https://doi.org/10.17234/INFUTURE.2019.3

Preliminary communication

Quantitative Analysis of Adjectives in the Russian Literary Corpus of Realism and Romanticism

Lorena Kasunić
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
lkasunic@ffzg.hr

Petra Bago
Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
pbago@ffzg.hr

Summary

Computational analysis of text is an increasingly important approach used by researchers in the field of digital humanities. A much-debated question is whether computational techniques such as text analysis, which is in fact a quantitative approach, is adequate for analysing literary texts, since literature is considered as a type of artistic expression. In the paper we highlight the importance of the application of computational analysis with a study conducted on a corpus of selected Russian literary texts from the periods of Realism and Romanticism. Texts included in the romantic subcorpus are "Eugene Onegin" by Alexander Pushkin and "A Hero of Our Time" by Mikhail Lermontov. Texts that constitute the realist subcorpus are "Anna Karenina" by Leo Tolstoy and "Crime and Punishment" by Fyodor Dostoevsky. The analyzed texts are translations into the Croatian language. The paper presents current methods and approaches used in computational literature analysis. The focus of this research is the analysis of adjective usage in romantic and realist texts, due to the fact that these two literary periods are based on distinctive poetic principles. The texts were analyzed using the programming language "Python". Part-of-speech tagging was accomplished with an online tagger for Croatian language. Considering that all texts are historical (because they originate in the 19th or early 20th century) difficulties with POS tagging are expected. Results of the research show more similarites in the usage of adjectives between the subcorpora then expected. The paper points out how quantitative methods "borrowed" from the field of natural language processing and statistics can be significant in drawing conclusions about literature and that numbers can be meaningful if interpreted competently.

Key words: digital humanities, quantitative methods, stylometry, POS tagging, Croatian, adjective comparison, Russian Romanticism, Russian Realism

Introduction

Computational analysis as one of the methods in digital humanities is applicable to all digital and analog objects that can be studied, meaning that it is not limited to text objects. However, given that the emphasis in this paper is on the analysis of text and language, computational analysis settings will be presented in this context. Computational analysis is a technique that is possible only if there is digitized text. In case the researcher only has analogue text, it must first undergo the digitalization process (Text Analysis Resources, 2016).

What can computational analysis provide in comparison to traditional study of texts without the help of computers? First, it allows us to read a large number of texts in a short amount of time. Here the term "reading" refers to the possibility of passing through the text while orienting to some specific parameters (e.g. the usage of given names as opposed to family names in Jane Austen's novels). Secondly, texts can be automatically classified. Computers are trained to identify whether the analyzed text is a dictionary, a Greek tragedy, a historical epic poem or a letter. It is also possible to determine the authorship of a particular text based on the analysis of the corpus of a presumed author/s. Thirdly, computational analysis makes it easier to see the link between time-distanced texts (Computational Textual Analysis, 2018). Moreover, computer programs can be used to visualize the collected data (using charts, tables, text annotations, etc.). Furthermore, such analysis can empirically and statistically confirm the validity of initial hypotheses, enabling theorists to obtain evidence for their hypotheses (Kerr, 2017).

Related work

Within the field of computational analysis, there are two main approaches: a quantitative and a qualitative approach. However, these approaches are mainly not so distinctive, and they often overlap. Hoover (2008) gives a definition of a quantitative approach claiming it is an approach to literary texts where features or elements of literary texts are numerically represented, applying strong, precise and widely accepted methods of mathematics to measurement, classification and analysis. The increase in the number of available digital texts raised the interest in this approach and stressed the innovation in the ways in which literary texts are treated.

Petrović and Vranešević (2015) defined a quantitative approach as an innovative way of reading literary texts. What is of interest to the quantitative approach to literary texts is the issue of authorship and style, but it is also concerned with some more specific and complex issues such as: genre, theme, tone of the text, periodization (Hammond, 2016).

One of the most popular application of the quantitative approach in literature is stylometry where literary styles are analyzed using the distant reading method (Laramée, 2018). It rests on the premise that authors write in a distinctive, machine-detectable unique way. Problems which stylometry studies are closest to those addressed by the science of literature, with particular interest in patterns and repetitions that are related to issues of interpretation, meaning and aesthetics. The process of stylometric analysis consists of several complex multifactorial stages of preprocessing, feature extraction, statistical analysis and presentation of results, often by visual means (Eder et al., 2016). The linguistic level and grammatical, orthographic, syntactic and morphological research of the text should also not be neglected. The possibilities of applying computational methods in the process of analysis are especially significant when it comes to drawing conclusions from data that even a professional reader (a university professor, literary critic or literary theorist) cannot “detect” by close reading - in this method, the researchers use just distant reading.

The data on the number of transitive verbs, adjectives or the total number of words in Dickens's Great Expectations can be both obtained manually and by quantitative approach. When counting number of parts of speech manually, there is a greater chance of mistakes and it takes a longer time.

In one of her researches on Kafka’s literary corpus Berenike Hermann (2017) conducted keyness analysis and compared extracted keywords from the texts of 4 modernistic German authors. She focused on word classes and noticed (based on the given results) the existance of a high frequency of lemmas that may perform “modal” functions in the discourse in the Kafka’s corpus. The research shows that quantitative exploration of single words is quite useful (Berenike Hermann, 2017).

Similarly, Algee-Hewitt et al. (2016) analyzed the frequency of combination of any two consecutive words that repeat themselves in observed nineteenth-century novels (canonical and non-canonical texts). This process of lingustic redundancy revealed significat difference between canonical and non-canonical texts: three-fourths of the canonical texts (from the Chadwyck-Healey collection) was less redundant than three-fourths of the non-canonical texts (collected from libraries). This tells us that authors who used language in a redundant way had a bigger chance of being forgotten and remaining unread (Algee-Hewitt et al., 2016).

Qiu and Zhang (2015) in their paper (where they propose new methods of word segmentation for Chinese novels) conclude the following: “For example, based on the analysis of syntax, major events can be extracted from the novel, the relationship between characters can be automatically detected, and sentiment of the author can be analyzed.” Just like in our research, Qiu and Zhang use POS tagging as a baseline segmentor.

Kutuzov (2010) on the other hand investigates word types, word tokens and types to token ratio. He compares K. Vonnegut's two novels and their Russian translations, using mostly statistical methods. Many experts who use computer tools in their study of literature are increasingly advocating that computational analysis should join with traditional studies, and that they should complement each other (Hammond, 2016). Quantitative approaches must be aligned with existing ideals and practices in the humanities. The presentation of the mere fact of how many nouns there are in a particular novel does not serve any purpose if it does not consider and clarify the meaning of such data. One appearance of a particular language feature can be much more interesting and important than ten occurrences of another language feature (Petrović, Vranešević, 2015). Quantitative analysis is great for detecting what is rarely or unconventionally used in specific texts. And this can only be detected by counting and comparing, which computing enables (Hoover, 2008).

Dataset

The research about the usage of adjectives in the corpus was conducted on the Croatian translations of works by Russian romanticists and realists. The corpus was divided into two subcorpora – a realist and a romantic corpus. The realist subcorpus consisted of the novels *Anna Karenina* by Leo Tolstoy (translated by Martin Lovrenčević) and *Crime and Punishment* by Fyodor Dostoevsky (translated by Iso Velikanović). The romantic subcorpus consisted of the verse novel, *Eugene Onegin* by Alexander Pushkin (translated by Ivan Trnski), and the novel *A Hero of Our Time* by Mikhail Lermontov (translated by Milan Bogdanović).

Romanticism and Realism as literary periods rest upon contrary principles. These periods do not share the same worldview nor do they perceive the social and cultural reality similarly. Main characteristics of Romanticism are: rejection of the idea of order and racionalism, emphasis on the emotions, irrationality, subjectivity. Authors are occupied with the idea of a genius, a hero, an exceptional individual who is a visionary creator (Croatian Encyclopedia, n.d.). They often connect that individual with the surrounding nature. Lyrical expressions, outburst of emotional states presented in a form of description (when talking about novels) - all these features are considered typical for the literary period of Romanticism.

Realism, on the other hand, tends to be coprehensive as much as possible. Character shaping is of crucial importance for realist authors. Much space is given to showing social, cultural, economical and political circumstances and conditions. Unlike Romanticism, Realism tries to reduce the ramification of the plot and wants to put a light on character's development (Croatian Encyclopedia, n.d.). Prose, especially novel, is the dominant tool of literary expression. Pushkin and Lermontov are canonical names in Russian romantic literature and because of that they can be considered as representatives of a typical Realism poetics. Taking texts of similar artistic and poetic value makes the comparison more accurate and precise. Description as a narrative technique is present both in Romanticism and Realism but in different ways. Romantic authors describe nature, feelings and melancolic atmosphere. Realist authors describe physical appearance, the space of the plot (wretched houses and flats, public houses etc.) on a very "realistic" way, without idealization or emotional enthusiasm. That is the reason why the usage of adjectives shoud differ in the romantic and the realist subcorpora.

These texts were chosen because they represent the culmination of Russian literature and are representatives of the periods from which they originate. All the texts selected, except *Eugene Onegin*, are deliberately prose because belonging to the same literary form requires the application of similar principles in the structure of the text. Poems and dramas have a specific structure and were therefore not considered. It is well-known that the novel as a form prevailed in Realism, while Russian Romanticism remains known for the texts analyzed in our research, even though Pushkin was a great poet. The main hypothesis was that the use of adjectives would differ in the analyzed subcorpora since Romanticism and Realism are based on, we may say, completely opposed poetics and modes of expression, as well as thematic preoccupations. The texts do not have a balanced number of tokens, so below in Table 1 we have provided the data on the size of the corpus, subcopora and the individual literary texts, punctuation included:

Table 1. Number of tokens in each literary text and subcorpus

	Romantic subcorpus		Realist subcorpus	
	Title of text	Number of tokens	Title of text	Number of tokens
	Eugene Onegin	31,516	*Anna Karennina*	370,282
	A Hero of Our Time	55,134	*Crime and Punishment*	229,328
Total	86,650		599,610	

(subcorpora)		
Total (corpora)	686,260	

Methodology

Python (version 3.6.0) and POS tagging (for the Croatian language[1]) were used in the process of computational analysis. All texts were downloaded from the *eLektire* website in the .txt format for further study[2].

The research was based on the application of statistical methods and methods used in natural language processing. The first step was to preprocess the texts - removing data such as footnotes, notes, titles, author's names, dictionaries of lesser-known words. It was necessary to obtain "pure" literary texts without any metadata, as they would affect the results of the research. Subsequently, the texts were processed using the online POS tagger, through which the texts were morphosyntactically tagged and the lemmatization was performed. All subsequent analyses were performed on the processed text.

It was necessary to check the accuracy of the classifier before the data was obtained. This was done by taking one segment of the text and tagging it manually by one annotator. First 500 tokens (including punctuation) from *Eugene Onegin* were chosen for manual tagging. Since the research was focused on adjectives, manual tagging did not go into deep morphosyntactic analysis, but only verified whether the parts of speech were correctly labeled.

Using the *Python* programming language, we obtained data on the frequency distribution of words and punctuation for each text, with reference to the tags used when tagging texts in Croatian. After that, only the adjectives were generated, in order to show which ones appeared most often in a particular text.

The tagger wrongly labeled some words as adjectives, that we excluded from the analysis. Methods used in the analysis of literary texts are mostly "borrowed" from natural language processing and statistics. In this study, the methods used were frequency distribution, tokenization, lemmatization and POS tagging.

Results

The first results that were obtained were related to the accuracy of the classifier itself. The website[3] of the Croatian POS tagger states that the accuracy for the Croatian language is 92.53%. We manually labeled the first 500 tokens (including the punctuation) of *Eugene Onegin* to evaluate the POS tagger and obtained the accuracy of 87.2%. When studying the data, it was apparent that the classifier did not make any mistakes in tagging punctuation, so the accuracy was calculated if the punctuation was excluded. The accuracy then dropped to 84.2%. Out of the 61 errors counted, 10 were related to the wrong tagging of adjectives. In Table 2 we provide the confusion matrix for POS tags.

Table 2. The confusion matrix for POS tags

		Actual class												
		N	V	A	P	R	S	C	M	Q	I	Y	X	Z
Predicted class	N	93	12	3	3	6	0	1	0	0	0	0	0	0
	V	2	64	3	2	4	0	0	0	0	0	0	0	0
	A	2	4	56	3	3	0	0	0	0	0	0	0	0
	P	2	0	2	45	0	0	0	0	0	0	0	0	0
	R	0	0	1	1	18	0	0	0	0	0	0	0	0
	S	0	0	0	0	0	22	0	0	0	0	0	0	0
	C	0	0	0	3	0	0	15	0	1	0	0	0	0
	M	0	0	1	0	2	0	0	6	0	0	0	0	0
	Q	0	0	0	0	0	0	1	0	3	0	0	0	0
	I	0	0	0	0	0	0	0	0	0	0	0	0	0
	Y	0	0	0	0	0	0	0	2	0	0	1	0	0
	X	0	0	0	0	0	0	0	0	0	0	0	0	0

[1] http://www.clarin.si/info/about/

[2] https://lektire.skole.hr

[3] http://www.clarin.si/info/k-centre/web-services-documentation/

	Z	0	0	0	0	0	0	0	0	0	0	0	0	113

If we look at this from another direction (specifically, if we are interested in how many words the classifier tagged as adjectives, when they were in fact some other part of speech), it comes up to 12 mistakes. Of the 61 errors in total, 22 are connected with adjective annotation.

Next, we calculated the frequency distribution of word types. Table 3 shows the frequency distribution of words and punctuation for each text.

Table 3. Frequency distribution of word types and punctuation

	Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*	Romantic subcorpus	Realist subcorpus
Verbs	4,422 (14.03%)	**11,838 (21.47%)**	45,541 (19.86%)	**77,475 (20.92%)**	16,260 (18.77%)	**123,016 (20.52%)**
Nouns	7,546 (23.94%)	8,945 (16.22%)	33,123 (14.44%)	62,494 (16.88%)	16,491 (19.03%)	95,617 (15.95%)
Adpositions	1.730 (5.49%)	3,435 (6.23%)	12,690 (5.53%)	23,449 (6.33%)	5,165 (5.96%)	36,139 (6.03%)
Adverbs	1,364 (4.33%)	3,660 (6.64%)	17,096 (7.45%)	24,184 (6.53%)	5,024 (5.80%)	41,280 (6.88%)
Adjectives	3,131 (9.93%)	3,642 (6.61%)	13,898 (6.06%)	25,218 (6.81%)	6,773 (7.82%)	39,116 (6.52%)
Pronouns	3,215 (10.20%)	7,284 (13.21%)	28,471 (12.41%)	49,032 (13.24%)	10,499 (12.12%)	77,503 (12.93%)
Numerals	470 (1.49%)	554 (1.00%)	1,886 (0.82%)	2,635 (0.71%)	1,024 (1.18%)	4,521 (0.75%)
Conjuctions	1,260 (4.00%)	4,617 (8.37%)	21,388 (9.32%)	32,904 (8.89%)	5,877 (6.78%)	54,292 (9.05%)
Interjections	52 (0.16%)	98 (0.18%)	480 (0.21%)	447 (0.12%)	150 (0.17%)	927 (0.15%)
Particles	666 (2.11%)	1,079 (1.96%)	6,217 (2.71%)	7,722 (2.09%)	1,745 (2.01%)	13,939 (2.32%)
Abbrevations	40 (0.13%)	9 (0.02%)	14 (0.01%)	64 (0.02%)	49 (0.06%)	78 (0.01%)
Residuals	25 (0.08%)	27 (0.05%)	28 (0.01%)	187 (0.05%)	52 (0.06%)	215 (0.04%)
Punctuation	**7,595 (24.10%)**	9,946 (18.04%)	**48,496 (21.15%)**	64,471 (17.41%)	**17,541 (20.24%)**	112,967 (18.84%)
Total	31,516	55,134	229,328	370,282	86,650	599,610

The morphosyntactic tagger is trained on the corpus of texts that are part of the Croatian Language Repository, hrWaC (the Croatian Web Corpus), and the Croatian National Corpus. Data on the first two corpora are freely available on the Internet and they were used to compute the relative frequency distribution of parts of speech in order to compare the results with the corpus of texts used in this research. In *Eugene Onegin*, the biggest difference in percentages is visible in the distribution of punctuation marks: in this text their relative frequency distribution is 24.1%, in hrWaC is 11.89%, or 14.29% in the Croatian Language Corpus. In *A Hero of Our Time*, except for punctuation, there are differences in representation of verbs (21.47% in the analyzed text, 16.05% in hrWaC and 14.82% in the Croatian Language Repository), nouns (16.22% in the analyzed text, 26% in hrWaC, 27.91 % in the Croatian Language Repository) and pronouns (13.21% in the analyzed text, 8.2% in hrWaC and 6.88% in the Croatian Language Repository). In the novel *Crime and Punishment* the biggest differences are again found in the punctuation (we may say that this is the leitmotif in all the analyzed texts), then in the nouns (14.44% is the relative frequency in the novel, the percentages for hrWaC and the Croatian Language Repository are identical to the ones previously mentioned) (12.41%). The same goes for *Anna Karenina*: the relative frequency distribution of nouns is 16.88%, punctuation marks 17.41%, and verbs 20.92%. As far as adjectives are concerned, the largest relative frequency distribution is 9.93% and refers to *Eugene Onegin*, while the rest of the texts have a similar relative frequency distribution for adjectives: *A Hero of Our Time* - 6.61%, *Crime and Punishment* - 6.06%, *Anna Karenina* - 6.81%.

Table 4. The number of individual adjectives in each text

	Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*
Descriptive adjectives	**2,961 (94.57%)**	**3,389 (93.05%)**	**12,589 (90.58%)**	**23,160 (91.84%)**
Possessive adjectives	53 (1.69%)	32 (0.88%)	379 (2.73%)	587 (2.33%)
Past participle	117 (3.74%)	221 (6.07%)	930 (6.70%)	1,471 (5.83%)
Total number of adjectives	3,131	3,642	13,898	25,218

Within the category of adjectives, an additional frequency distribution of certain adjectives (descriptive, possessive, and past participle) was performed, as can be seen in Table 4. The classifier recognizes these three types of adjectives and this is the reasoning behind this categorization, although in the grammar of the Croatian language there is a basic categorization of adjectives into descriptive, constructive and possessive. Given that the analysis is predominantly oriented on the presence of adjectives in each of these literary texts, we assembled the data on the 20 most common adjectives found in particular texts. From the results, it is evident that adjectives "sam" (Eng."alone") and "sav" (Eng. "entire") appear in all four texts. These are the most commonly used adjectives. In hrWaC, the frequency distribution for "sam" (Eng. "alone") is 1,261,043 (0.97% of the total number of adjectives in the corpus) and 5,303,295 (4.07%) for "sav" (Eng. "entire"). In the Croatian Language Repository frequency distribution for "sam" (Eng. "alone") is 66,518 (0.65% of the total number of adjectives in the corpus) and 288,819 (2.84%) for "sav" (Eng. "entire"). All the other 20 most common adjectives for each literary text can be seen in Table 5. As mentioned in the previous chapter, words (and letters) which were excluded from the table of adjectives are: "moj" (Eng. "mine"), "vaš" (Eng. "your"), "njen" (Eng. "hers"), "Svidrigajlov" (Eng. "Svidrigailov"), "Raskoljnikov (Eng. "Raskolnikov"), "njegov" (Eng. "his"), "l" (Eng. "l").

Table 5. 20 most common adjectives for each literary text

Eugene Onegin	*A Hero of Our Time*	*Crime and Punishment*	*Anna Karenina*	hrWaC	Croatian Language Repository
sav (Eng. entire)	sav (Eng. entire)	sam (Eng. alone)	sav (Eng. entire)	sav (Eng. entire)	sav (Eng. entire)
mlad (Eng. young)	sam (Eng. alone)	sav (Eng. entire)	sam (Eng. alone)	velik (Eng. big)	hrvatski (Eng. Croatian)
sam (Eng. alone)	velik (Eng. big)	isti (Eng. same)	dobar (Eng. good)	nov (Eng. new)	velik (Eng. big)
star (Eng. old)	čitav (Eng. intact)	cijel (Eng. whole)	isti (Eng. same)	dobar (Eng. good)	nov (Eng. new)
lijep (Eng. beautiful)	hladan (Eng. cold)	velik (Eng. big)	nov (Eng. new)	hrvatski (Eng. Croatian)	dobar (Eng. good)
mio (Eng. dear)	dobar (Eng. good)	posljednji (Eng. last)	lijep (Eng. beautiful)	sam (Eng. alone)	europski (Eng. European)
krasan (Eng. splendid)	mlad (Eng. young)	dobar (Eng. good)	velik (Eng. big)	mali (Eng. small)	sam (Eng. alone)
velik (Eng. big)	crn (Eng. black)	nov (Eng. new)	star(Eng. old)	isti (Eng. same)	državni (Eng. national)
nov (Eng. new)	čudan (Eng. strange)	neobičan (Eng. unusual)	mlad (Eng. young)	cijel (Eng. whole)	politički (Eng. political)

ruski (Eng. Russian)	isti (Eng. same)	čudan (Eng. strange)	veseo (Eng. merry)	ostali (Eng. remaining)	mali (Eng. small)
tanak (Eng. thin)	prav (Eng. real)	mali (Eng. small)	moguć (Eng. possible)	star (Eng. old)	isti (Eng. same)
sladak (Eng. sweet)	bijel (Eng. white)	pijan (Eng. drunk)	čitav (Eng. intact)	mlad (Eng. young)	američki (Eng. American)
čudan (Eng. strange)	štaban (Eng. headquartered)	osobit (Eng. special)	potreban (Eng. neccesary)	važan (Eng. important)	posljednji (Eng. last)
živ (Eng. alive)	uvjeren (Eng. convinced)	jasan (Eng. clear)	posljednji (Eng. last)	potreban (Eng. neccessary)	glavni (Eng. main)
drag (Eng. dear)	posljednji (Eng. last)	prav (Eng. real)	sretan (Eng. happy)	prav (Eng. real)	zagrebački (Eng. Zagreb's)
prost (Eng. vulgar)	pun (Eng. full)	mlad (Eng. young)	mali (Eng. small)	poznat (Eng. famous)	svjetski (Eng. worldwide)
bijel (Eng. white)	lijep (Eng. beautiful)	bolestan (Eng. sick)	visok (Eng. tall)	mnogi (Eng. many)	star (Eng. old)
hladan (Eng. cold)	smiješan (Eng. funny)	glup (Eng. stupid)	bijel (Eng. white)	europski (Eng. European)	ostali (Eng. remaining)
tih (Eng. quiet)	blijed (Eng. pale)	star (Eng. old)	miran (Eng. still)	glavni (Eng. main)	međunarodni (Eng. international)
dobar (Eng. good)	star (Eng. old)	strašan (Eng. terrible)	strašan (Eng. terrible)	visok (Eng. tall)	mlad (Eng. young)

Discussion

The application of the classifier on the corpus of literary texts showed that the classifier successfully performed part-of-speech tagging. The accuracy difference is 5.33% (accuracy of the classifier obtained on tested text fragment: 87.2%), compared to the data about the accuracy of the classifier available on its website (92.53%). One ought to keep in mind that here we are dealing with Croatian translations of historical texts created during the 19th and 20th century and therefore their language differs from typical contemporary Croatian language. To test the accuracy of the classifier a text fragment was taken from *Eugene Onegin* because the text (by its structure and language) deviates the most from everyday language, and therefore it is more likely that the classifier will make a mistake whilst tagging. Because of the lack of information on the dates of translations for *Anna Karenina* and *Crime and Punishment*, one cannot decidedly explain the wrong tagging by the fact that the translation of *Eugene Onegin* is the oldest from the analyzed texts, although this possibility should not be dismissed.

From the results of the relative frequency distribution, we can conclude that the literary texts which make up our corpus have a different number of nouns, verbs, pronouns and punctuation marks in relation to hrWaC and the Croatian Language Repository. *Anna Karenina*, *A Hero of Our Time* and *Crime and Punishment* use more verbs and pronouns and fewer nouns in relation to the above-mentioned corpora. With *Eugene Onegin*, it is a different situation. There are less verbs, pronouns and even conjunctions than in other novels. On the other hand, nouns and adjectives are more present than in other texts. This seems to be the case because this a verse novel that inherited a part of the lyrical influence. Although there is a plot, it is not the central aspect of the novel, and therefore there are fewer verbs. The greater quantity of adjectives and nouns can be explained by the presence of the lyrical mode of expression which strives for descriptiveness, expressiveness, and enumeration. A low percentage of conjunctions can be associated with a high percentage of punctuation marks, which are an essential element of writing in verse. Comparing the two subcorpora, the realist and romantic, and then considering the relative frequency distribution, *A Hero of Our Time* could be assigned to the realist rather than a romantic subcorpus. It must be emphasized that this conclusion was reached without considering the content of the text itself, so it is exclusively based on statistical data. Punctuation generally has a much larger share in all literary texts than in the Croatian Web Corpus and the Croatian Language Repository. This phenomenon is understandable as literary texts often use complex sentences, *stringing*, inserts, and these devices increase the use of commas, colons, ellipses, parentheses, etc.

If we observe the frequency distribution of a particular type of adjective (descriptive, possessive, and past participle), we can observe a uniformity of the ratios within all analyzed texts. The highest in frequency are descriptive adjectives, then past participles and finally possessives. From this, one can read the common feature of Realism and Romanticism - the aspiration to describe, whether the fictional world in which the action takes place, the characters that inhabit it, or feelings and emotional states. What is of greatest interest is which specific adjectives appear in a particular text. We produced lists of 20 most frequent adjectives for each text. Through our research, it became apparent that the adjectives on the top of the list are very similar, namely “sav” (Eng. “entire”) and sam (Eng. “alone”). The romantic subcorpus contains a lot of similarities and repetition of adjectives - both in *Eugene Onegin* and *A Hero of Our Time*, the following adjectives are present: “sav” (Eng. “entire”), “sam” (Eng. “alone”), “lijep” (Eng. “beautiful”), “bijel” (Eng. “white”), “velik” (Eng. “big”), “hladan” (Eng. “cold”), “čudan” (Eng. “strange”), “mlad” (Eng. “young”), “star” (Eng. “old”). Both texts have one or two pairs of mutually opposing adjectives: “mlad” (Eng. “young”) and “star” (Eng. “old”) in *Eugene Onegin* and *A Hero of Our Time* and “bijel” (Eng. “white”) and “crn” (Eng. “black”) in *A Hero of Our Time*. There are similarities in the realist subcorpus as well. Adjectives that appear both in *Crime and Punishment* and in *Anna Karenina* are “sam” (Eng. “alone”), “sav” (Eng. “entire”), “isti” (Eng. “same”), “nov” (Eng. “new”), “velik” (Eng. “big”), “posljednji” (Eng. “last”), “dobar” (Eng. “good”), “mali” (Eng. “small”), “mlad” (Eng. “young”), “strašan” (Eng. “terrible”), “star” (Eng. “old”). What is particularly important for these texts is that the names of characters (Raskolnikov, Svidrigailov and Vronski) are referred to as adjectives. The reason for this is in the suffixes -ov and -ski which are typical for adjectives in the Croatian language, and not for proper names. There are also pairs of adjectives: “velik” (Eng. “big”) - “mali” (Eng. “small”) (*Crime and Punishment*) and “star” (Eng. “old”) - “mlad” (Eng. “young”) (*Anna Karenina* and *Crime and Punishment*). It is interesting that the adjective “strange” appears in all the texts, except in *Anna Karenina*. In fact, a large number of adjectives are common to all four texts, although they belong to different literary periods. This could be perceived as one of the theories often emphasized by literary theorists and that is that one cannot draw clear boundaries between literary periods. Influences are always present and for no great literary text can be said to be a typical realistic or romantic text, for example.

What were the problems we encountered while conducting the research? The first problem occurred when using the classifier. Namely, the realist texts are much more extensive than the romantic ones and when they were supposed to be tagged, they were simply too large and the online classifier was blocked. Therefore, these texts had to be divided into several smaller text files that were then tagged and merged into a single text file that was used in further analysis. When launching the *Python* program code for each text, it was apparent that due to the size of *Anna Karenina* file and *Crime and Punishment* file, it took more time for *Python* to produce the data. This was especially noticeable in the part of the research where 20 of the most frequent adjectives were extracted. As for the manual tagging, there were problems with defining parts of speech for some words because they were outdated (e.g. “priljem”, “ponevju”, “stežne”, “zgolje”) and the annotator did not know the meaning of these words.

The results have shown that there are some differences in the use of adjectives between the two subcorpora, but they are not as drastic as expected. A possible answer might lie in the fact that what makes literary texts differ from one another are language phenomena that are not so great in number, i.e., those that are specific to a certain text. In this particular case that would mean that we should study the adjectives that are in the middle of frequency rankings - they are not so rare that they could be considered as exceptions and not so frequent that their occurrence could be attributed to the general method of structuring and using language in literary art. Given adjectives could be further observed in the surrounding context in which they are occur. It would also be interesting to see if are there any patterns in distribution of different types of adjectives throughout the romantic and the realist subcorpora. Perhaps a similarity in the choice of adjectives may confirm a certain intertextuality and literary influences. That could be a possible direction for future work on this or similar corpora of literary texts.

Conclusion

Quantitative approaches have their supporters, as well as their opponents. They contribute to the improvement and further development of practices within digital humanities. Examples of concrete application of methods in computational analysis show that empirical data can be of use in attempts to interpret literary texts.

We attempted to implement computational methods in order to test our hypothesis that the usage of adjectives differs between the two opposite literary periods. The research focused exclusively on the use of quantitative tools, such as frequency distribution. It also used methods from natural language processing (POS tagging, lemmatization, tokenization). The process of conducting the research has confirmed the usefulness of the quantitative approach in the interpretation of literature, but only if there is a human agent who will be able to interpret the obtained data. The paper endeavored to give empirical results and conclusions which can shed a light on the complicated question of boundaries between literary periods.

Although the beginnings of computer usage in studying historical texts (including literary ones) originate in the 1950s, there is still room for progress. More focus needs to be put on the development of new tools and methods, especially for texts that are not written in world languages such as English. There is still a lack of properly digitized machine-readable Croatian texts, especially historical ones from various (literary) periods. Research is mostly carried out on large canonical texts, and the less famous ones are neglected. There is a need for cooperation of experts from different fields - information science, linguistics, literature, computer science etc. However, despite all the obstacles encountered by digital humanists, computer analysis is increasingly being used as a new approach to literary texts, one that can provide a different point of view and encourage us to ask previously untold or overlooked questions.

References

Algee-Hewitt, M., Allison, S., Gemma, M., Heuser, R., Moretti, F., Walser, H. (2016). Canon/Archive. Large-scale Dynamics in the Literary Field. https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf (25.10.2019.)

Berenike Hermann, J. (2017). In a test bed with Kafka. Introducing a mixed-method approach to digital stylistics. // Digital Humanities Quarterly 11, 4. http://www.digitalhumanities.org/dhq/vol/11/4/000341/000341.html (14.3.2019.)

Computational Textual Analysis. (2018). https://guides.temple.edu/corpusanalysis (15.3.2019.)

Croatian Encyclopedia. http://www.enciklopedija.hr/ (24.10.2019.)

Eder, M., Rybicki, J., Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. // The R Journal 8, 1. https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf (20.7.2019.)

Hammond, A. (2016). Quantitative Approaches to the Literary. // Literature in the Digital Age: An Introduction / Cambridge: Cambridge University Press, 82-130

Hoover, D. L. (2018). Quantitative Analysis and Literary Studies. // A Companion to Digital Literary Studies / Schreibman, S., Siemens, R. (eds.). Oxford: Blackwell. http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.id=0&brand=9781405148641_brand (26.2.2019.)

Kerr, S. J. (2017). When Computer Science Met Jane Austen and Edgeworth. // NPPSH Reflections 1, 38-41

Kutuzov, A. (2010). Change of word types to word tokens ratio in the course of translation (based on russian translations of k. Vonnegut's novels). https://arxiv.org/ftp/arxiv/papers/1003/1003.0337.pdf (25.10.2019.)

Laramée, F. D. (2018). Introduction to stylometry with Python. https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python (22.7.2019.)

Petrović, B., Vranešević, D. (2015). Kvantitativna raščlamba Čudnovatih zgoda šegrta Hlapića Ivane Brlić-Mažuranić. // Šegrt Hlapić – od čudnovatog do čudesnog / Majhut, B., Narančić Kovač, S.; Lovrić, S. (eds.). Zagreb: Slavonski Brod: Hrvatska udruga istraživača dječje književnosti: Ogranak Maticea hrvatske, 251-267

Qiu, L., Zhang, Y. (2015). Word Segmentation for Chinese Novels. // Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9323/9538 (24.10.2019.).

Technopedia. https://www.techopedia.com/definition/13698/tokenization (19.7.2019.)

Text Analysis Resources. (2016). https://digitalhumanities.berkeley.edu/resources/text-analysis-resources, (15.3.2019.)

The Stanford Natural Language Processing Group. https://nlp.stanford.edu/software/tagger.shtml (20.7.2019.)