Travaux neuchâtelois de linguistique, 2024, 79, 21-36

Stylometry and Deep Learning: A case study on Milan Kundera's *Le Livre du rire et de l'oubli*

Federica BEGHINI
Università degli Studi di Padova, Université Côte d'Azur
federica.beghini@phd.unipd.it

This study aims to uncover the prototypical linguistic elements and patterns of Kundera's prose in his novel *Le Livre du rire et de l'oubli* (1979, Gallimard). The exploration employs statistical and machine learning techniques, including the application of Hyperbase in both its web and standard versions. Hyperbase provides deep learning features for text classification tasks (Savoy 2015; Tuzzi & Cortelazzo 2018), based on convolutional neural networks (Kalchbrenner *et al.* 2014; Kim 2014) which go beyond the process of convolution and incorporate an innovative deconvolution mechanism that extracts key linguistic markers essential for classification purposes (Vanni *et al.* 2018; Mayaffre & Vanni 2021). The training of the Hyperbase deep learning model involves an extensive corpus containing novels by 36 authors, including Kundera, thus encompassing the French literature landscape from 1960-2014. The study leads to the identification of linguistic markers related to vocabulary, morphosyntax, lexical and grammatical patterns, and lexico-grammatical structures. These markers are then examined to reveal the underlying aesthetic intentions of the author. The conclusion focuses on the contribution of deep learning and statistics in the context of this qualitative linguistic study of a literary text.

1. Introduction

This research endeavours to identify and analyse the distinctive textual elements of Kundera's prose in *Le Livre du rire et de l'oubli* (1979) through an integrated analysis. The term "integrated" is used because a combination of qualitative and quantitative approaches, including deep learning and statistical methods, were employed to offer a more comprehensive study.

Specifically, the amalgamation of new deep learning methodologies with traditional statistical analyses of textual data enabled the exploration not only of simple units, but also complex units or patterns. This approach facilitates a more profound and comprehensive linguistic study, encompassing lexical, grammatical, and lexico-grammatical patterns.

1.1 State of the art

In recent years, the fields of textometry[1] and stylometry[2] have witnessed a growing interest in approaches that go beyond the conventional study of simple units in a text (e.g., lexical form, lemma, morphosyntactic category, punctuation mark) and also aim to study the syntagmatic relations that they form among one another. To this end, over the past two decades, within the fields of stylometry

[1] A discipline that analyses text corpus by means of computer processing (Mayaffre *et al.* 2019; Pincemin 2012, 2020; Magri 2020).

[2] Stylometry is the term employed when textometric analysis is used to "characterise a style of writing" (Magri 2010; Holmes 1998).

and textometry there has been a continuous increase in interest in the study of complex units, referred to as "des sequences d'unités" ("sequences of units"), "des schémas de phrase" ("sentence patterns"), "patrons" or "motifs" ("patterns"), as per Legallois (2018).

Diverse approaches, including *statistical analysis of textual data* (Legallois 2018; Longrée *et al.* 2008, 2013) and *deep learning methods* (Mayaffre & Vanni 2021), have been proposed and explored. Notably, the application of deep learning has proven to be fundamental to the study of these complex entities (Mayaffre & Vanni 2021; Vanni *et al*. 2022; Thon *et al*. 2022), as will be shown in this research.

This study aligns with this trajectory of investigation. The subsequent section details the methodology adopted for this specific objective of identifying and examining the prototypical simple and complex linguistic units of Kundera's prose within his novel. In this context, "prototypical" denotes the most salient linguistic features that differentiate Kundera's writing from that of a representative sample of contemporary French novelists.

1.2 Methodology

As previously stated, our approach combines both qualitative and quantitative dimensions. More specifically, the textual data was collected using deep learning and then corroborated through a statistical exploration. These data were then subjected to qualitative scrutiny through linguistic analysis.

Our deep learning method corresponds to the text classification algorithm of Vanni *et al*. (2018), grounded in convolutional neural networks (CNN, Kalchbrenner *et al.* 2014; Kim 2014). In addition to the conventional *convolution* mechanism facilitating text classification, Vanni *et al.* (2018) have introduced an inverse *deconvolution* mechanism to identify the linguistic components on which the algorithm has based its choice of classification. For example, what textual units prompted the algorithm to attribute this text to author X instead of author Y? What linguistic features led it to classify this text under genre X rather than genre Y? In this sense, the deconvolution mechanism represents the innovative aspect of the method, as it tries to solve the so-called *black box problem.*[3] For an engineering-oriented exploration of the deconvolution method's functionality and its effectiveness in identifying linguistic markers, the interested reader can refer to the developer's research (Vanni *et al*. 2018; Vanni 2021).

In contrast to our previous work (Beghini 2022) where this deep learning method was tested using a *corpus-based* approach, this research is conducted using a *corpus-driven* approach, i.e., quantitative data are observed without pre-existing qualitative assumptions.

[3] The "black box" corresponds exactly to all the elements on which the classification algorithm makes its choices.

As a first step, a training corpus for the algorithm was built, encompassing 36 authors, including Kundera, and comprising 136 texts with a total of 11,478,630 occurrences.[4] In particular, the sub-corpus containing Kundera's work includes all of the author's novels, with the exception of the novel to be classified – *Le Livre du rire et de l'oubli* – amounting to 9 texts and 729,667 occurrences.

After training on the 36 authors, the algorithm was presented with the text earmarked for classification, with the specific task to attribute it to one of the authors on which it had been trained. The aim of this attribution task is not to authenticate the authorship of the text, a matter beyond doubt, but to detect the prototypical lexical, grammatical, and/or lexico-grammatical elements characteristic of Kundera's prose within the novel, using the deconvolution mechanism.

In this study, the algorithm has been trained on grammatical forms, lexical forms, and lemmas. The aim is to detect the simple and complex units (lexical, grammatical, or lexico-grammatical configurations) that enable the algorithm to recognise Kundera's prose in *Le Livre du rire et de l'oubli*: what makes the algorithm identify Kundera as the author of the text to be classified? Which linguistic elements *characterise* the idiolect of our author in relation to other authors of contemporary French literature? In other words, the goal is to identify the textual elements that cause the algorithm to recognize a text as the production of one author rather than another.

In contrast to the previous study that conducted 15 attribution tasks for the analysis of Kundera's essays and novels,[5] this paper extensively presents the results of a single attribution task, namely that of the novel *Le Livre du rire et de l'oubli*. In this way, it will be possible to present a more extensive exploration of the textual data detected by the algorithm. The selection of this particular novel was dictated by the fact that it had one of the highest accuracy rates.[6]

Although this paper exclusively showcases the classification task results for one of Kundera's texts, these results are nevertheless extremely important for the identification of prototypical elements of Kundera's novelistic prose in general.

---

4 Due to space constraints, for the complete list of 136 texts in this reference corpus we refer to Beghini F. (2023). À la recherche de la "pépite d'or". Étude textométrique de l'œuvre de Milan Kundera. Thèse de doctorat, Università degli Studi di Padova, Université Côte d'Azur. In this study, it will also be possible to examine in depth the guiding principles behind the construction of this corpus. Among the studies that we have considered for the compilation process are Biber *et al*. (1998), Rastier (2011), Lebart *et al*. (2019). The 136 novels constitute one of the sub-corpora of the corpus created within the framework of this thesis. This corpus comprises 189 texts, which have been incorporated into a database that is available on the beta version of Hyperbase web. Since these are texts from contemporary literature, the database is not open access for privacy reasons. However, access can be requested by contacting the Logométrie team of the BCL laboratory at the University of the Côte d'Azur.

5 We conducted this work on the collection of short stories, the 4 essays and the 10 novels in Kundera's production.

6 The "accuracy rate" refers to the percentage of accuracy or precision of the classification task performed by the algorithm (see section 2).

Indeed, in *Le Livre du rire et de l'oubli*, the classification algorithm recognised textual units that it had previously learned during the training phase from Kundera's other novels. A linguistic marker need not feature in all novel classifications to be considered prototypical. The linguistic markers identified in classifying one novel correspond to the elements characteristic of Kundera's prose as learned and identified by the algorithm.

The outcomes of this classification task correspond to what are defined as *activation zones* ("zones d'activation", Vanni 2021), i.e., text extracts that were particularly crucial to the classification task due to their containing significantly relevant linguistic markers. These extracts underscore the linguistic markers aligned with grammatical categories, lexical forms, and lemmas forming the basis for classification choices. This paper presents the results of the activation zones considered most significant by the algorithm, i.e., those that are characterised by a higher *TDS*,[7] examining their linguistic markers (lexical forms, lemmas, POS).

These linguistic markers then underwent statistical analysis to assess their significance. More specifically, these textual data were searched for in the corpus containing all of Kundera's novels, alongside the works of the other 35 authors selected for this study.

Finally, the results of this quantitative analysis, i.e., the combination of this new deep learning method with traditional statistical analysis, were examined from a qualitative point of view. This qualitative interpretation involved linguistic exploration that also focused on the aesthetic intention behind stylistic choices.

2. An integrated analysis

The classification algorithm attributed *Le Livre du rire et de l'oubli* (1979) to Kundera with a 72% confidence and an accuracy rate of 83.64%.[8] Our analysis of the linguistic markers derived from the deconvolution mechanism initially focused on simple lexical units and subsequently considered complex lexical and morphosyntactic units.

---

7 TDS stands for *Text Deconvolution Saliency*. It is an indicator of the relevance given by the algorithm to a particular linguistic marker. For more in-depth information, see Vanni, 2021.

8 The meaning of "accuracy rate" was explained in the previous section. The algorithm assigned a 72% probability of the novel being authored by Kundera. In other words, the algorithm suggests a 72% likelihood that *Le Livre du rire et de l'oubli* was written by Kundera. The remaining 28% is distributed across the other 35 novelists of the training corpus, with none of them having a notably higher percentage compared to the rest.

2.1 The lexical level: forms and lemmas

Forms and lemmas emerge as the most prevalent linguistic markers in this classification task.

Firstly, the recurrence of two feminine articles, "la" and "une", is observed in numerous key passages. Deep learning analysis indicates that, on the whole, feminine articles exhibit a higher activation rate or TDS[9] than masculine articles. Statistical analysis shows that these determiners are indeed particularly relevant in Kundera's novels in comparison to other authors, with a standard deviation of 16.4 for "une" and 53.8 for "la".[10]

In the activation zones[11] with the highest TDS, it is observed that the linguistic markers "une" and "la" are often highlighted by the algorithm in conjunction with the term "nostalgie" ("nostalgia"). The calculation of specificities further supports the significance of the forms "nostalgie" and "la nostalgie" (Fig. 1). The subsequent subsection will also present a relevant lexico-grammatical pattern encompassing this particular noun.

9 For a definition of TDS, refer to Section 1.2.

10 As explained in the methodology section (1.2), the deep learning results were explored through statistical analysis to determine their statistical significance. The terms "significance", "standard deviation", "specificities" and "overuse" all refer to statistical calculations aimed at identifying *linguistic specificities*. The *calculation of specificities* is integrated into most textometric software and is based on the hypergeometric model (Lafon 1980). This calculation takes into account the size of the corpus and of the text, word frequency in both the corpus and the text, revealing whether the word's usage is higher or lower than the norm (average usage) within a text. For further discussion, see the Hyperbase manual (Brunet 2011), where the statistical measures used are explained in detail. The purpose of this calculation is to identify the linguistic elements that are not only more frequent, but also have a greater specific significance in a particular text or author than in other texts or authors. The following article is also helpful to learn more about how these features work in the standard and web versions of the Hyperbase software: Vanni & Mittmann 2016.

11 See section 1.2 for the meaning of activation zones.

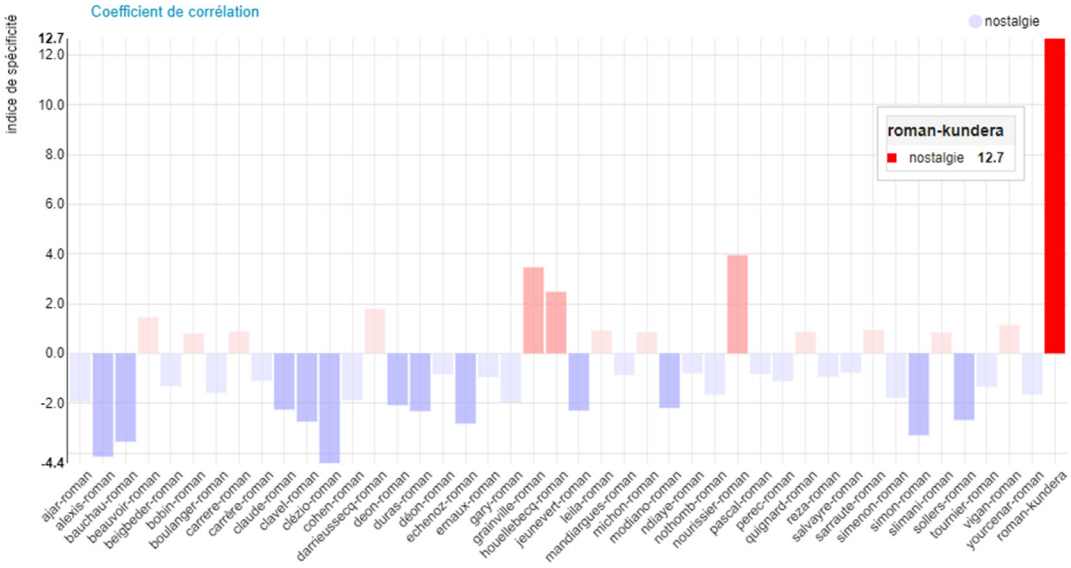

Figure 1: The standard deviation of "nostalgie"

Among the other forms and lemmas, a series of words linked to the realm of *Czech history* and *politics* become discernible: the form and lemma "communiste" ("communist"), the form "socialiste" ("socialist"), "peuple" ("people") – usually followed by "tchèque" ("Czech") – "adversaires" (as in "adversaires politiques", i.e., "political opponents") and "politique" (as in "opposition politique", i.e. "political opposition"), the forms "tchèques" ("Czech"), the form "tribunal" ("court"), pertaining to trials of traitors, and the lemma "allemand" ("armée allemande", i.e., "German army").

In another key passage, the form "étudiant" ("student") emerges, referencing the main character of the novel's fifth part. This figure is never referred to by a proper name but solely by the common noun "étudiant".

Subsequently, the activation zone that follows highlights a series of words that our semantic study associates with the exploration of the theme of "l'âge lyrique" ("the lyric age") or "l'accord catégorique avec l'être" ("the categorical agreement with being"),[12] namely "enfant" (standard deviation of 3.2), "intimité" ("intimacy", 5.1), "hocher la tête" ("to nod one's head", "LEM:hocher la tête",[13] standard deviation of 5.0), and "équipe" ("team", 2.1).

> […] renoncer à son intimité. Elle va avec eux à la salle de bains, quoique le premier jour elle ait refusé de les y accompagner parce qu'PRON:Masc:Sing:3pers lui répugner de faire sa toilette sous leurs regards. La salle de bains, une grande pièce carrelée, est le centre de la vie des enfants et de leurs pensées secrètes. D'un côté il y avait les dix cuvettes des

12 For the definition of the lyric age, lyricism and categorical agreement with being, see the subsequent section, "3. Qualitative analysis".

13 The abbreviation "LEM:" is the Hyperbase (web version) encoding for defining a lemma.

> waters, de l'autre dix lavabos. Il y a toujours une équipe assise sur les waters avec la chemise retroussée, et une autre […].[14]

This activation zone will be revisited in the subsequent subsection (2.2), which delves into the presence of morphosyntactic patterns. The connection between these terms and the theme of lyricism in Kundera's work will also be further explored (sections 2.2. and 3).

Moreover, "amour" ("love") recurs in several activation zones, its significance being statistically confirmed with an exceptional standard deviation of 37.6 (Fig. 2). This significant departure between Kundera and other authors is uncommon since, generally, when calculating specificity, an element is considered significant within a specific author or text if it surpasses the significance threshold of +2 or -2. Other words relating to this thematic core include the nouns "baiser" ("kiss"), "cœur" ("love", in some contexts) and the verbs "aimait" ("he/she loved"), and "faire l'amour" ("to make love").

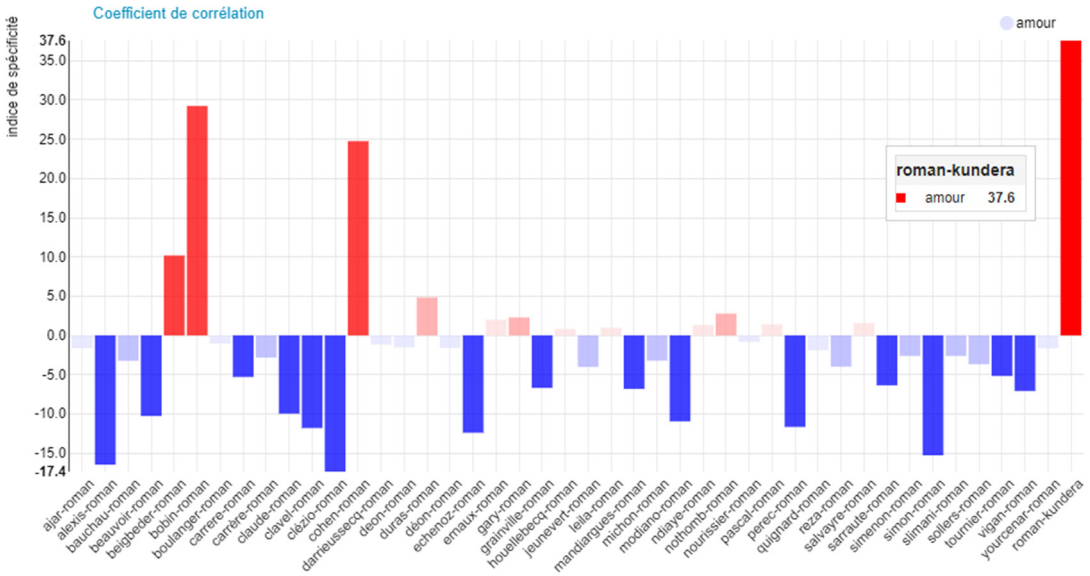

Figure 2: Standard deviation of "amour" in Kundera and the reference corpus

The subsequent subsection will introduce a lexico-grammatical structure in which "amour" is included. Finally, the algorithm extracted the lemma "rire"

---

14 Legend. Green: lemma. Blue: lexical form. Orange: grammatical category and punctuation (grammatical categories are coded according to the codes of the chosen tagger, spaCy). "[...] renoncer à son intimité. Elle va avec eux à la salle de bains, quoique le premier jour elle ait refusé de les y accompagner parce qu'il lui répugnait de faire sa toilette sous leurs regards. La salle de bains, une grande pièce carrelée, est le centre de la vie des enfants et de leurs pensées secrètes. D'un côté il y a les dix cuvettes des waters, de l'autre dix lavabos. Il y a toujours une équipe assise sur les waters avec la chemise retroussée, et une autre [...]" "She goes to the bathroom with them, though on that first day she had refused to accompany them there because it repelled her to wash herself with them looking on. The large, tiled bathroom is at the center of the children's lives and secret thoughts. On one side are ten toilet bowls, on the other ten washbasins. While one team sits with hitched-up nightshirts on the toilet bowls, another […]" (London: Faber & Faber 1996, trans. Aaron Asher, 240).

("laugh", 8.9) and the forms "insignificance" ("insignificance", 8.4) and "image" ("image", 12.8), which are also statistically relevant, as indicated by their standard deviations in parentheses.

Following the observation of simple lexical units (forms and lemmas), the subsequent subsection will deal with the analysis of lexical and lexico-grammatical patterns evident in the activation zones extracted by the algorithm.

2.2 Lexical and morphosyntactic patterns

The lexical patterns extracted include sequences of words, such as "un éclat de rire" ("a burst of laughter") and "jeune homme" ("young man"). The latter deals with the theme of lyricism, while the former pertains to the theme of laughter, specifically a particular type of laughter: the "rire des diables" ("laughter of the devils"; Kundera, Œ I, 984-985). Notably, "jeune homme" has a standard deviation of 15.19 and it is the most significant "adj. + noun" sequence in Kundera.

Furthermore, the grammatical sequence "adj. + noun" is among the grammatical patterns that emerged from the classification task, with statistical analysis revealing a standard deviation of 12.4 when compared to the reference authors. Another example of a lexico-grammatical pattern corresponding to this grammatical sequence is "une ADJ nostalgie", with a standard deviation of 5.1.

This morphosyntactic sequence also reoccurs in another grammatical pattern, namely [Det. + (Adj) + N1 + (Adj) + Prep + Det. + (Adj) + N2 + (Adj)], which was the subject of our previous study (Beghini, 2022). This pattern in its entirety was also found in *Le Livre du rire et de l'oubli*, not only as a grammatical sequence, but also as a lexico-grammatical sequence, in "la possibilité de l'amour physique" ("the possibility of physical love"), "l'ennemi de l'amour et de la poésie" ("the enemy of love and poetry").

> Car maintenant qu'il était mort, son mari n'avait plus qu'elle, plus qu'elle au monde ! C'est pourquoi, aussitôt qu'elle songeait à la possibilité de l'amour physique avec un autre, l'image de son mari surgissait, et avec elle une lancinante nostalgie et avec la nostalgie une immense envie de pleurer.[15]

[15] "Because now that he was dead, her husband had no one but her, no one but her in the entire world! That is why, the moment she even considered the possibility of physical love with another man, her husband's image suddenly appeared, and with it an agonizing yearning, and with that yearning an immense desire to weep." (London: Faber & Faber 1996, trans. Aaron Asher, 122).

In the next figure, we can see the words highlighted in one of the most relevant activation zones.[16] In addition to the noun phrase "la possibilité de l'amour physique", the phrase "immense envie de pleurer" ("an overwhelming desire to cry") and the term "nostalgie" appear.

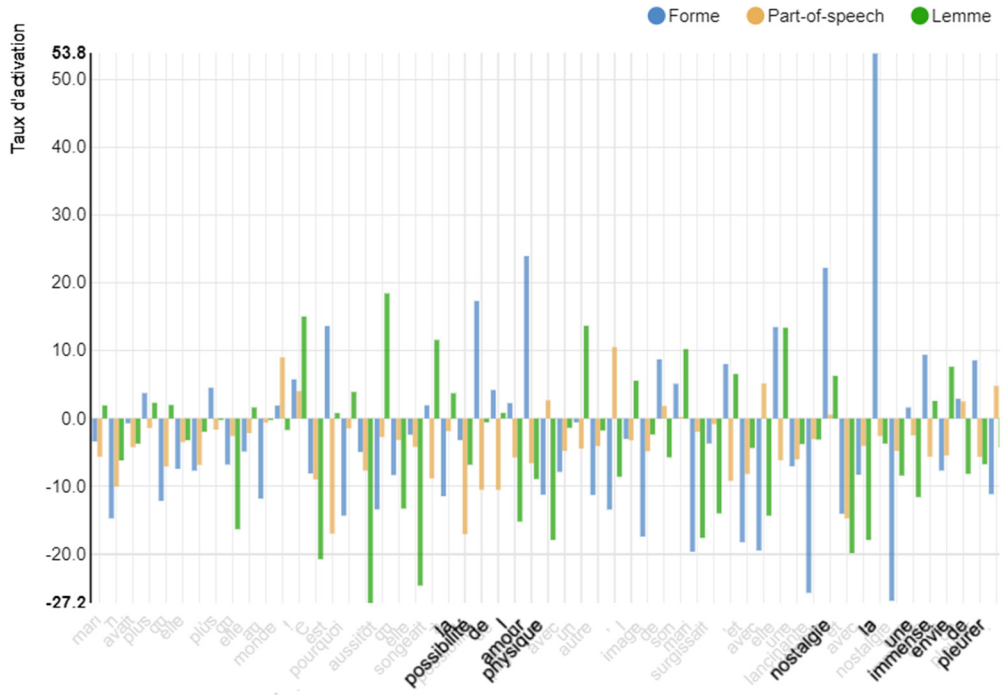

Figure 3: An activation zone in the novel *Le Livre du rire et de l'oubli*

Statistical analysis confirms the relevance of the pattern "Noun + Prep. + Det. + amour" in Kundera's work (Fig. 4), with a remarkable score of 13.6.[17] As for the statistical study of the other highlighted phrase, "immense envie de pleurer", the structure "LEM:envie *** VERB:Inf"[18] was detected, with a standard deviation of 8. The relevance of the word "nostalgie" was previously underscored in the section on lexical forms and lemmas (2.1).

16 As observed earlier, the reference to the most relevant activation zones means the activation zones displaying higher TDS values. The relevance of the different linguistic markers in this activation zone is illustrated in the graph in Figure 3, indicated by the activation rate values ("taux d'activation").

17 The phrase "Prep. + Det. + amour" has an even higher significant score (24.6).

18 In Hyperbase Web, when used in conjunction with the spaCy tagger, asterisks signify the presence of one or more words.

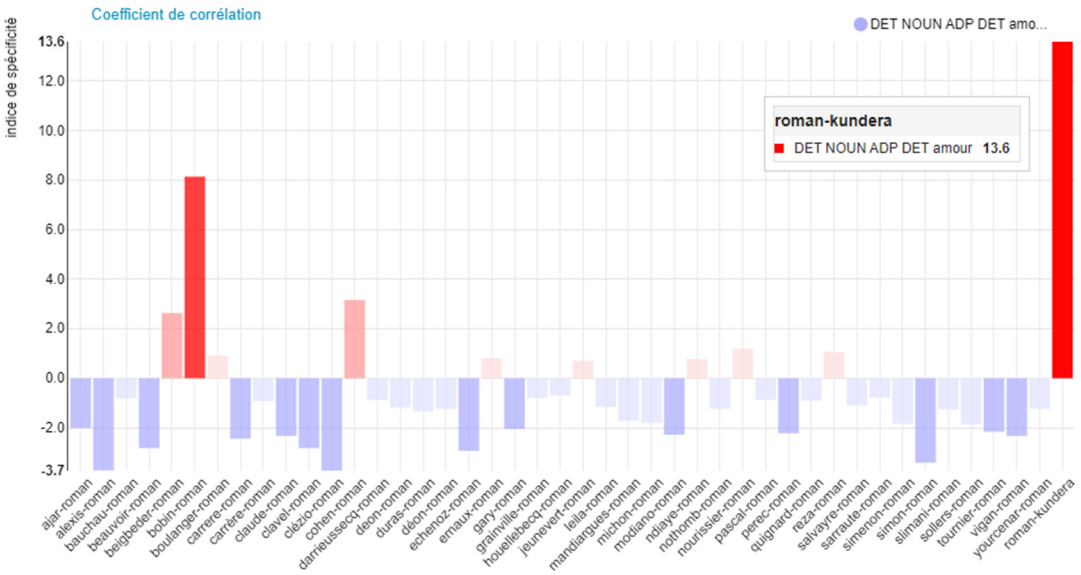

Figure 4: "Det + N_1 + Prep + Det + amour"

Statistical analysis uncovered several similar structures: "l'illusion de l'amour" ("the illusion of love"[19]), "l'anxiété de l'amour" ("the anxiety of love"), "l'enivrement de l'amour" ("the intoxication of love"), "l'absolu de l'amour" ("the absolute of love"), "l'agressivité de l'amour" ("the aggressiveness of love"), "l'éclat de l'amour" ("the brilliance of love"), "la voix de l'amour" ("the voice of love"), "la comédie de l'amour" ("the comedy of love"). Specifically, among these expressions, "l'absolu de l'amour" ("the absolute of love") stands out as the most significant, with a standard deviation of 16.62 compared to the reference authors.

In one of the activation zones of the previous subsection (2.1), the lexico-grammatical pattern "PRON:Masc:Sing:3pers *** LEM:répugner"[20] is highlighted. This sentence pattern has a standard deviation of 2.1 and, among its occurrences, one of the most frequent sequences is "lui LEM:répugner", with a standard deviation of 5.7. When our search was restricted to "lui LEM:répugner", this sequence was usually preceded by a relative pronoun – "*PRON:Rel* lui LEM:répugner" (3.8) – or a personal pronoun – "*PRON:Masc:Sing:3pers* lui LEM:répugner" (2.2). In our statistical analyses, the verb "répugner" stands out significantly in Kundera's work (3.6) compared to the other authors.

Another activation zone contains the pattern "DET ??? où", where the three question marks represent a single word according to the software's indication.

19 The translations in parentheses are literal translations of these noun phrases, they do not belong to the official translation.

20 The term "répugner" means "to repel" or "to be repulsed".

> Elle avait pénétré dans DET:Ind:Masc:Sing:Art monde où il existait une mystérieuse échelle de valeurs qu'elle ne comprenait pas.[21]

The sequence's standard deviation was found to be significant, with a value of 6.9.[22] Further research attempted to assess the relevance of preceding linguistic markers, but statistical evidence ruled this out. However, if we consider the corresponding grammatical category instead of the form "penetré", the resulting sentence pattern becomes "VERB ???? DET ??? où", with a standard deviation of 5.5. Substituting the lemma "où" with its grammatical category yields the morphosyntactic pattern "VERB ???? DET ??? PRON:Rel", showing a standard deviation of 6.5. Statistical research also affirmed the significance of the noun phrase "DET NOUN PRON:Rel" (6.4), as shown by Fig. 5.

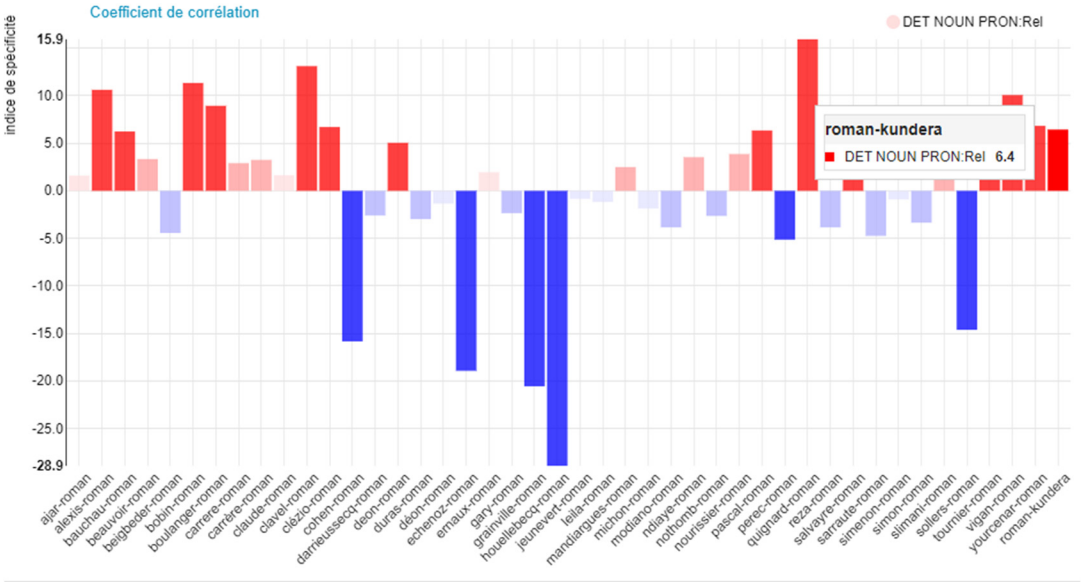

Figure 5: Det. + N + Pron. rel.

In summary, this analysis identified a lexico-grammatical pattern, "DET ??? où" (6.9). Further analysis revealed the salience of the grammatical patterns "(VERB ???) DET ???? PRON:Rel" (6.5) and "DET NOUN PRON:Rel" (6.4).

3. Qualitative Analysis

In this section, we provide a summary of our integrated analysis and offer a more detailed presentation of our qualitative analysis. We focus primarily on the lexicon and subsequently delve into the lexical, lexico-grammatical, and grammatical patterns.

---

21 "She had entered into a world with a mysterious scale of values she did not understand." (London: Faber & Faber 1996, trans. Aaron Asher, 174).

22 As explained in section 1.2, the statistical calculations conducted in this study are focused on exploring specificities and are based on the hypergeometric model.

The forms and lemmas detected by the classification algorithm and subsequently substantiated by statistical analysis can be associated with some of the central themes of Kundera's work.

However, prior to exploring the themes related to these terms, it is worth taking into account the salience of *feminine articles*. The overuse of feminine articles most likely correlates to a tendency towards the use of abstract language. These quantitative data can therefore be explained by the presence of discourses that encompass abstract, existential, or philosophical dimensions. In French, nouns denoting abstract concepts are frequently feminine (Brunet 1988: 193).

As far as *lexical terms* are concerned, our quantitative analysis unveiled common nouns employed to designate characters whose proper names remain unmentioned. For example, the student in the fifth part of the novel is never designated by his proper name, but only by the common nouns "étudiant" or "homme", the latter in the adjectival phrase "jeune homme". A statistical analysis of usage contexts also revealed this tendency in other novels. In Kundera's works, compared to the reference corpus, there is an overuse of common nouns denoting a character based on their occupation (e.g., "le peintre", "the painter" in *La Vie est ailleurs*) or family relationships ("la belle-soeur", "the sister-in-law" in *L'Identité*).

In all likelihood, this stylistic decision was taken because characters in Kundera's novels are leveraged to explore existential themes. The use of proper names can certainly aid in placing them within a context and making them seem real, but our author is not concerned with realism. For the sake of existential exploration, a realistic description is not necessary; any information provided about the characters serves to describe and define their existential code, and often it is not even necessary to know their names.

As for the other forms and lemmas, what themes emerge from a qualitative study of our quantitative analysis of the novel? Our research revealed the themes of *lyricism* or *categorical agreement with being* ("age lyrique" or "accord catégorique avec l'être"*)*, *love*, *laughter* and *insignificance*. The other terms mentioned above, which cannot immediately be traced back to these two macro-themes, belong to Czech history and politics or, more generally, to the sphere of feelings.

Various words adopt distinct meanings in Kundera's context, diverging from their dictionary or common usage definitions. Consequently, a semantic analysis employing interpretive semantics methodologies (Rastier 1987) is useful in defining inherent and afferent semes (Rastier 1987: 43-44.). Thanks to this semantic analysis, it became possible to trace several of the words detected by the algorithm back to particular themes, and this methodological approach holds particular relevance for the theme of lyricism.

An in-depth definition of the theme of the "âge lyrique" goes beyond the scope of this article. For the purposes of the present study, it suffices to state that

Kundera's sense of the theme of lyricism pertains to a proclivity for disregarding fragments of reality that do not conform to one's consciously chosen conceptual framework. The opposite of this existential stance is represented by the anti-lyric age, characterised on the contrary by a tendency towards demystification and critical thinking.[23]

Deep learning analysis highlighted a number of terms that, when contextualised, align with this theme. They pertain to the domains of youth and childhood, sometimes also correlated with inexperience ("enfant", "jeunesse", "jeune(s)", "enfance"). Other terms belong to contexts of use in which reference is made to the opposition between public and private; in particular, the categorical agreement with being and lyricism are associated with a lack of privacy and the loss or dilution of the individual identity in a community ("intimité", "équipe"). "Hocher la tête" is a symbolic gesture characterising the categorical agreement with being, which we find in various passages of the novel (Kundera 2016, Œ I: 989, 1017, 1080, 1083).

Both laughter and insignificance can be considered key concepts for Kundera's entire oeuvre (Kundera 2016, Œ II, Biographie de l'œuvre 1297; Ricard 2003: 63). More precisely, the lemma "rire" can have two different meanings: /rire$_1$/, referred to as "le rire des anges" ("the laughter of the angels"), is aligned with a lyrical attitude, while /rire$_2$/, "le rire des diables" ("the laughter of the devils"), embodies a demystifying, anti-lyric posture.[13]

Finally, the quantitative relevance (via deep learning and statistics) of words linked to the theme of love is highly significant in comparison with our reference authors. Moreover, the word "amour" allows us to transition from simple to complex units since both deep learning and statistics detected a recurrence of the lexico-grammatical patterns "Noun + Prep. + Det. + amour" (standard deviation of 16.6) and "Prep. + Det. + amour" (24.6).

These patterns are variations of the grammatical pattern identified by Beghini (2022), i.e., [Det. + (Adj.) + N1 + (Adj.) + Prep. + Det. + (Adj.) + N + (Adj.)]. The frequency and the significance of this textual element reveal a stylistic preference for secondary predication and a tendency towards nominalisation. This stylistic choice serves dual aesthetic aims: the first is geared towards synthesis and is characterised by precision and clarity, corresponding to a desire to define, i.e., in accordance with the Latin etymon "de-finire", to set limits, to circumscribe a portion of reality with rigour and accuracy. The second tendency is aligned with the desire to defy the limits of the finiteness of a definition and thus to allow a glimpse of the presence of potential afferent semes. Indeed, certain nominal groups acquire new meanings or expand their semantic scope when considered in the context of other passages in Kundera's work (cf. Beghini 2022).

23 For further information, see Kundera (2016, Œ II: 705-707, 716) and Ricard (2003).

In addition, our study found other lexico-grammatical and grammatical variations of this pattern. Specifically, the grammatical configurations "Verb + ? + Det. + ? + Rel. pron." and "Det + N + Rel. pron." emerged, their relevance is supported by statistical analysis. Particularly noteworthy is the expanded nominal group featuring a relative proposition as a noun modifier, bearing particular significance in Kundera's works (6.5).

This expanded nominal group constitutes the second type of expanded noun phrase identified in our analyses, with the first being "Det. + N1 + Prep. + Det. + N2". Additionally, the "adj. + noun" noun phrase recurs. The prominence of these nominal structures is heightened by a statistical exploration that indicates a dearth of the grammatical category of nouns in Kundera's works compared to the reference authors.

Finally, another lexico-grammatical pattern was observed: "PRON:Rel lui LEM:répugner" (3.8). Our qualitative research associated this recurrent pattern with the theme of lyricism. Indeed, the passages in which this lexico-grammatical sequence emerges represent a character displaying repulsion, often directed at another character or a collective marked by a lyrical disposition. In other circumstances, it indicates a character's reaction to a "laideur esthétique" ("aesthetic ugliness"), or a "laideur physique" ("physical ugliness"). More precisely, in the latter case, this repulsion either needs contextualisation within the portrayal of recurring existential reflections on the interplay between body and soul (Kundera, Œ II, 81-83; Beghini 2023: 281-288), or it reflects aversion towards a form of lyrical stance (as in the case of Sabina when she speaks of the "laideur esthétique" of the communist regime, Kundera, Œ I: 1341).

Finally, as regards lexical patterns, the most noteworthy ones bring us back to the themes mentioned earlier in this section: "jeune homme" relates to lyricism and "l'absolu de l'amour" to the themes of love and eroticism.

4. Conclusion

The aim of our study was to identify a number of prototypical elements in Kundera's prose by analysing the results of a deep learning classification task applied to one of Kundera's novels. To achieve this, an integrated methodology was adopted, combining a qualitative analysis through deep learning with traditional statistical approaches.

The deep learning deconvolution mechanism highlighted the textual elements that set Kundera's prose apart from that of contemporary authors and could therefore be considered characteristic of his idiolect. The linguistic markers extracted by the algorithm and validated through statistical analysis enabled us to identify forms, lemmas, as well as lexical, grammatical, and lexico-grammatical patterns.

More specifically, through the study of forms and lemmas, we identified a propensity for abstract language and, in particular, an inclination towards the exploration of certain existential themes. Furthermore, through the integrated analysis of forms, lemmas, and/or grammatical categories, we gained insight into the interpretation of certain lexical, grammatical, and lexico-grammatical patterns that correspond to nominal structures, offering a qualitative perspective that aims to define their aesthetic intent. Consequently, this approach facilitated our investigation into a realm that marks one of the new frontiers of textometry, namely the analysis of complex units.

In future studies, we will also investigate the classifications of other narrative texts and essays to identify further textual elements, both simple and complex units, that are prototypical of Kundera's writing.

BIBLIOGRAPHY

Beghini F. (2022): Stylométrie, ADT et deep learning. Une étude de cas sur la prose romanesque de Milan Kundera. JADT 2022, Naples (Vadistat Press).

Beghini F. (2023): À la recherche de la "pépite d'or". Étude textométrique de l'œuvre de Milan Kundera. Thèse de doctorat, Padoue, Nice (Università degli Studi di Padova, Université Côte d'Azur).

Biber D., Conrad S. & Reppen R. (1998), Corpus linguistics, Investigating Language, Structure and Use, Cambridge (Cambridge Approaches to Linguistics).

Brunet, É. (1988): Le vocabulaire de Victor Hugo. Paris-Genève (Slatkine-Champion).

Brunet, É. (2011): HYPERBASE. Nice (Université Côte d'Azur). http://ancilla.unice.fr/bases/manuel.pdf (19.08.2023).

Holmes, D.I. (1998): The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3), 111-117.

Kalchbrenner, N. *et al*. (2014): A Convolutional Neural Network for Modelling Sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 1, 655-665.

Kim, Y. (2014): Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.

Kundera M. (1996): The Book of Laughter and Forgetting. London (Faber and Faber).

Kundera M. (2016, 2020): Œuvre I et II. Paris (Gallimard).

Lafon, P. (1980): Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1, 127-165.

Lebart, L., Pincemin, B. & Poudat, C. (2019): Analyse des données textuelles. Québec (PUQ).

Legallois, D. (2018): Les motifs lexico-grammaticaux: une nouvelle approche en stylistique. In Stylistique et méthode. Quels paliers de pertinence textuelle? Lyon (PUL).

Longrée, D. & Mellet, S. (2013): Le motif : une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours. Langages, 189, 65-79.

Longrée, D., Luong X. & Mellet S. (2008): Les motifs : un outil pour la caractérisation topologique des textes. In S. Heiden & B. Pincemin (eds), JADT 2008, Lyon (Presses de l'ENS).

Magri V. (2010): Stylistique et statistiques. Le corpus textuel et Hyperbase. In J. Wulf & L. Bougault (eds), Stylistique? Rennes (Presses universitaires de Rennes).

Magri, V. (2020): La linguistique et le nombre. Le Français Moderne – Revue de linguistique Française, CILF (conseil international de la langue française).

Mayaffre, D., Pincemin, B. & Poudat, C. (2019): Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours. Langue française, 203(3), 101-115.

Mayaffre, D. & Vanni, L. (eds) (2021): L'Intelligence artificielle des textes. Des algorithmes à l'interprétation. Paris (Honoré Champion).

Pincemin B. (2012): Sémantique interprétative et textométrie, Texto ! Textes et Cultures, 17(3), http://www.revue-texto.net/index.php?id=3049 (19.08.2023).

Pincemin, B. (2020): La textométrie en question. Le Français Moderne – Revue de linguistique Française, CILF (conseil international de la langue française). https://shs.hal.science/halshs-02902088 (14.08.2023).

Rastier, Fr. (1987, 2009): Sémantique interprétative. Paris (PUF).

Rastier, Fr. (2011): La mesure et le grain. Sémantique de corpus. Paris (Honoré Champion).

Ricard, F. (2003): Le Dernier Après-midi d'Agnès: essai sur l'œuvre de Milan Kundera. Paris (Gallimard).

Savoy, J. (2015): Estimating the Probability of an Authorship Attribution. Journal of the American Society for Information Science and Technology. doi: 10.1002/asi.23455 (16.08.2023).

Thon, V., Vanni, L. & Longrée D. (2022): Le deep learning auxiliaire de l'ADT dans le choix de textes à étiqueter en vue d'un corpus de comparaison". JADT 2022, Naples (Vadistat press).

Tognini-Bonelli, E. (2001): Corpus Linguistics at Work. Amsterdam, Philadelphia (John Benjamins Publishing Company).

Tuzzi, A. & Cortelazzo M. (eds) (2018): Drawing Elena Ferrante's Profile. Padova (Padova University Press).

Vanni L. & A. Mittmann (2016): "Cooccurrences spécifiques et représentations graphiques, le nouveau "Thème" d'Hyperbase", JADT 2016 – Statistical Analysis of Textual Data, Nice, 295-305. https://hal.science/hal-01359413 (14.08.2023).

Vanni L. (2021): De l'analyse statistique de données textuelles aux réseaux de neurones artificiels. Vers des motifs linguistiques profonds, Intelligence artificielle [cs.AI]. Nice (Université Côte d'Azur), https://theses.hal.science/tel-03621264v2 (14.08.2023).

Vanni L., Ducoffe M., Mayaffre, D., Precioso, F., Longrée, D. *et al*. (2018): Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis. 56th Annual Meeting of the Association for Computational Linguistics, Melbourne. https://hal.science/hal-01804310 (14.08.2023).

Vanni L., Guaresi, M. & Magri V. (2022): Convolution et marqueurs multidimensionnels. Description des représentations genrées dans un corpus de films français. 16th International Conference on Statistical Analysis of Textual Data (JADT 2022), July 2022, Naples (Vadistat press).