W17-2204