Stylométrie, ADT et deep learning. Une étude de cas sur la prose romanesque de Milan Kundera

Federica Beghini

Stylométrie, ADT et deep learning. Une étude de cas sur la prose romanesque de Milan Kundera

Federica Beghini[1]

[1]Università degli Studi di Padova,
Université Côte d'Azur (Laboratoire BCL) – federica.beghini@phd.unipd.it

Abstract

This article presents a study of characteristic morphosyntactic elements of Milan Kundera's novel production, which are detected by linguistic, statistical and machine learning approaches. The specificity of this contribution is to propose, in addition to the traditional statistical methods, a deep learning training on a database comprising Kundera's novels and a representative sample of French contemporary novels, using grammatical categories as sole representation of the texts. This study led to the identification of a morphosyntactic pattern, which was then examined to reveal the aesthetic intention behind its use.

Keywords: French linguistics, stylometry, French literature, deep learning.

Résumé

Cet article présente une étude des éléments morphosyntaxiques caractéristiques de la production romanesque de Milan Kundera, qui sont détectés par des approches linguistiques, statistiques et d'apprentissage profond. La spécificité de cette contribution est de proposer, outre les traditionnelles méthodes d'exploration statistique, un entraînement d'apprentissage profond sur une base comprenant les romans de Kundera et un corpus représentatif des romans contemporains, en utilisant les catégories grammaticales des mots comme seule représentation des textes. Cette étude a permis d'identifier un patron morphosyntaxique, qui a été ensuite examiné pour révéler l'intention esthétique qui en sous-tend l'emploi.

Mots clés : linguistique française, stylométrie, littérature française, deep learning.

1. Introduction

Ces dernières années, le domaine des études textométriques a connu un intérêt croissant pour les approches qui, outre l'étude traditionnelle des unités simples d'un texte (forme lexicale, lemme, catégorie morphosyntaxique, ponctuème), visent également à examiner les relations syntagmatiques que celles-ci forment entre elles. Ainsi, dans les deux dernières décennies, la textométrie a abordé la question de l'étude des unités complexes, ou, selon les termes de Legallois (2018), « des séquences d'unités », « des schémas de phrase », des « patrons » ou des « motifs ». À cette fin, différentes démarches ont été proposées et explorées, relevant de l'ADT (Legallois, 2018 ; Longrée et *al*. 2008, 2013) et du *deep learning* (Magri 2020 ; Mayaffre et Vanni, 2021). Cette étude s'inscrit dans cette lignée de recherche : son objectif correspond à l'identification des patrons prototypiques de la prose de Kundera, plus précisément des schémas de phrase morphosyntaxiques les plus significatifs de son écriture, par rapport à celle d'un

échantillon représentatif de la littérature française contemporaine à notre auteur. Dans ce but, une approche innovante des méthodes du *deep learning* pour l'étude des unités complexes d'une œuvre littéraire a été élaborée, employée conjointement aux outils d'exploration statistique.

2. Une méthodologie intégrée

La notion de « motif », définie par Longrée *et al*. (2008 ; 2013), Legallois (2018), Mayaffre et Vanni (2021), correspond à l'association récurrente de certains éléments textuels, continus ou discontinus, pouvant appartenir à différents niveaux linguistiques (lexique, catégories grammaticales, syntaxe) qui viennent former des configurations ou des patrons lexico-grammaticaux. Dans cette étude, nous ne concentrerons notre analyse que sur le plan morphosyntaxique, à savoir sur les unités complexes les plus saillantes des catégories grammaticales.

Cette recherche porte sur les neufs romans de l'*Œuvre I et II* de Kundera (2016, Gallimard), pour un total de 823.261 occurrences. L'analyse de ce corpus se sert d'un étalon de référence qui a été compilé pour qu'il puisse représenter de façon homogène la langue littéraire française du genre romanesque de la période contemporaine à l'activité littéraire de cet auteur (1960-2019). Ce modèle de référence contient 35 romanciers pour un total de 9.418.136 occurrences.

L'étude des configurations morphosyntaxiques a été menée à l'aide d'une approche *corpus-based* : l'exploration quantitative du corpus s'est basée sur des hypothèses dérivant d'une étude qualitative de l'œuvre de Kundera. Plus précisément, par le biais de l'exploration statistique et de l'apprentissage profond, nous avons mis à l'épreuve nos considérations qualitatives concernant la récurrence d'un certain type de groupe nominal, à savoir [Dét. + N1 + Prép. + Dét. + N2].

Ces analyses quantitatives ont été menées à l'aide du logiciel Hyperbase (version web et version standard) et elles s'appuient à la fois sur l'exploration statistique et sur l'apprentissage profond pour une classification de textes (Savoy, 2015) – plus précisément, une tâche d'attribution – qui utilise des réseaux neuronaux convolutifs (*convolutional neural network*) (Kalchbrenner *et al*., 2014). En effet, depuis les travaux de Vanni *et al*. (2018), ces modèles ne se limitent plus à la tâche de classification, mais autorisent également un mécanisme inverse, celui de la déconvolution, en mesure d'extraire les éléments textuels à la base de cette classification, à savoir les *marqueurs linguistiques* sur lesquels repose son choix.

Ainsi, en ce qui concerne le *deep learning*, nous avons effectué autant d'entraînements que le nombre de romans de Kundera à classifier, à savoir neuf, seulement sur les catégories grammaticales. Pour chacune de ces neuf analyses, l'algorithme s'est entrainé sur une base contenant à la fois le corpus de référence et tous les romans de Kundera sauf celui à classifier. Par exemple, si le roman à classifier était *L'Identité*, il était enlevé de la base et, après l'entraînement, il était soumis à l'algorithme pour la tâche de classification ; la même procédure a été employée pour tous les romans de Kundera. Les textes de notre auteur ont tous été reconnus comme étant de lui, avec des taux de précision variant entre 40% et 70%, avec l'exception du premier roman, qui est attribué à Kundera, mais avec un taux de précision plus bas, de 24%. Enfin, tous les résultats de ces neuf tâches de classification ont été examinés et, dans la prochaine section, nous présenterons des marqueurs linguistiques qui reviennent et qui peuvent donc être considérés comme des éléments linguistiques distinctifs de la prose de Kundera par rapport à celle de notre échantillon représentatif de la prose romanesque contemporaine à l'auteur.

3. ADT et *deep learning*. Les configurations morphosyntaxiques

Tout d'abord, nous avons examiné la structure [Dét. + N1 + Prép. + Dét. + N2] à l'aide des méthodes statistiques. Un calcul des spécificités (fig. 1) a confirmé sa significativité dans les romans de Kundera et une surutilisation dans son œuvre par rapport à ce qu'on observe dans la langue littéraire contemporaine, avec un écart réduit de +35.4.

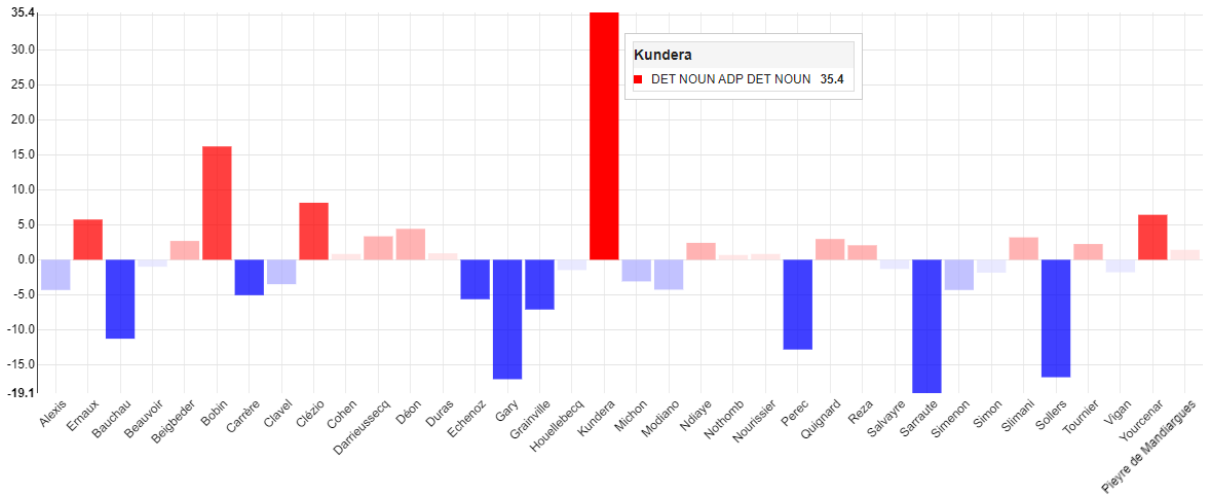

Fig. 1 – Le schéma [Dét. + N1 + Prép. + Dét. + N2] (Hyperbase – 2022)

Ce « schéma de phrase » a également été relevé dans les classifications du *deep learning* : dans les zones d'activation les plus significatives – c'est-à-dire qui présentent un TDS élevé (Vanni *et al*., 2018), – le processus de déconvolution a détecté des suites de codes qui le rappellent.

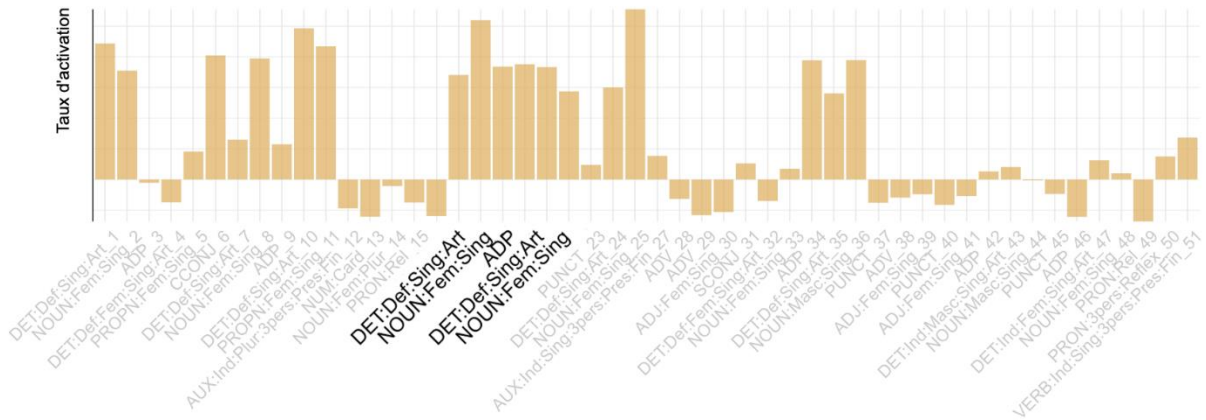

Fig. 2 – Le deep learning. Le marqueur linguistique [Dét + N + Prép + Dét + N] (Hyperdeep – 2022)[1]

Étant donné l'impossibilité d'inclure toutes les sorties des neuf classifications, nous rapportons ici sous forme discursive les marqueurs linguistiques qui, en plus d'avoir un TDS élevé, sont communs aux neuf classifications. La même séquence de composants de la structure que nous venons de relever se succède (fig. 2), bien que souvent seules des parties du patron apparaissent, comme [N + Prép + Dét + N] et [Prép + Dét + N]. Les résultats qui ne se réfèrent pas spécifiquement à cette suite de catégories grammaticales, tels que [N...N] ou [Dét + N... Dét +

[1] « Pour avoir la certitude que l'amitié érotique ne cède jamais à **l'agressivité de l'amour**, il ne voyait chacune de ses maîtresses permanentes qu'à de très longs intervalles. » (Kundera, 2016, p. 1149).

N], détectent néanmoins la saillance de la succession des substantifs. Enfin, d'autres zones d'activation mettent en évidence des variantes concernant la présence des adjectifs qui précèdent ou suivent les noms ; le cas d'adjectifs antéposés est le plus significatif, comme dans le marqueur [Prép + Dét + Adj + N].

Ainsi, la suite de codes relevée par une étude qualitative du texte et vérifiée par la statistique a également été détectée par l'apprentissage profond. En plus de démontrer sa significativité, le *deep learning* nous a également permis d'en identifier des variantes [Prép + D + *Adj* + N] et de relever la saillance de la catégorie grammaticale du substantif.

3.1. Le retour au texte. À la recherche des motifs morphosyntaxiques

Les sorties du *deep learning* ont orienté notre étude du concordancier, dans laquelle nous avons donc décidé d'inclure des variantes contenant l'adjectif antéposé et postposé. Ce retour au texte par le biais de l'analyse des cooccurrences sert à l'identification des constantes de ce schéma grammatical, afin de dégager l'intention esthétique qui en sous-tend l'emploi. Pour des raisons d'espace, nous ne rapportons qu'une sélection des exemples les plus significatifs.

1	Curieuse alliance :	**la froide impersonnalité de la technique**	et les flammes de l'extase.
2	Imprimer une forme à une durée, c'est	**l'exigence de la beauté,**	mais aussi celle de la mémoire.
3	Elle s'identifia donc, avec une sorte de passion masochiste, à	**la vulnérable nudité de son visage**	sur lequel s'imprimaient toutes les traces de ses souffrances.
4	Je corrige donc mon diagnostic : « Le malade souffre de	**la déformation masochiste de sa mémoire. »**	En effet, il ne se souvient que des situations qui le rendent mécontent de lui-même.
5	… il ne comprend rien à	**la valeur de l'insignifiance.**	Voilà ma réponse à ta question sur le
6	En cela nous sommes tous égaux : Banaka, Bibi, moi et Goethe.	**L'irrésistible prolifération de la graphomanie**	parmi les hommes politiques, les chauffeurs de taxi, les parturientes, les amantes, les assassins, les voleurs […]
7	Les trois hommes nus, guidés par	**les injonctions muettes de son regard,**	se penchaient sur la petite et lui enlevaient le reste de ses vêtements.
8	Il imaginait	**l'histoire de son corps :**	il était perdu parmi des millions d'autres corps
9	Il faut comprendre	**le paradoxe mathématique de la nostalgie :**	elle est la plus puissante dans la première jeunesse quand le volume de la vie passée est tout à fait insignifiant.
10	Son bâillement n'avait ni commencement ni fin, c'était	**le bâillement infini de la mélodie wagnérienne :**	la bouche se fermait sans se clore tout à fait, elle s'ouvrait encore et encore, tandis que, à contretemps, les yeux aussi s'ouvraient et se fermaient.
11	[…] ils savaient fort bien qu'elles se répétaient avec	**la stupéfiante régularité de la probabilité statistique.**	Ils savaient que quelqu'un allait certainement

Figure 3 – Les occurrences du schéma [Dét. + (Adj) + N1 + (Adj) + Dét. + N2]

Au niveau syntaxique, les occurrences de la figure 3 partagent la même construction, qui correspond à un **groupe nominal étendu** : un groupe nominal dans lequel le nom tête est modifié par un groupe prépositionnel modifieur du nom [Dét + N_1 + Prép + Dét + N_2]. Cette structure peut s'actualiser dans trois cas différents, selon que le nom tête est N_1 ou N_2 (Riegel, M. *et al*., p. 346-349) ; dans notre cas, le nom tête correspond à N_1.

D'un point de vue syntaxique, N_2 est toujours le complément du nom de N_1, mais la relation entre les deux noms change. Il existe des relations de sujet à prédicat, où un verbe ou un adjectif est nominalisé (2, 4, 6, 7), par exemple la tournure « c'est l'exigence de la beauté » a été préféré à « la beauté exige que … ». La préposition *de* qui relie les deux substantifs peut indiquer d'autres types de relation, notamment la relation de possession au sens large, plus précisément de *propriété* ou de *qualité* (3, 8, 11), par exemple « la vulnérable nudité de son visage » a été choisi à la place de « son visage se caractérise par une nudité vulnérable ». Enfin, une relation attributive se forme entre les deux noms quand la préposition *de* instaure « un rapport de

catégorisation discursive […] entre un nom à valeur générale classifiante et le référent particulier désigné par le nom tête complément » (Riegel, M. *et al*., 1994, p. 348) : on ne dit pas que « la technique est froide et impersonnelle », mais on parle de « la froide impersonnalité de la technique » et ainsi de suite (1, 3, 9). De plus, le marqueur linguistique détecté par le *deep learning* (fig. 2) s'inscrit dans ce troisième cas : à « l'amour est agressif », « l'agressivité de l'amour » a été préféré. Dans de nombreux cas, cette relation attributive est également caractérisée par un langage figuré : par exemple, la nostalgie est comparée à un paradoxe mathématique. On trouve également une tendance à la personnification (« la froide impersonnalité de la technique » ; « l'agressivité de l'amour ») et parfois à la concrétisation des concepts abstraits (« les flammes de l'extase »).

Le dénominateur commun de ces schémas est la présence d'une **prédication seconde** et d'un processus de **nominalisation**. En effet, un syntagme nominal a été préféré à la prédication première : par exemple, « son regard enjoignit aux trois hommes de » ou « de son regard, elle enjoignit aux trois hommes de » sont synthétisées par le GN « les injonctions muettes de son regard ». Dans le cas des relations de propriété, l'énoncé « elle se répétait avec la stupéfiante régularité de la probabilité statistique » a été choisi au lieu de « elle se répétait avec une régularité stupéfiante semblable à celle de la probabilité statistique ». Dans le troisième cas, la mémoire n'est pas déformée d'une façon masochiste, mais il est question de « la déformation masochiste de sa mémoire ». Ainsi, nous remarquons une tendance vers la nominalisation, le nom étant préféré au verbe ou à l'adjectif.

Dans certaines de ces occurrences, notamment dans les constructions attributives et dans certains des cas où il existe une relation de propriété entre les deux noms, on a affaire également à ce que Molinié définit comme la **métonymie de l'abstraction** (2011, p. 106). En effet, N_1 correspond à une propriété de N_2, propriété qui devient alors désignation. Par exemple, la structure « l'impersonnalité de la technique » permet d'accorder une plus grande importance à cette propriété, l'impersonnalité, qui aurait autrement été moins mise en valeur, si elle avait été présentée sous forme d'adjectif dans le syntagme « la technique impersonnelle » ou la phrase « la technique est impersonnelle ».

3.1.1. L'intention esthétique

Aller vers la nominalisation, vers le concept, permet ainsi de conserver les deux noms dans leur signification intégrale : avec « la technique est impersonnelle » il y aurait eu une déperdition du sens du concept signifié par l'adjectif. De plus, la prédication seconde et la nominalisation donnent lieu à un effet de synthèse : à une prédication première est préférée la force synthétique du groupe nominal étendu. Cette structure condense plusieurs significations relatives à un objet donné, en l'occurrence « la technique », de manière à évoquer dans l'esprit du lecteur une série de composants sémiques qui peuvent aider à mieux cerner sa réflexion. Il s'agit parfois de connotations qui prennent forme grâce à d'autres passages de son œuvre.

L'effet esthétique qui en résulte nous semble concorder avec sa poétique : ces schémas de phrases présentent un style concis qui, malgré sa précision et sa brièveté (ou grâce à elles) parvient à contenir – ou plutôt à donner à entrevoir – l'infinie richesse des connotations d'un thème ou d'une impression. De cette façon, ce choix assure ainsi une exploration du sens qui n'est jamais arrêtée (une « irrésistible prolifération » de ses chemins) et qui se découvre continuellement stimulée.

Ce foisonnement du contenu sémantique de ces syntagmes nominaux rappelle le concept de variations sur un thème que Kundera lui-même théorise en référence à sa poétique : son œuvre

est un grand roman contenant un certain nombre de *grands* thèmes qui sont développés sous forme de variations. Ces thèmes, « tels des fleuves, parcourent toute l'étendue de l'œuvre kundérienne sans se laisser interrompre par aucune frontière, ramassant sur leur passage et répandant dans chaque roman qu'ils irriguent un limon sémantique proprement inépuisable » (Ricard, 2003, p. 62). Ainsi, l'analyse existentielle de ces thèmes, ou, selon les termes de Kundera, l'exploration « d'une portion jusqu'alors inconnue de l'existence » est au cœur de sa poétique romanesque et c'est l'intention esthétique qui guide ses choix (Kundera, 2016, p. 707).

Il n'est donc pas inattendu que la significativité de cette structure ait été vérifiée par les approches statistiques et d'apprentissage profond, vu qu'elle réalise l'intention de saisir le foisonnement et la complexité des significations de plusieurs thèmes et situations existentielles, dont la composition sémique peut se trouver continuellement renouvelée et enrichie de nouveaux sèmes.

4. Conclusion

Nos hypothèses qualitatives, concernant la significativité d'une structure nominale particulière, ont été vérifiées grâce à l'apport à la fois de l'exploration statistique et de l'apprentissage profond ; plus précisément, le *deep learning* nous a permis d'en identifier des variantes et de relever, plus en général, la saillance des groupes nominaux.

Par conséquent, cette méthodologie intégrée nous a permis d'identifier et de définir un motif morphosyntaxique significatif de sa prose qui la distingue de celle de ses contemporains et de l'interpréter d'un point de vue linguistique et littéraire, de façon à déceler l'intention esthétique qui en sous-tend l'emploi. Cette démarche étant une nouvelle approche encore à explorer, nous visons à l'approfondir dans de futures études pour pouvoir détecter et examiner d'autres patrons morphosyntaxiques.

Bibliographie

Kalchbrenner N., Grefenstette E. et Blunsom P. (2014), « A convolutional neural network for modelling sentences », *52th Annual Meeting of the Association for Computational Linguistics*, p. 655–665.

Kundera M. (2011), *Œuvre I et II*, Paris, Gallimard, Coll. Bibliothèque de la Pléiade, 2016.

Legallois D. (2018), « Les motifs lexico-grammaticaux : une nouvelle approche en stylistique », in *Stylistique et méthode. Quels paliers de pertinence textuelle ?*, Lyon, Presses universitaires de Lyon.

Longrée D., Luong X. et Mellet S. (2008), « Les motifs : un outil pour la caractérisation topologique des textes », in Serge Heiden, Bénédicte Pincemin (eds.), *JADT 2008*, Lyon, Presses de l'ENS.

Longree D., Mellet, S. (2013), « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraseologie de la langue au discours », *Langages* 189, p. 65–79.

Magri V. (2020), « Le deep learning comme défi pour identifier le style d'un écrivain : l'exemple de Jean Giono », *JADT 2020*, Jun 2020, Toulouse, France. ⟨hal-02936437⟩.

Mayaffre D. et Vanni L. (2021), *L'intelligence artificielle des* textes, Paris, Honoré Champion.

Molinié G. (2011), *Éléments de stylistique française*, Paris, PUF.

Ricard F. (2003), *Le Dernier Après-midi d'Agnès*, Paris, Gallimard.

Riegel, M., Pellat J.-C. et Rioul R. (1994), *Grammaire méthodique du français*, Paris, PUF, 2011.

Savoy J. (2016), « Estimating the probability of an authorship attribution », *JASIST*, 67(6), p. 1462–72.

Tuzzi A. Cortelazzo M. (eds. ; 2018), *Drawing Elena Ferrante's Profile*, Padova, PUP.

Vanni, L. *et al*. (2018), « Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis », Melbourne, *56th Annual Meeting of the Association for Computational Linguistics*, ⟨hal-01804310⟩.