Analele Universității de Vest din Timișoara. Seria Științe Filologice
2024, Nr. 62/1, pp. 9-25
https://doi.org/10.35923/AUTFil.62-1.02

STYLOMETRY FOR DUMMIES: UNDERSTANDING HOW STYLO PACKAGE WORKS ON ROMANIAN TEXTS

Roxana PATRAS[*]
Lucretia PASCARIU[**]
Alexandra ILIE[***]

[*], [**], [***]"Alexandru Ioan Cuza" University of Iași
[*] roxana.patras@uaic.ro
https://orcid.org/0000-0003-2534-8663
[**] lucretia.pascariu@gmail.com
https://orcid.org/0000-0002-1109-8205
[***] alexandrailie216@gmail.com
https://orcid.org/0009-0008-4284-0653

Stylometry for dummies: understanding how Stylo package works on Romanian texts

In the absence of clear strategies, such as "thematic foregrounding" (a means of highlighting the thematic aspects and of backdropping the stylistic aspects of a text), Stylo Package should be used with care, as a diagnosis means. It is, thus, useful to evaluate, through statistical lens, traditional assumptions on periods, subgenres, authors' co-influencing and so forth, but these results should be credited only as prelims. Beside studies that could replicate on Romanian novel corpora previously proposed workflows, we believe that evaluation of the tool's performance as far as a precise language is concerned (in our case Romanian) should be done by means of multilingual stylometry, which has yielded very nice results for "lesser-resourced" languages too. We entitled the present contribution, "stylometry for dummies", in a self-ironical mood, but also for several other reasons, such as to indicate the experimental level of our approach to both the tool's and the testing material's features, as well as to point at the perplexity that the novice researcher experiences in front of large, easy-to-get and easy-to-discard sets of "results" produced by the Stylo Package. After spending some time with combining features, changing parameters, optimizing processing and other tasks, we believe that beginners should keep a diary in which self-observation while using the tool and progress with understanding the tool's computational premises should mix.

KEYWORDS:

Stylo package; imposters; oppose; style; stylometry.

Does the concept of "style" need new theoretical premises?

Drawing from rhetorical treatises, where elocutio was understood as a strategy of dosing discursive effects in order to persuade the audience, the concept of "style" was, until the beginning of the 20th century, decoded either in a prescriptive key (in the line of Du Marsais, "style is a sum of figures") or in a psychological key (in the line of Buffon, continued by Dilthey, Spitzer, Wellek and Warren, "style is man himself"). Later on, there was a "descriptive" turn that came under the influence of positivism and formalism, although there are studies confirming that it was not until the middle of the last century that a methodology based on measuring the most frequent words was patented (Primorac et al., 2023; Herrmann, Van Dalen-Oskam, & Schöch, 2015). There was, however, a critical insight into the potential of such an approach, manifested in the original but isolated and somewhat eccentric contributions of some scholars like Wicenty Lutoslawski (1898) or Georgy Udny Yule (1939). This is also the case with Romanian philology. Tudor Vianu's initiative to build a dictionary of Eminescu's "language", supported by a pool of occurrences and their contexts (Vianu, 1968), was/is also regarded as eccentric and, to this day, insufficiently exploited. Despite a lack of tradition and specialized research hubs, Romanian computational stylistics has emerged shyly, with several contributions on authorship attribution (Dinu, Popescu & Dinu, 2008; Modoc & Gârdan, 2020, Nitu and Dascălu 2024) and chronolostylometry (Teodorescu & Bolea, 2018). Some of them are based on ready-made packages for stylometry, while others are based on hybrid approaches that combine „handcrafted linguistic features, ranging from surface indices like word frequencies to syntax, semantics, and discourse markers, with contextualized embeddings from a Romanian BERT encoder" (Nitu & Dascălu, 2024).

Based on a quantitative approach to stylistic descriptors, "computational stylistics" (CS) (stylometry) can highlight, through statistical evaluations, the authenticity or authorship of a text (Potthast et al., 2010; Savoy, 2020, p. 3-16). For several reasons, most approaches of this type still problematize the concept of "style":

a. despite the fact that stylometry is not focused on "function words" alone anymore, there is still a strong reliance on the fact that, applied on non-lexical units (function words), stylometry operates with a different concept of style, which needs further specification and analysis;

b. the plethora of tools, variables (features) and methodologies/practices adopted by the expert community often requires an integrative explanation, a motivation and a legitimation en bref;

c. despite the apparent homogeneity, the field of stylometry remains segregated between "advanced stylometry", which focuses on data processing and measurement optimization (Hermann, Jacobs & Piper, 2021), and less advanced —usually pronouncedly empirical and "intuitive"—stylometry, which is concerned, starting from certain theoretical assumptions (periodization, gender signal, etc.), with "profiling" discursive phenomena (Savoy, 2020, p. 10-16; Misini et al., 2022), with observing data behaviors when they enter different statistical scenarios.

One of the most successful DH domains, stylometry is able to bridge various levels of expertise, even if its researcher-friendly and super-inclusive orientation might bring about remarks as the following:

> Despite its clear "empirical" orientation, CS notably includes approaches that emphasize "hermeneutic" processes, i. e. interpretative and subjective dimensions modeled within a computational approach [...]. One of the challenges for the future of CS will be integration of these different epistemic traditions, with clearer attention paid to the relationship between research aims and research methods. Rather than argue for a single best-model of research practices, our vision for the field is deliberately heterodox when it comes to theory and methodology. (Hermann, Jacobs & Piper, 2021, p. 452)

Scholars have also pointed out that the explanations for the success of stylometric results are generally “inadequate”, since they stand on two assumptions (psycholinguistics and variationist sociolinguistics) that cannot be generalized to carefully edited texts:

1. the frequent use of grammatical structures is a matter of subconscious/ unconscious behavior or deeply ingrained linguistic habits (Hoover, 2001; Karsdorp, Kestemont & Riddell, 2022);

2. individual stylistic variations (idiolects) engender from dialectal variations and the relationship between the two is hierarchical.

As a theoretical alternative, the situationist approach assumes that the high frequencies of (functional) words indicate, in fact, an embedded rhetorical relationship between the author and “their” target public, which remains active even when the text is modified to a greater or lesser extent (lemmatized, stemmatized or even translated). The authors’ commitment to addressability implies that the changes of target audience will drive to *register variation*, which is what the stylometric tools detect (rather than idiolects). Regardless of whether the text is original, translated (Rybicki, 2012), or scaled down to stems or lemmas, the striking results of stylometry are seeded in a given communicative context:

> The study of register variation, especially as pursued in the tradition of multidimensional register analysis, provides a theoretical basis for explaining stylometry, both for the application of these methods in general and for the findings of individual studies. In other words, rather than explaining stylometric results as a form of individual dialect variation, […] stylometric results are a form of individual register variation. Stylometric methods do not work primarily because all authors use distinct dialects of the same language — although this may well be the case — but because all authors use slightly different registers of the same language in a given communicative context... [it is because of] this type of individual register variation, as opposed to individual dialect variation, that generally underlies the successful applications of standard stylometric methods for authorship analysis. (Grieve, 2023, p. 52)

For its comprehensiveness and relevance, for its emphasis on the empirical and observational dimension of style analysis, regardless of the type of tools that we used, we propose the following statement as a working definition for the concept of “style”: “Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively” (Herrmann, Van Dalen-Oskam, & Schöch, 2015, p. 44).

Quick options for amateur stylometrists (Stylo package)

Stylometry has developed by nuancing or adapting principles already stated in the 19th century (Lutosławski, 1898): sampling (text samples with the same number of words are comparable), chronology (text samples belong to the same period), and comparison (between text samples of the same kind). There are statistical and computational analysis tools such as RStudio, Signature[1], JGAAP (Java Graphical Authorship Attribution Program)[2], Lexos, StyloR, which have had a significant impact

[1] *The Signature Stylometric System* is a software produced by Peter Millican in 2014, accessed 26 May 2022: https://www.philocomp.net/texts/signature.htm.

[2] JGAAP is a tool that allows even the inexperienced to apply machine learning techniques to establish the authorship

on style research in recent decades (Juola 2008, 2015; McCarthy & O'Sullivan, 2021; Kabala, 2020; Kedwards *et al.*, 2021; Dockray-Miller *et al.*, 2021; Kestemont *et al.*, 2015; Jungmannová & Plecháč 2024; etc.).

For those who are not yet familiar with the Stylo package options, we provide here a brief introduction. As far as the file format and pre-processing techniques are concerned, it is worth mentioning the fact that that the testing material for this study consisted of 185 files compiled from three corpora of novels published during the long 19th century (ELTEC-rom, Hai-Ro, Pop-Lite). The files are available in TEI-xml format. Although the Stylo package also allows the processing of .xml files, we decided to simplify the test materials and convert all files to *.txt* format. Also, being familiar with how these corpora were edited, we tried to prevent casual faults by automatically removing special characters and some typo errors: Ș/ș-cedilla, Ț/ț-cedilla, the varieties of graphic signs used to transcribe, in editable format, the "z" sound (dz, D/d-cedilla, D/d-comma). When starting to use this tool, it is recommended to create a project (e.g. "StyloForDummies") that must contain, in its turn, a folder entitled "corpus", where the files to be analyzed are saved (Eder, Rybicki &Kestemont, 2016). For more complex analyses (e.g. *oppose* or *classify*), sub-directories such as *primary_set*, *secondary_set*, *training_set* should be also created. If the directory "corpus" remains inside another directory, there is the option to change the directory. Nice graphical representations highly depend on renaming files consistently, according to a project protocol. For instance, while for profiling views (the whole collection) we used AuthorName_Title.For detecting subgenre signals, we renamed the files by adding a set of tags (subgenre_AuthorName_Title) that have been found in adjacent resources (*Dicționarul Cronologic al Romanului Românesc*, 2nd edition, 2023; Hai-Ro 2020[3]). However, for even more visually appealingvisualizations, data produced with Stylo package can be exported to Gephi.

Installing Stylo can be easily done on all operating systems (Eder, Rybicki & Kestemont, 2016), with a small exception for MAC, which requires X11 as a supporting tool[4]. The package features and functions are accessible either via a GUI (with options for language, pronoun elimination, MFW threshold, culling, sampling, analysis type, measured distance, view formatting), or from the command line, when the analysis aims at more than just profiling phenomena [*classify()*, *imposters()*, *oppose()*]. For the neophytes who, at the news that the tool is an out-of-the-box solution, are already enthusiastic, it should be noted that a more advanced use of the package requires R skills (Eder, 2018). A mention should be made about default language options: since there is no option for Romanian, we have tried almost everything "closely" related (Latin, Italian or French) and then we opted for the neutral option "Other". Because certain factors such as the chosen distance[5] or the culling method[6] affect the profiling results and then the choice of observation units in subsequent iterations (Papke *et al.* 2023), we eventually decided to apply Classic Delta because it is the most "simple, *intuitive*, and straightforward [emphasis added]" (Karsdorp, Kestemont & Riddell, 2022,

of a text. JGAAP was developed in the Evaluating Variation in Language (EVL) lab at Duquesne University, accessed 26 May 2022: https://github.com/evllabs/JGAAP.

[3] Shorthand "DCRR".

[4] Just to extend the limits of our patient experimenting, we used, comparatively, two OS.

[5] Burrows' Delta, Canberra, Cosine Delta, Manhattan, minmax, nearest neighbor, etc.

[6] E.g. the option to delete frequent verbs such as "a simți"/ *to feel* or "a iubi"/ *to love*, which might be too strong subgenre signals for sentimental novels.

Stylometry and the Voice…) and because it has already produced robust results for texts that are relatively long as in our case study. However, other distances such as Cosine has also been tested on Romanian corpora to evaluate authorship attribution and it did slightly better than Classic Delta, Cosine Delta and Eder's Delta (Schöch 2022). While it is beyond the scope of this introductory article to find the proper stylometric measure for Romanian 19th-century novels and for corelated specific features, we are aware that future contributions will need a stronger focus on the optimization of distances for various literary phenomena that had a historical and localized existence such as literary movements, literary communities, subgenres, and literary periods. However, even for the best-trained literary historians, this is known to be a difficult task that implies clear strategies of corpus modelling and, as shown by Calvo Tello (2021) and Schöch (2023), a "multi-level classification" approach. For instance, some subgenres proved, in test visualizations, more cohesive than others. Sociologically speaking, the clusters might be motivated not only by textual, but also by extratextual phenomena, for instance the sociology of literary success that explains the influence of a VIP author on imitators.

As indicated above, the similarity and dissimilarity among several texts included in a collection (*corpus*), is based on the statistical calculation of word frequency (MFW: most frequent words), more precisely on the similarity of the text's vectors as established by the word frequencies (Eder, Rybicki, & Kestemont, 2016, p. 15). The basic statistical metrics available in the Stylo Package are the following: 1. Cluster Analysis (CA)/ Hierarchical Agglomerative Clustering; 2. Consensus Tree - Bootstrap Analysis; 3. PCA. For functions such as *oppose*, the user can test corresponding metrics and visualizations. Results can be visualized differently (scatterplot, tree diagram, marked areas), but the real challenge comes in the interpretation, which must take into account possible errors or biases arising from the reduction of the data to a "model", as well as mitigation solutions for this unavoidable abstraction processes, for instance focusing only on a small number of MFWs, considered as defining the problem under analysis, or keeping rather than removing some lexical terms in culling operations. Recent studies have shown that gradual abstraction/ simplification of original texts—lemmatization, removing NER, then replacing less frequent words with their more frequent synonyms—is not only effective in terms of processing but also in terms of subgenre association (Sobchuk & Šeļa, 2024).

It has already been noted that for exploratory stages when the researcher wants to have access, at a glance, to the entire data structure of the corpus, CA is preferable (Kaufman & Rousseeuw, 2005; Eder, 2017). The method detects compact clusters consisting of more stylistically similar/solid texts, which can then be constituted into units of observation for more refined analyses. For example, if highlighted in a *horizontal CA tree view*, the solidarity between two authors writing novels of the same subgenre (i.e. *historical*) should be then followed up and checked by performing some kind of sampling or by using other metrics appropriate for smaller data sets (PCA). Basically, the dendrogram concatenates clusters, placing very similar texts on shorter branches (the right side of pics produced with Stylo) and then, as the size of the branches increases, in larger and larger clusters. PCA is a complementary technique that "enables the *intuitive* visualizations of corpora [emphasis added]" (Karsdorp, Kestemont & Riddell, 2022, *Stylometry and the Voice…*). In order to place each object in a two-dimensional geometric space, the simplification of a matrix of x texts having y variables

(MFW) is achieved by reducing the number of variables to only 2, with single values on the x and y axes. Experts consider PCA analysis to be "the prototypical approach to text modelling in stylometry, because we create a much smaller model which we know beforehand will only be an approximation of our original dataset" (Karsdorp, Kestemont & Riddell, 2022, *Stylometry and the Voice…*). Bootstrap analysis results from aggregating the data from a set number of iterations and indicates in how many of these iterations a particular association is evident.

Profiling 4 subgenres in a collection of Romanian novels

Neatly balanced, our test material comprised 185 files, representing novels of various types[7] published roughly in the time span of 1850-1930 and previously categorized as belonging to four prominent subgenres of the popular novel (two-thirds)[8] and to 1 category of unlabeled novels (one-third): a. *hajduk* (51 texts), b. *historical* (28 texts), c. *mystery – crime – sensational[9][mcs]* (32 texts) and d. *sentimental* (22 texts), e. *not-labeled* (52 texts). We considered these classifications entered by previous (traditional) resources such as dictionaries, lexicons or literary histories as *"ground truth" subgenres* (Sobchuk & Šeļa, 2024).

In line with the findings of Burrows (1987, 2002), it is generally accepted that pronouns do not belong with function words and that "removing them typically stabilizes attribution results" (Karsdorp, Kestemont & Riddell, 2022, *Stylometry and the Voice…*). Nevertheless, the culling operation should be performed with great care because:

> Author signals, genre characteristics, narrative perspective, plot elements, and many other aspects confound the definition of stylometric similarity. The task of amplifying desired aspects and controlling for others is thus both a theoretical and an empirical one, and it should not be taken lightly. (Päpcke *et al.*, 2013, p. 292)

Hence, we pretested the effectiveness of auctorial and generic associations by contrasting two analysis scenarios (with pronouns vs. pronouns culled) for the entire collection of 185 novels and then for some of the subgenres appreciated as markedly pronominal (e.g. sentimental). This testing required the compilation of a stopwordlist[10], which helped at "skipping" unwanted words, in our case pronouns (with all forms) and common errors in a list of 5000 MFW previously generated by Stylo.

[7] Defined types are the following: feuilleton – volume; complete – incomplete; c. finished – unfinished.

[8] Subgenre labels were established either by the DCRR 2023 (for all four popular classes) or by extending research results of other thematic corpora (e.g. the case of the hajduk novel).

[9] Until further studies will highlight the differences between these subgenres, we have decided to aggregate the DCRR labels "mysteries", "crime" and "sensational" as "mcs". We chose to integrate crime, mystery and sensational novels under a single label because, in our collection, each of these subgenres is represented by a relatively small number of texts.

[10] 1. We examined the "wordlist.txt" file generated by Stylo and identified the most common typos, spelling and phonological errors that required manual cleaning;

2. We cleaned up the text collection using Python scripts;

3. Extracted short-form pronouns from compact constructions (e.g., "ducându-și" becomes "ducându" + "-și") and generated a new "wordlist.txt" with separate entries for long-form and short-form pronouns;

4. We manually checked the "wordlist.txt" file generated by Stylo by adding a "#" in front of all the words that would be omitted from the analysis. For example, for the pronoun "atâția", which can also be misspelled (without diacritics), we insert "#" in front of all possible forms: "#atâția", "#atatia", "#atâtia";

5. Once the stopword list was ready, we applied the code line *stylo(features = "wordlist_stopwords_fara_pronume_2.txt"*.

Highlighting the stylistic signal for subgenres such as *hajduk*, *sentimental*, *historical* and *mysteries – crime – sensational* (*mcs*) was performed with CA and Bootstrap-Consensus Tree from 100 to 1100 MFW, with iterations at every 200 MFW to empirically observe the different positioning of texts when increasing the number of features. For the 20 entries of N.D. Popescu, a subgenre signal was noticed, namely the separation of hajduk novels (*Iancu Jianu..., Aga Gruia..., Corbea..., Codreanu..., Moartea lui Iancu Jianu..., Radu Anghel..., Tunsu..., Boierii..., Bujor..., Iancu Jianu Haiducul1912..., Iancu Jianu Haiducul...*), historical (*Bătălia..., Manole..., Mircea...*), and war novels (*Santinela..., Sora..., Primul...*). Only one novel by this very prolific author breaks the subgenre intuition, while another does not combine with any of the subgenres (*Radu...*). Similarly, N.G. Rădulescu-Niger's productions appear in two different areas of the graphic representation, which would coincide with the ages of the author's artistic development: a. *Procurorul...* (1888) b. *Fuga..., Măriuca..., Foamete...* (1896) și c. *Romanul unei iubiri* (1920) și *Orfanii...* (1913).

Since the scenario considering the inclusion of pronouns indicates a subgenre signal at 100 MFW, we decided to test the solidarity of the novels labeled in DCRR 2023, also considering a byway comparison with previous genre-theme classifications as "ground-truth" referents (e. g. Hai-Ro). Checking the distances among novels with the same subgenre label implied setting-up several sub-collections for each. Subgenre divisions were explored by keeping under observation the MFW bounder limit at which some associations stabilize, for instance, an anonymous novel constantly viewed in the company of a group of authors or an author. To easily visualize subgenre associations, we inserted in front of the original file name (Ighell_SimionLicinsky.txt) the DCRR label (mcs_Ighell_SimionLicinsky.txt). Consequently, the Hai-Ro corpus, which was built on thematic criteria, contains novels labeled as "hajduk", "mcs" "historical", as well as a set of novels which, not being indexed yet, kept their original file names. The set of labels provided by the dictionary provides a matrix of "expectation and recognition" that has always been the groundwork of genres (Compagnon, 2001, sec. 2). However, by amplifying desired aspects and controlling for others, our approach to subgenres is a user-oriented theoretical frame that coordinates "immanent formal analysis of the individual text with the twin diachronic perspective of the history of forms and the evolution of social life" (Jameson 1981, 92). Thus, in the following (sub)chapters, we will check, comparatively, the situationist and the essentialist approaches to literary genres, by delineating, for the former approach, each subgenre's outreach and by looking for possible prototypical items, for the latter.

*

By keeping an eye on groupings of authors and genres, we started from profiling the whole corpus, without sampling and using the "classical" distance Delta (*see* Patras & Pascariu, 2024, as well as other data and results available at https://github.com/lucretiapascariu/Profiling-Genre-Signals-in-a-collection-of-Romanian-Novels-with-Stylo). Then, we chose a few units of observation (e.g. hajduk genre solidarity), which required narrowing the collection (subgenre sub-directories), sampling some sequences (the reference ground was the size of the shortest text), and applying more sensitive distances (Canberra, Eder, Cosine Delta, minmax) or multivariate analysis (PCA cov., PCA corr.). Because it makes processing easier and the meaning of the words remains the same, we opted for the lemmatization of the collection, although we are aware that the performance of the tool decreases (Savoy, 2020, p. 57-59) due to the fact that it lacks training for Romanian, there are still typo errors in the texts, and because lemmatization, as a preprocessing option, involves the removal of a lot of information. For instance, in the lemmatized sequences, the feminine article “o” transformed into the masculine indefinite article “un” creates a lot of data noise.

The hajduk cluster differentiates itself / stands out very quickly, maintaining its solidarity regardless / whether (or not) we increase the number of features and even if we change the distances. If the proximity of the *hajduk* and *mcs* labels was not surprising, the hajduk genre being branded from its inception as both “national” and “sensational”, we were instead intrigued by the association *mcs-sentimental-historical*, the sentimental novel placing itself as a kind of pivot in the bottom cluster.

*

In order to shed light on these placements, we used two Stylo functions: *oppose ()* and *imposters ()*. The former is appropriated for the enhancement of common/ distinct features between various subgenre pairings and draws from an extensional definition of the concept of “genre”. The latter, following Aristotle’s idea that “genres” must be filled with a substance and that generic labels are ontologically anchored signs (Schaeffer, 2006), draws from an intentional definition of the concept. In a nutshell, we are approaching popular subgenres extensively and intensively, both by looking for their outline (what differentiates them from other genres) and by looking for a prototypical item for each of them. At the same time, our intention is to crosscheck close-reading and distant reading results, that is, the manual attribution of dictionary labels and the output of statistical analysis.

The tests with the *oppose ()* function have shown the homogeneity of *historical* and *sentimental* clusters and the relative heterogeneity of *hajduk* and *mcs* clusters. The lists of markers and antimarkers for the *mcs-sentimental* pair contains a majority of non-lexical words (adverbs, prepositions, conjunctions, etc.) and pronouns: „d”, „precum”, „vr”, „îndată”, „său”, „di”, „ei”, „acolo”, „s”, „trei”, „patru”, „foarte”, „dupe”, „despre”, „unu”, „ba”, „qua”[ca], „multu”, „quât”[cât], „aici”, „dinaintea”, „daqua” [dacă]. However, after a careful browsing of these lists, we realized that the organization of words in several categories could provide a better semantic insight: 1. Time; 2. Place; 3. Persons; 4. Body related; 4. Abstract notions; 5. Verbs; 6. Objects; 7. Others. With a view to a deeper exploration of tendencies, we compiled 3 tables showing comparatively historical vs. sentimental, historical vs. mystery_crime_sensational, and hajduk vs. mystery_crime_sensational. Accordingly, we noticed the preference, in synonymic pairs, for the Slavic etymon (*vreme-timp*, *glas-voce*, *ceas-oră*), for the abstract nouns *fire-death* (*foc-moarte*) in *historical-hajduk* and for the abstract nouns *air-life* (*aer-viață*) for *mcs-sentimental*; we also noticed the avoidance of female persons, indoor spaces and emotion-cognition verbs in *historical* and *hajduk*.

historical-sentimental	Preferred	Avoided
Time	vreme [time], clipa [moment], ceas [hour], noapte [night], acuma [now]	timp [time], ora [hour], minut [minute], niciodată [never]
Place	cetate [citadel], pământ [land], curte [yard], deal [hill], vale [valley],	camera [room], București [Bucharest], salon [salon]

	loc [place], zid [wall], sat [village], Moldova, țara [country], aici [here], sus [up]	
Person (+pronouns)	vodă [voivode], bătrân [elder], dușman [enemy], popor [people], căpitan [captain], oștile [armies], călugăr [monk], părinte [parent], fiu [son], călăreț [rider], oștean [soldier], neam [people], Ștefan, român [Romanian], său [his/ hers], dânșie [him/ her]	femeie [woman], Elena, doamna [lady], bărbat [man], însa [her], o [a], Zoe
Person-Related (+person adjectives)	sânge [blood], picior [leg], trup [body], domnesc [lordly]	drag [lovely], fericit [happy], trist [sad]
Abstract Notions (+non-person adjectives)	moarte [death], parte [part], veste [news], lupta [fight], ajutor [help], fuga [flight], porunca [command], greu [difficult], **foc [fire]**	idee [idea], amor [love], fericire [happiness], ființa [being], frumos [beautiful], dulce [sweet], natura [nature], nimic [nothing], lume [people], iubire [love], sigur [secure], vis [dream], adevărat [true], plăcere [pleasure], fel [form], rău [evil], trai [lifestyle], viața [live], vorba [speeches], suflet [soul], dragoste [lov], **aer [air]**
Verbs	striga [to shout], porni [to start], sosi [to arrive], aduna [to gather], afla [to find out], pune [to put], trage [to pull], trimite [to send], apuca [to grab], purta [to carry]	iubi [to love], părea [to seem], crede [to believe], simți [to fell], scrie [to write], gândi [to think], uita [to forget], privi [to look], vorbi [to talk], suferi [to suffer], râde [to laugh], citi [to read], ști [to know], spune [to say], plăcea [to like], deveni [to become]
Objects	Cal [horse], arma [weapon]	floare [flower], lucru [thing]
Other	Peste [over], asupra [on], dela [from], către/ cătră [to], până [to], dintre [between], spre [to], vr- Trei [three], patru [four]	nu [no], d- [from], na, așa [like that], d [from],

historical-mcs	Preferred	Avoided
Time	clipă [moment], apoi [than], acuma [know], vreme [time], când [when], odată [once]	minut [minute], moment [moment], ora [hour], data [date], an [year], seara [evening], timp [time]
Place	Țara [county], cetate [citadel], sub [under], pământ [land], vale [valley], Moldova, deasupra [above], cale [way], mijloc [middle]	aci [here], casa [house], camera [room], afară [out]
Person (+pronouns)	dușman [enemy], vodă [voivode], popor [people], neam [people], fiecare [each], oștean [soldier], oștile [armies], român [Romanian], bătrân [elder], căpitan [captain], călăreț	ei [hers], femeie [woman], o [an], cineva [someone], persoana [the person], familie [family], amic [friend], acesta [this], vre-

	[rider], dânșie [him/ her]	[some], cine [who], însa [but]
Person-Related (+person adjectives)	glas [voice], fața [face], chip [countenance], trup [body], frunte [forehead], braț [arm], sânge [blood], fire [hairs], picior [leg], domnesc [lordly]	voce [voice]
Abstract Notions (+non-person adjectives)	luptă [struggle], lung [long], adânc [deep], moarte [death], putere [power], veste [news], alb [white], gând [thought], suflet [soul], greu [difficult], întreg [whole], nou [knew], sânt [saint], nădejde [hope], grai [speaking/ language/ dialect], **foc [fire]**	fel [kind], interes [interest], foarte [very], bine [well], sigur [secure], amor [love], mult [a lot], idee [idea], **aer [air]**
Verbs	porni [to start], aduna [to gather], ridica [to lift], ieși [to exist], cuprinde [to contain]	găsi [to find], întreba [to ask], spune [to tell], înțelege [to understand], observa [to observe], crede [to believe], iubi [to love]
Objects	cal [horse], arma [weapon]	lucru [thing], trăsură [carriage], ban [money]
Other	dela [from], până [to], printer [amoung], peste [over], in [in], spre [towards], iar [and]	dupe [after], ceva [something]

hajduk-mcs	Preferred	Avoided
Time	vreme [time], atunci [then], când [when], atunci [then]	ora [the hour], moment [moment], seara [the night], cândva [sometime], minut [minute], puțin [a little], oră [hour]
Place	țara [country], pădure [forest], pământ [land], codru [forest], drum [road], sus [up], loc [place], divan [sofa], mijloc [middle], sat [village], județ [county]	ușa [door], aici [here], camera [room]
Person (+pronouns)	haiduc [hajduk], Jianu, căpitan [captain], Iancu, boer [boyar], vodă [voivode], Tudor, român [Romanian], grec [Greek], ceata [gang], hoț [thief], frate [brother], tâlhar [bandit], țăran [peasant], neam [people], voinic [sturdy/ lad], om [human], Vladimirescu, poterași [policemen], l- [him], ast- [this], câte- [some]	femeie [women], cineva [someone], amic [friend], doamna [lady], dânsa [her], domn [mister], vreun [any], persoana [person], familie [family], mama [mother], acesta [this], voi [you]
Person-Related (+person adjectives)	chip [countenance], mâna [hand]	privire [glance/ look]
Abstract Notions (+non-person adjectives)	stăpânire [rule], drept [low], moarte [death], **foc [fire]**	lucru [thing], fericire [hapiness], plăcere [pleasure], interes [interest], puțin [a little], idee [idea], ceva [something], **aer [air]**
Verbs	eși [to exit], striga [to shout], ridica [to raise]	ști [to know], observa [to observe], întreba [to ask], aștepta [to wait], părea [to seem], crede [to believe],

		privy [to look], repeat [to repeat], vorbi [to speak]
Objects	cal [horse], arma [weapon]	
Other	dela [from], peste [over], spre [to], in [in], ci [but], cât [as much], iar [again], vre-[any-], ași [would]	după/ dupe [after], dinaintea [before], qua/ca [like], quât/cât [as much], asupra [on], vreun [some], câtva [somehow]

These results might also confirm that genres defined by their plots or settings (like hajduk and historical) may provide a clearer thematic signal than genres defined by their target audience or evoked emotions (like mystery-crime-senzational and sentimental) (Sobchuk & Šeļa, 2024).

The second function (General Imposters) is based on the intuition that we should not evaluate static vocabulary features (MFW), but rather evaluate whether one pair of documents is significantly closer than another from the same text pool (Eder, 2018). Based on a few iterations, we contrast the tested text, on the one hand, and:

a. a group of texts written by possible authors;

b. a selection of "imposters", i.e. authors who could not have written the text in evaluation. Accordingly, each candidate is given a score between 0 and 1; a score above 0.5 may suggest that the authorship check for a particular author was successful. Unlike the classical analysis performed above, in which we generated overview visualizations of the complete collection, this fine-tuning step requires knowledge of R.

We applied *imposters ()* to the four subgenres, taking as a reference point a "prototype" novel for each. For the present analysis we dodged issues regarding correct chronology and critical reception regarding specific cases, so the prototype candidates were decided upon by simply comparing various PCA results (from 100 to 1,100) for samples of 1,000 words (distance Delta). For each subgenre under analysis, we could notice clustering trends (toward the center of the PCA): *historical* novels cluster quite harmoniously between 100-1,100 MFW; *sentimental* novels group between 900-1,100 MFW; *mystery-crime-sensational* novel agglomerate at 900 MFW; and the *hajduk* novels at 300-500 MFW. Of these, the most heterogeneous seems to be the hajduk cluster. Since we suspected that the overall impression might have been affected by our choice of a more comprehensive hajduk corpus — numbering 46 novels, of which 28 novels are labelled in the DCRR as *hajduk*, 15 novels not yet included in the DCRR and 5 novels are labeled in DCRR as either *mcs* or *historical*—we ran several views only for the DCRR *group*. Surprisingly, the PCA is very scattered as well, showing several prominent groups, but not the clear solidarity of the other subgenres. Accordingly, we settled on the following prototype items:

a) *Iancu Jianu Haiducul* (1912) by N. D. Popescu as *hajduk* prototype;

b) *Femeie fără suflet* by Anonymous as *mcs* prototype;

c) *Cu paloșul* by Radu Rosetti as *historical* prototype;

d) *Elena* by Dimitrie Bolintineanu as *sentimental* prototype.

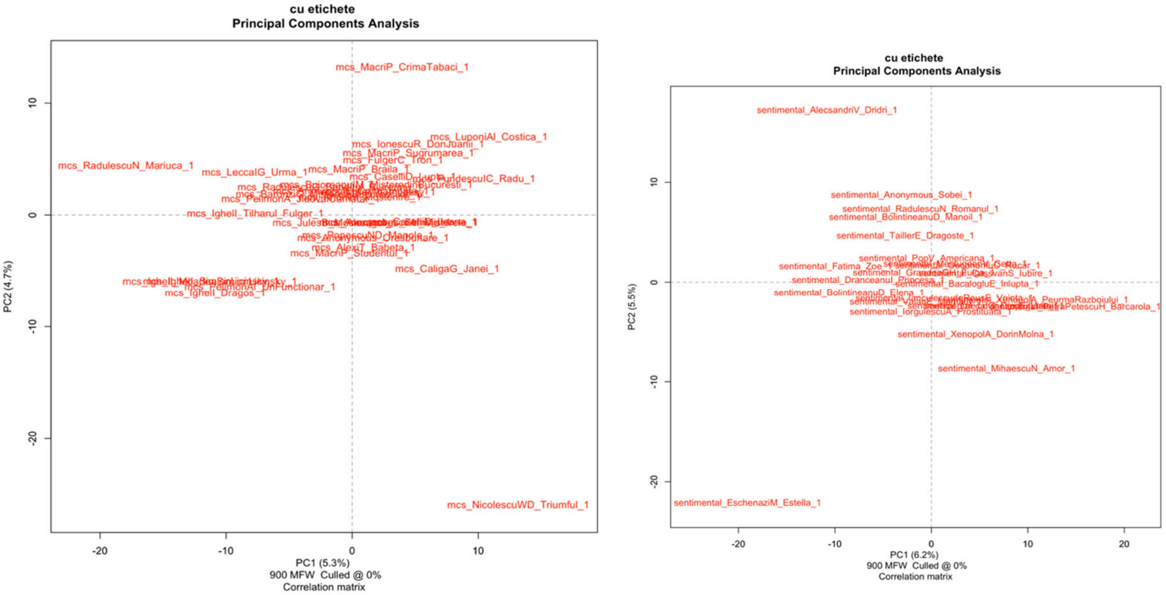
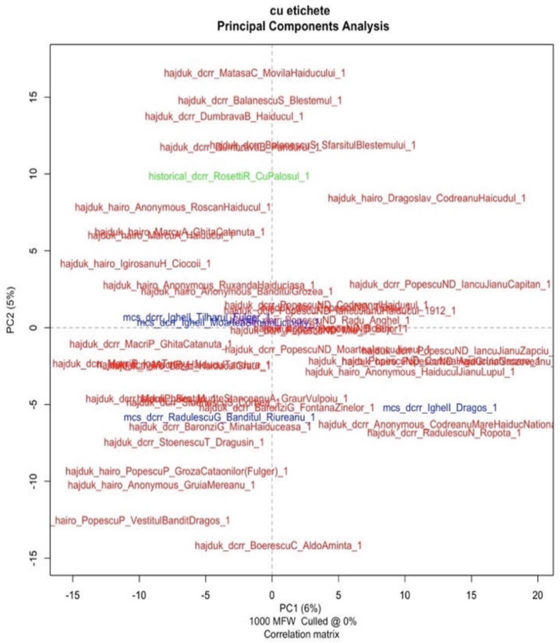
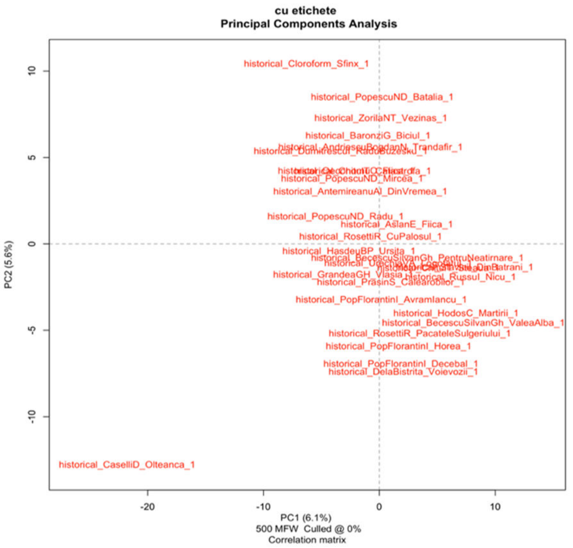


Figure 1: Prototypical item for hajduk, historical, mcs, and sentimental

We have excluded novels by these authors from the tested subcollections, thus the analyses ran as follows:



Test 1: *Iancu Jianu Haiducul* vs. 34 "imposters"

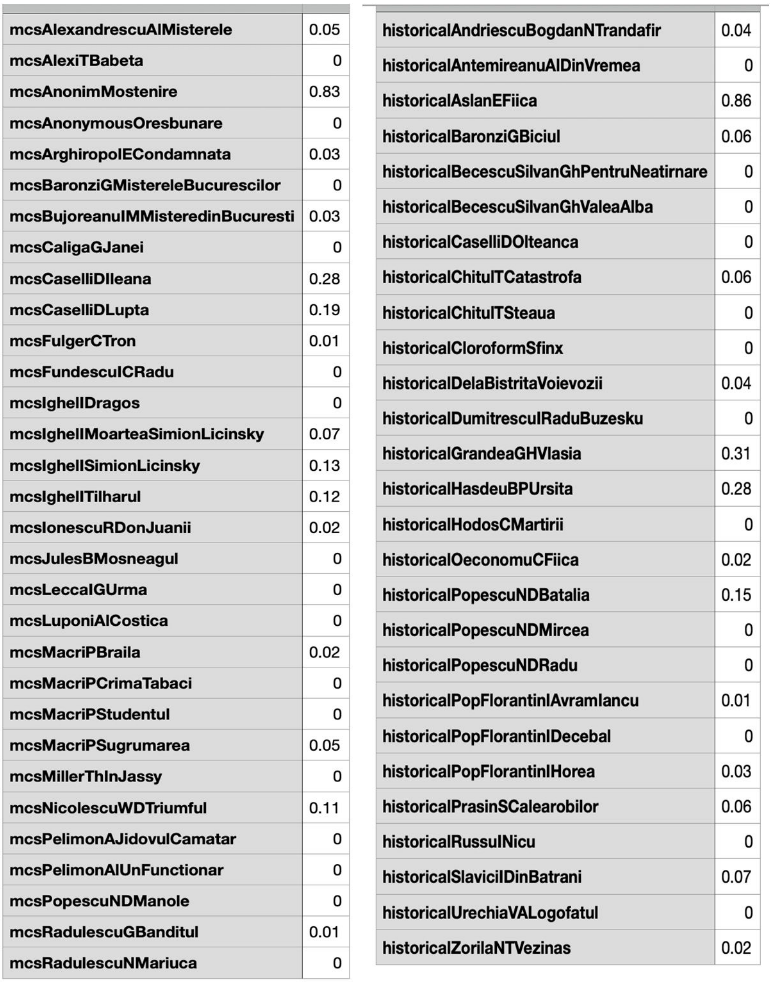


Test 2: Cu Paloșul vs. 27 "imposters"

Test 3: *Femeie fără suflet* vs. 31 “imposters”

Test 4: *Elena* vs. 22 “imposters”

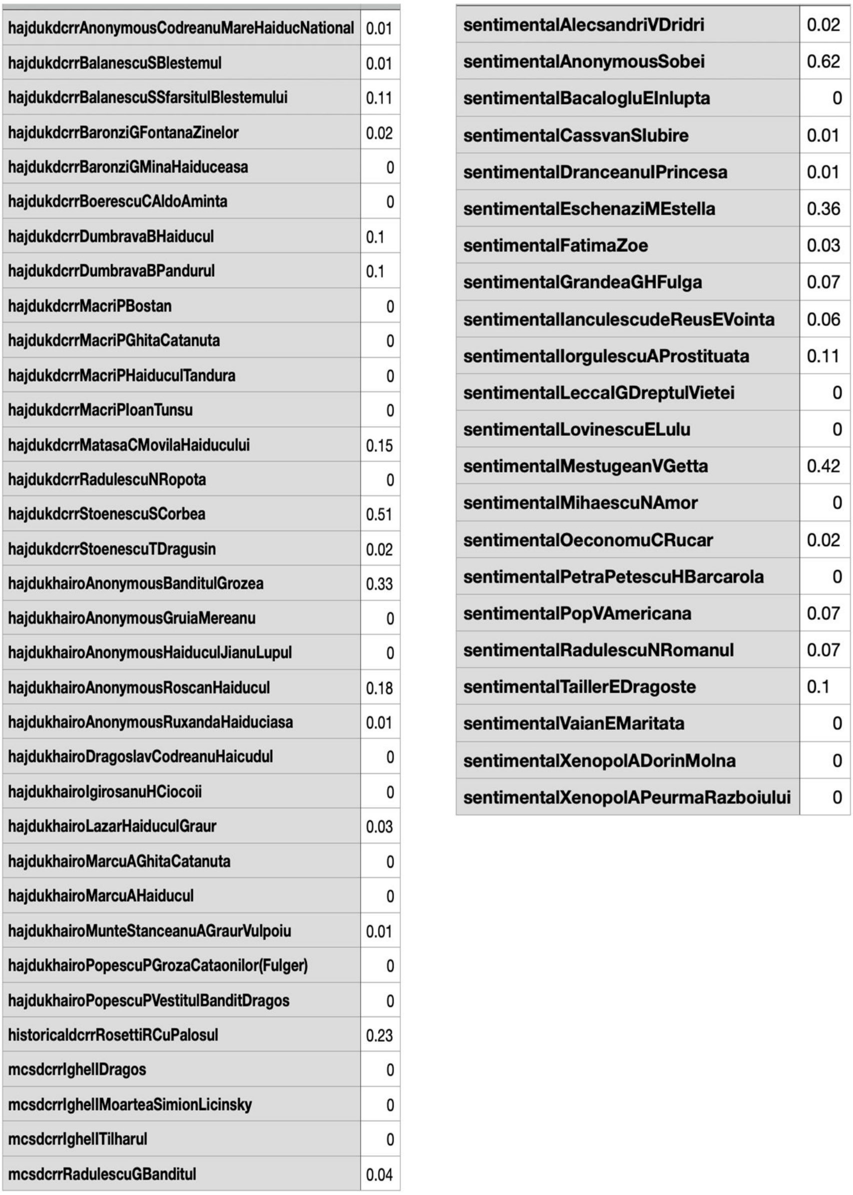
Generally used for authorship attribution rather than for genre modeling, the *imposters ()* method was useful, in this specific case, to profile a symptomatology of imitation within the boundaries of a subgenre. Even if some of the scores are over 0.5 and therefore, they look fit for in-depth authorship attribution, we chose to keep an eye on all non-null values returned for 100 MFW. Test 1 (*hajduk*) returned 16 nonnull scores (47.05%), Test 2 (*historical*) returned 14 non-null scores (51.85%), Test 3 (*mcs*) returned 15 nonnull scores (48.38%), and Test 4 (*sentimental*) returned 14 nonnull scores (63.63%). These preliminary results indicate the relative influence of an item in a group that is sharing the same label, Bolintineanu’s *Elena* (published in 1862) showing as a prototypical novel for the sentimental subgenre. With all defendable reservations regarding the choice of the prototype item—which might have been different, if the publication chronology was observed—the other results suggest two dynamics: one of imitation, the other of distancing from the presumed prototype. Moreover, for two subgenres (*mcs* and *historical*) the values peak up to 0.8, while for *hajduk* and *sentimental* only to 0.5-0.6.



mcsAlexandrescuAlMisterele	0.05
mcsAlexiTBabeta	0
mcsAnonimMostenire	0.83
mcsAnonymousOresbunare	0
mcsArghiropolECondamnata	0.03
mcsBaronziGMistereleBucurescilor	0
mcsBujoreanuIMMisteredinBucuresti	0.03
mcsCaligaGJanei	0
mcsCaselliDIleana	0.28
mcsCaselliDLupta	0.19
mcsFulgerCTron	0.01
mcsFundescuICRadu	0
mcsIghelIDragos	0
mcsIghelIMoarteaSimionLicinsky	0.07
mcsIghelISimionLicinsky	0.13
mcsIghelITilharul	0.12
mcsIonescuRDonJuanii	0.02
mcsJulesBMosneagul	0
mcsLeccaIGUrma	0
mcsLuponiAlCostica	0
mcsMacriPBraila	0.02
mcsMacriPCrimaTabaci	0
mcsMacriPStudentul	0
mcsMacriPSugrumarea	0.05
mcsMillerThInJassy	0
mcsNicolescuWDTriumful	0.11
mcsPelimonAJidovulCamatar	0
mcsPelimonAlUnFunctionar	0
mcsPopescuNDManole	0
mcsRadulescuGBanditul	0.01
mcsRadulescuNMariuca	0

historicalAndriescuBogdanNTrandafir	0.04
historicalAntemireanuAlDinVremea	0
historicalAslanEFiica	0.86
historicalBaronziGBiciul	0.06
historicalBecescuSilvanGhPentruNeatirnare	0
historicalBecescuSilvanGhValeaAlba	0
historicalCaselliDOlteanca	0
historicalChituITCatastrofa	0.06
historicalChituITSteaua	0
historicalCloroformSfinx	0
historicalDelaBistritaVoievozii	0.04
historicalDumitrescuIRaduBuzesku	0
historicalGrandeaGHVlasia	0.31
historicalHasdeuBPUrsita	0.28
historicalHodosCMartirii	0
historicalOeconomuCFiica	0.02
historicalPopescuNDBatalia	0.15
historicalPopescuNDMircea	0
historicalPopescuNDRadu	0
historicalPopFlorantinIAvramIancu	0.01
historicalPopFlorantinIDecebal	0
historicalPopFlorantinIHorea	0.03
historicalPrasinSCalearobilor	0.06
historicalRussuINicu	0
historicalSlaviciIDinBatrani	0.07
historicalUrechiaVALogofatul	0
historicalZorilaNTVezinas	0.02

Figure 2: Imposters scores for mcs and historical



hajdukdcrrAnonymousCodreanuMareHaiducNational	0.01
hajdukdcrrBalanescuSBlestemul	0.01
hajdukdcrrBalanescuSSfarsitulBlestemului	0.11
hajdukdcrrBaronziGFontanaZinelor	0.02
hajdukdcrrBaronziGMinaHaiduceasa	0
hajdukdcrrBoerescuCAldoAminta	0
hajdukdcrrDumbravaBHaiducul	0.1
hajdukdcrrDumbravaBPandurul	0.1
hajdukdcrrMacriPBostan	0
hajdukdcrrMacriPGhitaCatanuta	0
hajdukdcrrMacriPHaiduculTandura	0
hajdukdcrrMacriPIoanTunsu	0
hajdukdcrrMatasaCMovilaHaiducului	0.15
hajdukdcrrRadulescuNRopota	0
hajdukdcrrStoenescuSCorbea	0.51
hajdukdcrrStoenescuTDragusin	0.02
hajdukhairoAnonymousBanditulGrozea	0.33
hajdukhairoAnonymousGruiaMereanu	0
hajdukhairoAnonymousHaiduculJianuLupul	0
hajdukhairoAnonymousRoscanHaiducul	0.18
hajdukhairoAnonymousRuxandaHaiduciasa	0.01
hajdukhairoDragoslavCodreanuHaicudul	0
hajdukhairoIgirosanuHCiocoii	0
hajdukhairoLazarHaiduculGraur	0.03
hajdukhairoMarcuAGhitaCatanuta	0
hajdukhairoMarcuAHaiducul	0
hajdukhairoMunteStanceanuAGraurVulpoiu	0.01
hajdukhairoPopescuPGrozaCataonilor(Fulger)	0
hajdukhairoPopescuPVestitulBanditDragos	0
historicaldcrrRosettiRCuPalosul	0.23
mcsdcrrIghelIDragos	0
mcsdcrrIghelIMoarteaSimionLicinsky	0
mcsdcrrIghelITilharul	0
mcsdcrrRadulescuGBanditul	0.04

sentimentalAlecsandriVDridri	0.02
sentimentalAnonymousSobei	0.62
sentimentalBacalogluEInlupta	0
sentimentalCassvanSIubire	0.01
sentimentalDranceanuIPrincesa	0.01
sentimentalEschenaziMEstella	0.36
sentimentalFatimaZoe	0.03
sentimentalGrandeaGHFulga	0.07
sentimentalIanculescudeReusEVointa	0.06
sentimentalIorgulescuAProstituata	0.11
sentimentalLeccaIGDreptulVietei	0
sentimentalLovinescuELulu	0
sentimentalMestugeanVGetta	0.42
sentimentalMihaescuNAmor	0
sentimentalOeconomuCRucar	0.02
sentimentalPetraPetescuHBarcarola	0
sentimentalPopVAmericana	0.07
sentimentalRadulescuNRomanul	0.07
sentimentalTaillerEDragoste	0.1
sentimentalVaianEMaritata	0
sentimentalXenopolADorinMolna	0
sentimentalXenopolAPeurmaRazboiului	0

Figure 3: Imposters scores for hajduk and sentimental

A bit of autoethnography and conclusions

In terms of actual results, the tests on our collection of 185 novels have indicated, first of all, that Stylo package is useful for "profiling" subgeneric signals in Romanian novel corpora and in showing, rather straightforwardly, that "strong" genres such as the *hajduk* are constellated by leading figures not only in terms of population size (*see* N.D. Popescu's case), but also according to other criteria. However, in the absence of clear strategies, such as "thematic foregrounding" (a means of highlighting the thematic aspects and of backdropping the stylistic aspects of a text), Stylo should be used with care, as a diagnosis means. It is, thus, useful to evaluate through statistical lens traditional assumptions on periods, subgenres, authors' co-influencing etc., but these results should be credited only as preliminary.

There are studies which showcase the fact that choosing the right distance metric is crucial for improving genre clustering; in fact, various combinations of algorithms have reflected that Jensen–Shannon divergence is the best distance metric for genre recognition, Euclidean—the worst. Also, lexical simplification (replacing rare words with their most common synonyms) has not led to a noticeable improvement in genre recognition (Sobchuk & Šeļa, 2024). The bag of words approach (5000 MFW rather lesser MFW) and lemmatization—that we experimentally tried in this study, along with other combinations of features and preprocessing tasks that we have patiently performed—does very well on subgenre clustering, which was the focus of the present contribution. Therefore, beside studies that could replicate on Romanian novel corpora the kind of pipeline described in Sobchuk & Šeļa 2024, we believe that evaluation of the tool's performance as far as a precise language is concerned (in our case Romanian) should be done by means of multilingual comparative stylometry. This approach has yielded very nice results for "lesser-resourced" languages[11] too (*see* Schöch, 2024 for Hungarian and Ukranian).

We entitled the present contribution "stylometry for dummies" in a self-ironical mood, but also for several other reasons: to indicate the experimental level of our approach to both the tool's and the testing material's characteristics; to point at the perplexity that the novice researcher experiences in front of large, easy-to-get and easy-to-discard sets of "results" produced by the Stylo Package. After spending some time combining features, changing parameters, optimizing processing and other tasks, we believe that beginners should keep a diary in which self-observation while using the tool and progress with understanding the tool's computational premises should mix into a sort of "survival guide".

References:

Burrows, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford University Press.

Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing, 17*(3), 267–287.https://doi.org/10.1093/llc/17.3.267.

Calvo Tello, J. (2021). *The Novel in the Spanish Silver Age: A Digital Analysis of Genre Using Machine Learning.* Bielefeld University Press. https://doi.org/10.1515/9783839459256.

Compagnon, A. (2001). *Théorie de la littérature: la notion de genre.* https://www.fabula.org/compagnon/genre.php (Accessed: 10 November 2022).

DCRR. (2023). *Dicționarul cronologic al romanului românesc de la origini până în 2000* (Vol. I-II). Presa Universitară Clujeană.

Dinu, L., Popescu, M., & Dinu, A. (2008). Authorship Identification of Romanian Texts with Controversial Paternity. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* Marrakech, Morocco. http://www.lrec-conf.org/proceedings/lrec2008/pdf/862_paper.pdf (Accessed: 10 October 2023).

[11] The phrase "lesser-resourced" language, as well as "less(er)-used" language, is debatable as shown by Roxana Patras in her lecture DIGITAL HUMANITIES and "LESSER-RESOURCED" LANGUAGES:appropriation, imitation, localization of concepts, methods and tools, delivered at the ESUDH 2024 (https://esu-ct.conference.ubbcluj.ro/lectures/#lecture-patras).

Dockray-Miller, M., Drout, M. D. C., Kinkade, S., & Valerio, J. (2021). The Author and the Authors of the 'Vita Ædwardi Regis:' Women's Literary Culture and Digital Humanities. *Interfaces: A Journal of Medieval European Literatures*, *8*, 160-213. https://doi.org/10.54103/interfaces-08-09.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *R Journal, 8*(1), 107–121. https://doi.org/10.32614/RJ-2016-007.

Eder, M. (2017). Visualization in Stylometry: Cluster Analysis Using Networks. *Digital Scholarship in the Humanities, 32*(1), 50-64. https://doi.org/10.1093/llc/fqv061.

EDER, M. (2018). *Authorship Verification with the Package Stylo*. https://computationalstylistics.github.io/docs/imposters (Accessed: 17 February 2024).

Grieve, J. (2023). Register Variation Explains Stylometric Authorship Analysis. *Corpus Linguistics and Linguistic Theory, 19*(1), 47-77. https://doi.org/10.1515/cllt-2022-0040.

Hai-Ro: Proiect Bilateral România-Franța. (2020). https://proiectulbrancusihairo.wordpress.com/ (Accessed 27 April 2024).

Herrmann, J. B., Van Dalen-Oskam, K. H., & Schöch, C. (2015). Revisiting Style, a Key Concept in Literary Studies. *Journal of Literary Theory, 9*(1), 25-52. https://doi.org/10.1515/jlt-2015-0003.

Herrmann, J. B., Jacobs, A. M., & Piper, A. (2021). Computational Stylistics. In D. Kuiken & A. M. Jacobs (Eds.), *Handbook of Empirical Literary Studies* (pp. 451–486). De Gruyter.

Hoover, D.L. (2001). Statistical Stylistics and Authorship Attribution: An Empirical Investigation. *Literary and Linguistic Computing, 16*(4), 421-444. https://doi.org/10.1093/llc/16.4.421.

Jameson, F. (1981). *The Political Unconscious: Narrative as a Socially Symbolic Act.* Routledge.

Jungmannová, L., & Plecháč, P. (2024). Unsigned Play by Milan Kundera? An Authorship Attribution Study. *Digital Scholarship in the Humanities*, https://doi.org/10.48550/arXiv.2212.09879.

Juola, P. (2008). Authorship Attribution. *Foundations and Trends® in Information Retrieval, 1*(3), 233-334. http://dx.doi.org/10.1561/1500000005.

Juola, P. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities*, *30*(Suppl 1), i100–i113. https://doi.org/10.1093/llc/fqv040.

Kabala, J. (2020). Computational Authorship Attribution in Medieval Latin Corpora: the Case of the Monk of Lido (ca. 1101–08) and Gallus Anonymous (ca. 1113–17). *Language resources & Evaluation*, 54, 25-56. https://doi.org/10.1007/s10579-018-9424-0.

Karsdorp, F., Kestemont, M., & Riddell, A. (2022). *Humanities Data Analysis: Case Studies with Python*. Princeton University Press.

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding Groups in Data. An Introduction to Cluster Analysis*. Wiley.

Kestemont, M., Moens, S., & Deploige, J. (2015). Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux. *Digital Scholarship in the Humanities*, *30*(2), 199-224. https://doi.org/10.1093/llc/fqt063.

Lutosławski, W. (1898). Principes de stylométrie appliqués à la chronologie des oeuvres de Platon. *Revue des Études Grecques*, *11*(41), 65. https://www.persee.fr/doc/reg (Accessed: 27 May 2023).

McCarthy, R., & O'Sullivan, J. (2021). Who Wrote Wuthering Heights?. *Digital Scholarship in the Humanities*, *36*(2), 383-391. https://doi.org/10.1093/llc/fqaa031.

Misini, A., Kadriu, A., & Canhasi, E. (2022). A Survey on Authorship Analysis Tasks and Techniques. *SEEU Review, 17*, 153-167.

Modoc, E., & Gârdan, D. (2020). Style at the Scale of the Canon. A Stylometric Analysis of 100 Romanian Novels Published between 1920 and 1940. *Metacritic Journal for Comparative Studies and Theory*, *6*(2), 48-63. https://doi.org/10.24193/mjcst.2020.10.03.

Nitu, M., & Dascălu, M. (2024). Authorship Attribution in Less-Resourced Languages: A Hybrid Transformer Approach for Romanian. *Applied Sciences, 14*(7), 2700. https://doi.org/10.3390/app14072700.

Päpcke, S., Weitin, T., Herget, K., Glawion, A., & Brandes, U. (2023). Stylometric Similarity in Literary Corpora: Non-authorship Clustering and Deutscher Novellenschatz. *Digital Scholarship in the Humanities*, *38*(1), 277-295. https://doi.org/10.1093/llc/fqac039.

Patras, R. (2024): DH and lesser-resourced languages: appropriation, imitation, localization of concepts, methods and tools. *ESUDH 2024*. https://esu-ct.conference.ubbcluj.ro/lectures/#lecture-patras.

Patras, R., & Pascariu, L. (2024). Profiling-Genre-Signals-in-a-collection-of-Romanian-Novels-with-StyloR [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.10890870.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An Evaluation Framework for Plagiarism Detection. *Proceedings of the 23rd international conference on computational linguistics (Posters)* (pp. 997-1005). Beijing: Association for Computational Linguistics. https://aclanthology.org/C10-2115.pdf (Accessed: 10 May 2023).

Primorac, A., Arias, R., Patraș, R., Eglāja-Kristone, E., Van Dalen-Oskam, K., Herrmann, B., Schöch, C., & François, P. (2023). Distant Reading Two Decades On: Reflections on the Digital Turn in the Study of Literature. *Digital Studies/Le champ numérique*, *13*(1), 1-24. https://doi.org/10.16995/dscn.8855.

Rybicki, J. (2012). The Great Mystery of the (Almost) Invisible Translator: Stylometry in Translation. In M. P. Oakes & M. Ji (Eds.), *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research* (pp. 231-248). John Benjamins Publishing.

Schaeffer, J.-M. (2006). *¿Qué es un género literario?* (N. Campos Plaza & J. Bravo Castillo, Trans.). Akal. (Original work published 1989)

Savoy, J. (2020). *Machine Learning Methods for Stylometry. Authorship Attribution and Author Profiling*. Springer.

Schöch, C. (2022). *The European Literary Text Collection (ELTeC)*. Belgrade Training School, March 22, 2022. https://distantreading.github.io/eltec-slides/.

Schöch, C. (2023). What is Genre Analysis? In C. Schöch, J. Dudar, & E. Fileva (Eds.), *Survey of Methods in Computational Literary Studies (D3.2)*. Trier: CLS INFRA. https://doi.org/10.5281/zenodo.7892112.

Sobchuk, O., & Šeļa, A. (2024). Computational Thematics: Comparing Algorithms for Clustering the Genres of Literary Fiction. *Humanities and Social Sciences Communications, 11*, 438. https://doi.org/10.1057/s41599-024-02933-6

Teodorescu, H. N. L., & Bolea, S. C. (2018). Stylometric and Topic Analysis of a Historical Text. *ROMJIST*, *21*(2), 99-113. https://www.romjist.ro/full-texts/paper584.pdf (Accessed 24 September 2024).

Vianu, T. (1968). *Dicționarul limbii poetice a lui Mihai Eminescu*. Editura Academiei Republicii Socialiste România.

Yule, G. U. (1939). On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship. *Biometrika*, *30*(3-4), 363-90. https://doi.org/10.2307/2332655.