Metodološki zvezki, Vol. 5, No. 1, 2008, 81-93

Usage of Multivariate Analysis in Authorship Attribution: Did Janez Mencinger Write the Story “Poštena Bohinčeka”?

Marko Limbek[1]

Abstract

This paper uses different techniques of multivariate analysis in authorship attribution and shows that statistical methods can be successful in the field of stylometry and that useful results can be obtained.

1 Introduction

Unique solutions in the analysis of texts cannot be achieved with the use of subjective methods that depend on personal evaluation, therefore certain objective methods which would assure a level of certainty »beyond reasonable doubt« are called for. The aim is to obtain data for statistical analysis by the quantification of the characteristics of the texts.

In this case of authorship attribution the intention is to determine what distinguishes one author from the other authors in order to describe the author's personal style. Realistic results cannot always be guaranteed, but at least there is a wider choice of different techniques available provided. The usage of statistical methods in literature is very interesting and can be of great use in solving real questions. Some basic facts about the development of stylometry can be found in Holmes (1997), while in the last hundred years several authors have been exploring this field. Different techniques have been developed, ranging from less to more sophisticated.

The simplest technique is measuring the word length and the sentence length, which was done some time ago by Mendenhall (1887) and Yule (1938). This technique is simple and easily determinable, but not very reliable. The second

[1] Student of Statistics, University of Ljubljana, Kongresni trg 12, 1000 Ljubljana; marko.limbek@gmail.com

group of techniques uses the vocabulary distribution, which was extensively developed by Holmes (1991). It focuses on the distribution of the vocabulary frequency, especially on *hapax legomena* and *hapax dislegomena*, words that appear once or twice in the text and provide a good insight in the richness of the vocabulary. Holmes also deals with Sichel's distribution and Yule's characteristic K.

The last group of methods includes multivariate methods, used efficiently by Binongo (2003). In the style analysis it is important to obtain the fixed mark of the author's style to find the permanent characteristics, and those characteristics must be independent from the content which changes from one text to another. The main point is in using function words, such as pronouns, auxiliary verbs, prepositions, conjunctions, determiners and other closed-class words that form the skeleton of the text and do not have content. In this manner the method is somehow the opposite of exploring vocabulary distribution. An important fact is that the author cannot avoid using function words and moreover uses them unconsciously; they can be found even in the simplest texts, they do not change with the development of the language and they do not possess referential meaning, which is why they represent a truly objective source for determining the author's specifics. In the case in question, these function words are used as variables.

The background of linguistic phenomena has not been explored since the focus of the paper is in using methods of multivariate analysis. The relevance of using statistical methods should however be warranted too.

Slovenian authors have so far also been exploring the field. Dović (2002) has been using both method of measuring sentence and word length as well as cumulative method while performing an interesting analysis on possible plagiarism, whereas Primoz Jakopin has done some really extensive research especially in the field of enthropy and has also established a new corpus called New word. In this sense the paper somehow represents extension of their work by including methods of multivariate analysis into the national arsenal.

2 The problem

Multivariate analysis is often used for different kinds of authorship attribution. If there is a text for which it is unknown whether it belongs to one author or the other, usually the process is that the text and the text samples of both authors are analysed and compared and hopefully it is possible to determine to whom it is more likely to belong. In this particular case there is a story on the table for which the authorship is unknown and the examination is suitable to see whether it belongs to a certain author. The story in question is »Poštena Bohinčeka« from 1860 and the possible author is Slovenian writer Janez Mencinger. There are four other texts available that were undoubtedly written by him, and which originate from the same period around 1860, thus the time factor cannot serve as an

explanation for the difference in texts. All five texts will be analysed using the multivariate analysis methods, the differences and similarities between them will thus be determined and in the end the conclusion will be reached whether »Poštena Bohinčeka« was written by him or not.

The statistical units used consist of blocks of precisely 1000 words into which the available text has been cut. A computer programme written in *Perl* is then used to count the occurrences of each function word in each block, thus obtaining the distribution of all the function words. With such data it is now possible to start the multivariate analysis.

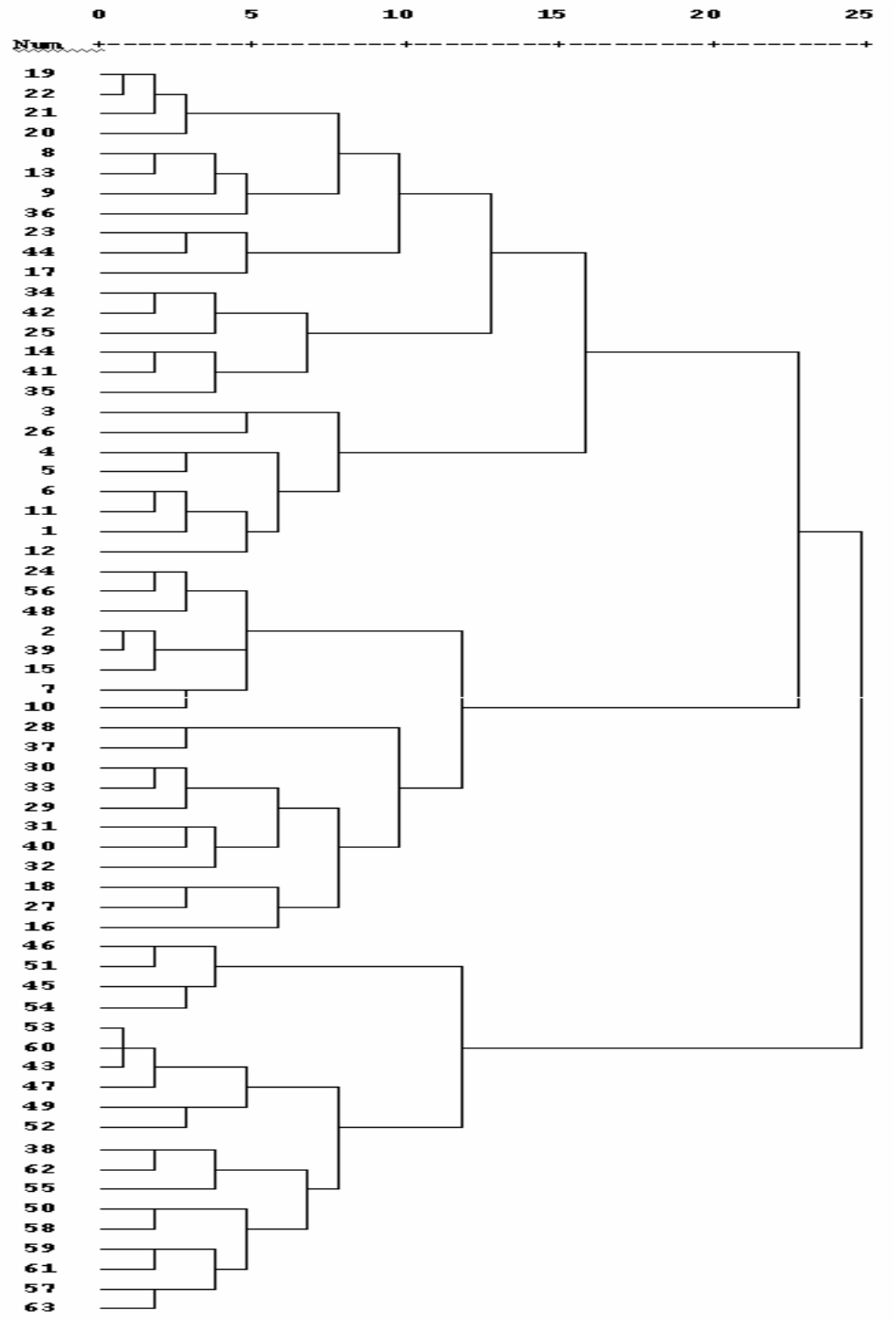
3 Data

There are five short stories at the disposal, the first four being Mencinger's: »Jerica« with more than 8.000 words, »Vetrogončič« with more than 11.000 words, »Človek toliko velja, kot plača« with more than 12.000 words and »Bore mladost« with more than 13.000 words. The main one, »Poštena Bohinčeka«, contains more than 19.000 words. In the manner described, »Poštena Bohinčeka« is divided into 19 blocks of thousand words, starting from the first word, while the words above 19.000 are to be neglected. The statistical result will not be harmed! In the same manner, 8 blocks are obtained from »Jerica«, 11 blocks from »Vetrogončič«, 12 blocks from »Človek toliko velja, kot plača« and 13 blocks from »Bore mladost«, which makes 44 blocks altogether from Mencinger and 19 from »Poštena Bohinčeka«. The total number of blocks is 63. The same number of words in each block also eliminates the basic need for the normalisation of variables.

There was a small additional project of how to compile a set of 50 most frequent function words in Slovenian language which would be chosen for variables. An existing list of some 200 function words had to be checked for their frequency through some bigger corpus and the cut off point was set after 50 words. The choice of words should be further discussed, since determining a basic set of function words is an important step to be made for any language. It resulted in the following stop words:

```
”ne”, “ki”, “le”, “tako”, “da”, “je”, “naj”, “ali”, “kar”, “k”,
“in”, “po”, “pri”, “proti”, “si”, “bo”, “v”, “iz”, “s”, “med”,
“cez”, “ko”, “kakor”, “kako”, “ker”, “z”, “pred”, “jaz”, “nic”,
“do”, “pa”, “ti”, “to”, “ga”, “brez”, “mu”, “bi”, “ni”, “kaj”,
“kadar”, “za”, “nihce”, “vse”, “preden”, “se”, “tudi”, “od”,
“ravno”, “na”, “o”.
```

At this point another *Perl* written programme is used to obtain the frequency of each function word in each block and to fill the matrix. Thus the preparation of data is complete and the analysis in SPSS can now continue.



Arranging units in order:
Bore mladost 1-13
Človek 14-25
Jerica 26-33
Vetrogonči 34-44
Poštena Bohinčeka 45-63

Figure 1: Dendrogram using Ward's Method.

4 Results

4.1 Cluster analysis

Clustering is classification of units into different groups, based on similarity of units, so that similar data is collected in the same group. The process is done in steps and each step can be observed in the belonging dendrogram. The method used is Ward's linkage with the least square distance. Arranging the 63 units in orderly blocks amounts to "Bore mladost" 1-13, "Človek" 14-25, "Jerica" 26-33, "Vetrogončič" 34-44 and "Poštena Bohinčeka" 45-63. As can be seen, dendrogram is in two parts, lower one, amost completely composed of blocks of »Poštena Bohinčeka«, and upper one, almost completely composed of blocks of other four stories. It is true, that blocks 48 and 56 of »Poštena Bohinčeka« are found in upper part, but all of other blocks are correctly put together. That means that common characteristics of the same text have been found and it also shows that other four stories are more similar, since they are mixed together in upper part.

4.2 Principal component analysis

Principal component analysis is used to obtain useful information out of multidimensional data. These multidimensional date are in a special way contracted in order to retain as much information as possible and to be somehow visible in less dimensions, preferably just two or three. Data structure can be seen. Results of principal component analysis are considered as the principal element in showing that Mencinger, in fact, was not the author of »Poštena Bohinčeka«. Performing principal component analysis according to 50 variables (frequency of selected words) shows that scree plot (in Appendix 2) breaks between the third and the fourth point and the first three components contain 27,584% of variance explained. The eigenvalues and variance explained of first ten components are listed in Table 1.

Table 1: Total variance explained.

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	4,929	10,953	10,953
2	4,215	9,367	20,319
3	3,269	7,264	27,584
4	2,343	5,207	32,791
5	2,202	4,893	37,684
6	2,000	4,445	42,129
7	1,913	4,250	46,380
8	1,825	4,056	50,436
9	1,725	3,833	54,269
10	1,583	3,518	57,787

Table 2: Component matrix.

words	C_1	C_2	C_3	English tranl.
ne	,354	**,682**	,129	no
ki	**-,512**	,218	,277	which
le	**,539**	-,149	,153	only
tako	**,401**	,147	,034	so
da	,297	-,118	**,624**	that
je	-,357	**-,763**	,138	is
naj	,050	-,029	,209	should
ali	,212	-,162	-,194	or
kar	**,498**	,019	,315	just
k	-,053	,000	,260	to
in	-,321	**,438**	,056	and
po	,250	-,079	-,358	after
pri	,115	-,009	**,525**	at, by
proti	-,275	,227	-,311	against
si	,293	-,054	**-,563**	you (are)
bo	,271	**,555**	,124	will be
v	**-,516**	,246	,183	in
iz	-,176	,258	,348	from
s	-,168	,003	-,033	with
med	-,248	,266	**,540**	between
cez	-,347	,131	-,353	over
ko	-,091	-,386	,021	when
kakor	,265	-,360	-,181	like
kako	**,489**	,136	-,160	how
ker	,002	-,076	-,092	because
z	-,296	,066	-,084	with
pred	-,360	**,441**	,136	before
jaz	**,421**	,300	-,061	me
do	-,006	**,490**	-,185	until
pa	**,427**	**,553**	,237	yet
ti	,307	,357	-,351	you
to	,394	-,245	,030	this
ga	,007	-,325	,353	him
mu	,175	**-,484**	-,072	him
bi	**,629**	,258	-,139	would
ni	,081	**-,582**	,260	is not
kaj	**,610**	-,070	,034	what
za	,272	-,190	-,032	for
vse	,093	-,051	,242	all
se	-,154	,150	-,327	is
tudi	,327	,170	,359	too
od	-,393	,264	-,009	from
ravno	**,480**	,064	-,141	exactly
na	-,314	,130	-,237	on
o	,097	,068	**,430**	about

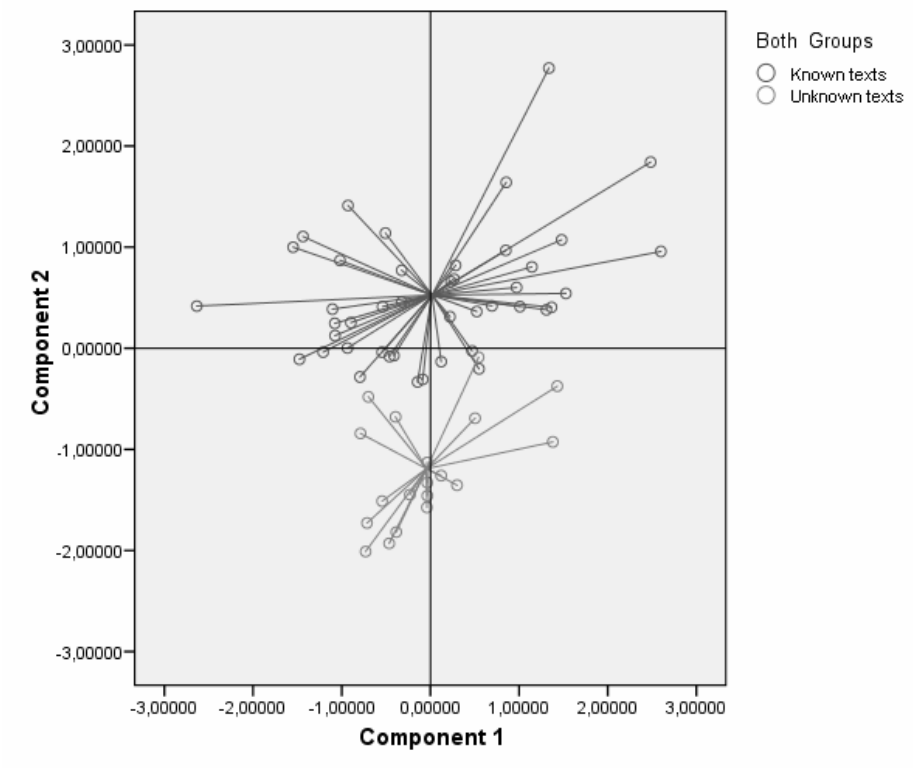


Figure 2: Scatter plot of the first and the second component.

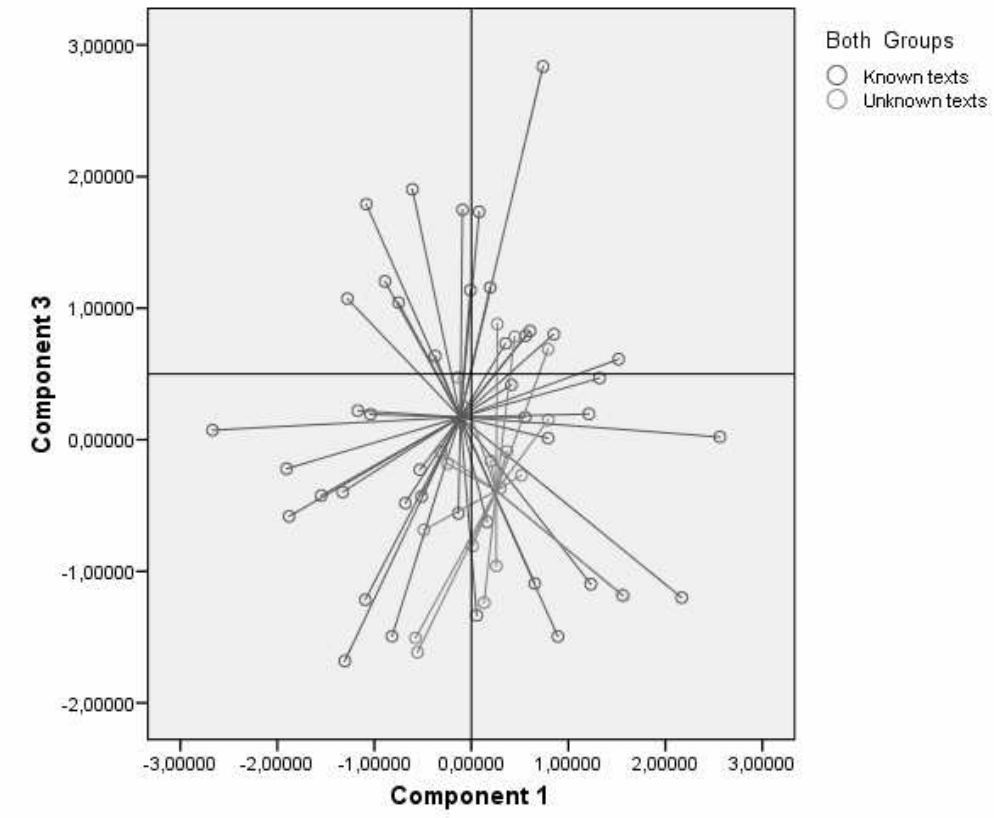


Figure 3: Scatter plot of the first and the third component.

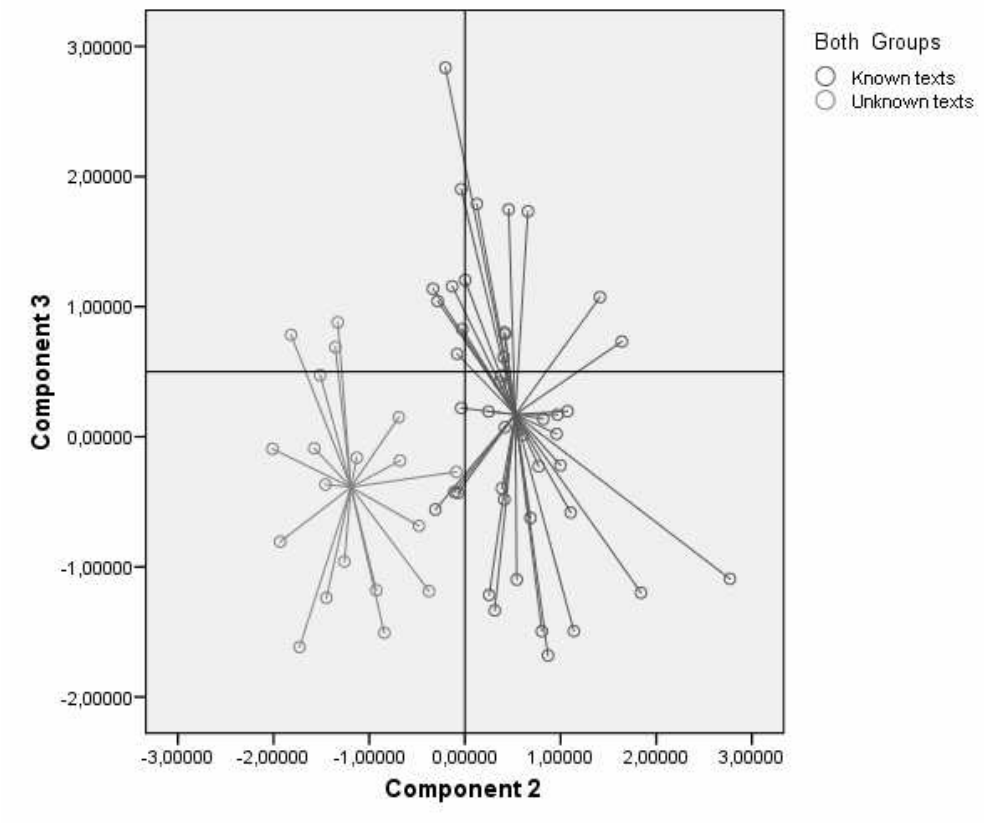


Figure 4: Scatter plot of the second and the third component.

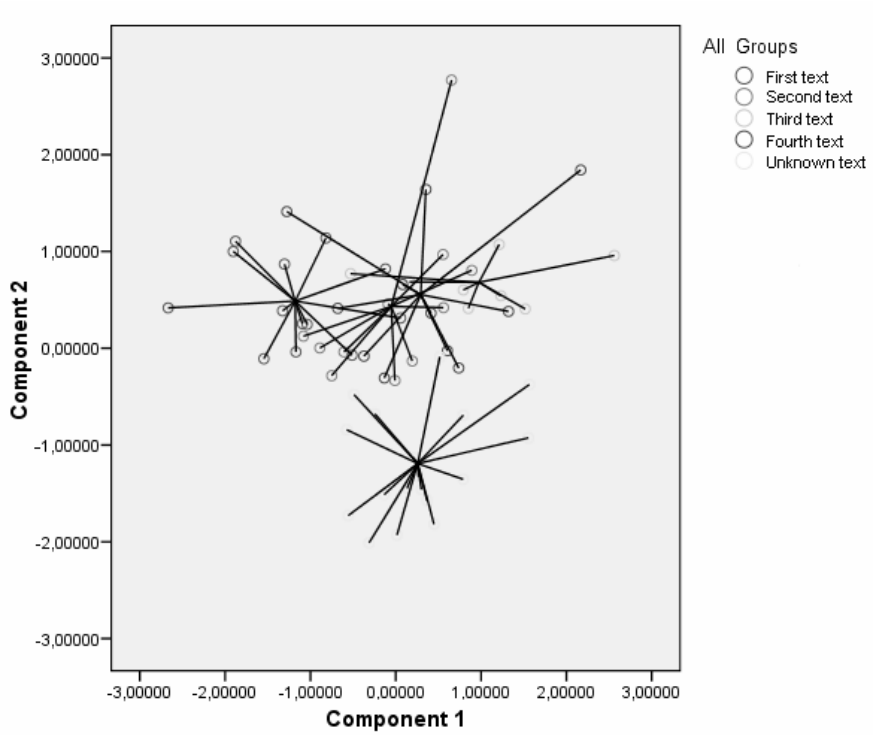


Figure 5: Similarity among Mencinger's stories and difference between both groups.

The component loading matrix shows which words correlate with the components the most. Those with the absolute value greater than 0,4 are in bold type. By looking at scattergrams of the first and the second loading component, the first and the third loading component and the second and the third loading component where each unit (block of a text) is labeled by »known« and »unknown« author, it is evident that the second component clearly divides the units into two groups, where the larger upper group represents four texts by Mencinger and the smaller lower group represents »Poštena Bohinčeka«. It can be concluded that each group was written by a different author, also taking into consideration very distinctive centroids.

When drawing each story separately, it can be seen that Mencinger's four stories are quite interlaced, whereas »Poštena Bohinčeka« differs from them. The same interlacement can be observed in 3D perspective.

4.3 t-test

The t-test is used to test the hypothesis that the means of two groups are equal or in other words that both groups are similar to each other. However performing t-test on two groups that are not similar not only confirms the existence of significant differences between both groups but also points out the single variables, that distinguish groups the most. The first group is represented by known words and the second by unknown words.

When performing the t-test, as shown in Appendix 2, 16 out of 50 variables made the distinction between groups. These variables are: **ne, ki, je, in, bo, v, iz, med, kakor, pred, nič, pa, brez, ni, kadar, nihče.**

This is a relatively sufficient proof that there is a statistical difference between means of both groups of texts, which confirms the hypothesis. Now the last step is performing another test with a discriminant analysis by using these t-test-identified variables, as well as variables suggested by the principal components analysis.

4.4 Discriminant analysis

Discriminant analysis is usually used to find linear combinations of variables, that would distinguish predefined classes. Here it is used mainly to confirm that two sets of words, known words and unknown words, are different. Coefficients of linear combinations will of course also be set.

As indicated, two discriminant analyses are performed. The results of the principal component analysis (on second component) suggest the distinguishable variables **ne, je, in, bo, pred, do, pa, mu, ni**, which distinguish the “known” and “unknown” texts the most. Therefore, these words are used for the first discriminant analysis, D_1. The results of the t-test suggest the variables **ne, ki, je,**

in, bo, v, iz, med, kakor, pred, nič, pa, brez, ni, kadar, nihče, which are used for the second discriminant analysis, D_2. The result is as follows: there is a difference between groups according to both analysis, it also shows that the group of variables, obtained with the t-test, is a better distinguisher. These variables classify original cases 100%, whereas the PCA classify group classifies the cases only 96,8%. Wilks' Lambda is also much higher with PCA group.

Table 3: Discriminant analysis: standardised loadings of the discriminant variables, % of correctly classified, Wilks' Lambda, X^2, significant level.

	D_1	D_2	
ne	,146	,367	no
ki	/	,660	which
je	,181	-,419	is
in	,405	,107	and
bo	,248	,034	will be
v	/	,317	in
iz	/	,169	from
med	/	,153	between
kakor	/	-,307	like
pred	,131	,108	before
nic	/	-,328	no
do	-,050	/	until
pa	,871	,855	yet
brez	/	,182	without
mu	-,450	/	him
ni	-,373	,013	not
kadar	/	-,574	when
nihce	/	,125	noone
% correctly classified	**96,8%**	**100%**	
Wilks' Lambda	**0,324**	**0,134**	
X^2	**63,696 (9)**	**106,659 (16)**	
Sig	**,000**	**,000**	

5 Conclusions

Are the authors of the analysed stories really different? The statistical results obtained by four different approaches confirm the hypothesis that the unknown author is not Mencinger. Furthermore, the variables that distinguish the stories the most have been identified. It must be emphasised that other criteria such as the historical time of writing, the theme of the stories and the literary style do not

differ and cannot influence the results obtained. We have thus managed to show, using four statistical approaches, that Janez Mencinger is not the author of "Poštena Bohinčeka", and it would be interesting to see who is!

Acknowledgements

The author would like to thank PhD Prof. Miran Hladnik, UL FF, for his contribution of texts and advice in the literary field and to PhD Prof. Anuška Ferligoj, UL FDV, for her extensive help with the statistical methods and other advices. There are also some other experts in the field who have shown interest in the project.

References

[1] Binongo, J.N.G. (2003): Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance*, **16,** 9-17.

[2] Dabagh, R.M. (2007): Authorship attribution and statistical text analysis. *Metodološki zvezki*, **4**, 149-163.

[3] Dović, M. (2002): Podbevšek in Cvelbar: Poskus empirične preverbe namigov o plagiatorstvu. *Slavistična revija*, **50**, 233-249.

[4] Holmes, D.I. (1991): A stylometric analysis of mormon scripture and related texts. *J.R. Statist.*, **155**, 91-120.

[5] Jakopin, P. (2003): Nizkoentropijski jezikovni model na besedilih Cirila Kosmača in Ivana Cankarja. *Slovenski roman*, **21**, 421-428.

Appendix

1 Independent Samples Test

Table 4: Means, standard deviation, t, sig. of known and unknown variables.

WORDS		MEANS		STANDARD DEV.			
		known	unknown	known	unknown	t	Sig.
ne	no	12,39	7,53	3,558	3,935	4,819	,000
ki	which	6,55	4,37	3,209	2,773	2,569	,013
le	only	2,52	2,32	1,861	1,057	,453	,652
tako	such	4,8	3,74	2,426	1,821	1,703	,094
da	that	16	13,47	5,532	3,323	1,847	,070
je	is	44,39	63	13,464	16,111	-4,743	,000
naj	should	1,77	1,79	1,508	1,512	-,040	,968
ali	or	2,89	3,95	2,137	1,985	-1,846	,070
kar	just	3	2,53	1,88	1,264	1,002	,320
k	to	2,11	2,11	1,498	1,729	,019	,985
in	and	39,05	31,74	8,488	5,425	3,452	,001
po	after	5,16	6	2,623	2,809	-1,143	,257
pri	at/by	2,82	2,42	1,756	2,063	,781	,438
proti	against	1,27	0,89	1,468	1,049	1,014	,315
si	(you) are	4,09	5,47	3,588	2,458	-1,529	,132
bo	will be	5,2	2,47	3,593	1,712	3,151	,003
v	in	17,8	14,79	4,892	4,131	2,340	,023
iz	from	4	2,53	2,035	1,124	2,959	,004
s	with	2,93	3,11	1,576	2,424	-,338	,736
med	between	2,02	0,89	1,406	0,875	3,229	,002
cez	over	1,48	1,32	1,592	1,157	,398	,692
ko	when	2,95	3,68	2,09	2,11	-1,268	,210
kakor	like	6,09	8,74	2,311	2,903	-3,855	,000
kako	how	3	2,26	2,323	1,695	1,245	,218
ker	because	3,86	3,84	2,174	2,911	,032	,974
z	with	4,77	4,89	2,666	2,961	-,161	,872
pred	before	2,64	1,16	1,63	1,5	3,382	,001
jaz	me	1,68	1,37	1,653	1,383	,723	,472
nic	nothing	0,93	2,26	1,404	1,695	-3,242	,002
do	until	1,98	1,32	1,517	1,108	1,710	,092
pa	yet	17,18	5,95	5,978	3,749	7,556	,000
ti	you	2,64	1,95	2,373	1,715	1,141	,258
to	this	3,36	3,89	2,114	2,331	-,887	,378
ga	him	6,09	6,63	3,388	2,91	-,605	,547
brez	without	1,23	0,47	1,309	0,697	2,361	,021
mu	him	5,89	8,26	3,604	3,364	-2,450	,017
bi	would	9,45	7,68	4,752	2,75	1,514	,135
ni	is not	7,16	10,53	3,154	4,948	-3,251	,002
kaj	what	3,77	4	3,256	2,494	-,271	,787
kadar	when	0,07	0,68	0,255	0,885	-4,263	,000
za	for	4,09	4,74	2,089	2,446	-1,069	,289
nihce	noone	0,57	0	0,974	0	2,531	,014
vse	all	3,41	4,05	2,213	1,957	-1,095	,278
preden	before	0,3	0,26	0,632	0,452	,201	,841
se	is	24,27	23,42	5,302	5,326	,584	,561
tudi	too	6	4,84	2,988	2,588	1,467	,148
od	from	3,98	3,26	2,758	1,91	1,025	,309
ravno	exactly	1,64	1,11	1,844	1,049	1,173	,245
na	on	10,41	9,89	3,694	4,067	,492	,624
o	about	1,37	0,74	1,662	0,933	1,553	,126

2 Scree plot

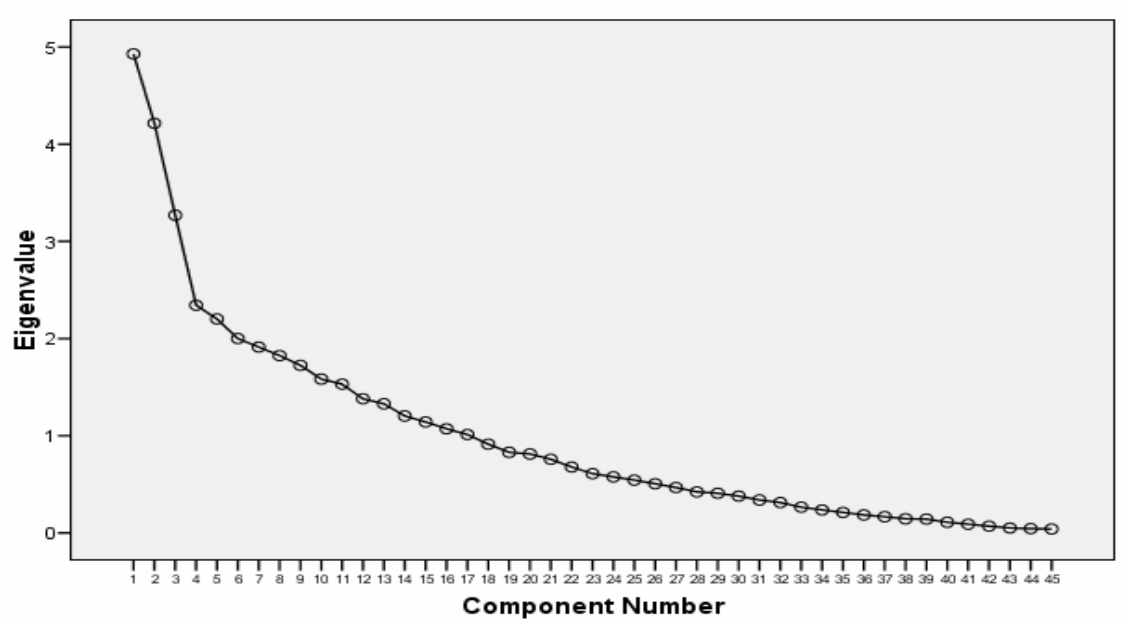


Figure 6: Scree plot.

3 Frequency table and histogram of variable “ne”

To illustrate the normal distribution of variables a histogram of variable “ne” has been added. The curve on the histogram represents continuous normal distribution and the columns represent discrete distribution of variable “ne”. It can be seen that the heights of the columns try to follow the normal curve and so the conclusion is as expected that the distribution of one variable is more or less normal.

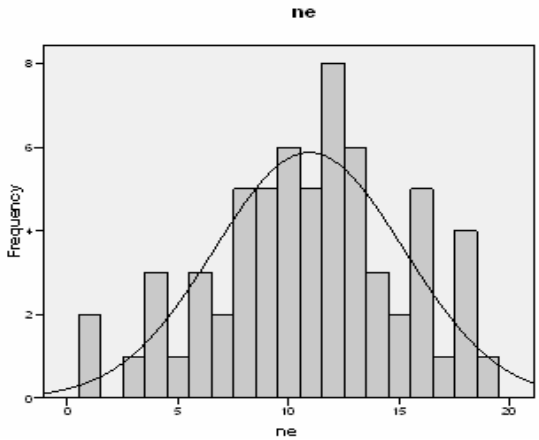


Figure 7: Histogram showing normal distribution of variable “ne”.