CSÉVE ANNA[1] – KALCSÓ GYULA[1,2] – MIHÁLY ESZTER[1]

STYLOMETRIC ANALYSIS OF THE CORRESPONDENCE OF ZSIGMOND MÓRICZ[1]

[1]Digital Humanities Centre, Petőfi Literary Museum
[2]Department of Hungarian Linguistics, Eszterházy Károly Catholic University

1. Goals

Computational stylometric analysis is most often used to identify authorship (Grieve 2007; Eder et al. 2016), but it can also be used to study the textual similarity or difference between texts from different periods in an author's life, between the works of the same author, etc. (Pennebaker–Stone 2003; Can–Patton 2004; Lancashire–Hurst 2018; Gomez-Adorno et al. 2018). The most commonly used method for measuring stylistic similarity is distance measurement (for a historical overview, see Moisl 2015), which is usually carried out with specific software. One of the most commonly used tools in stylometric analysis is the statistical software and programming language R and its various components (so-called packages). There is a package specifically developed for stylometric analysis called Stylo (Eder et al. 2016). The main purpose of this paper is to study stylometric differences between the letters written by Zsigmond Móricz to his wife, Janka, and other recipients between 1902 and 1913 by using the Stylo R package's distance measuring methods (Eder et al. 2017) with the toolkit Szöveglabor (see below). The following are also the first stylometric experiments of the Digital Humanities Centre of Petőfi Literary Museum. Our hypothesis is that there are stylometrically detectable differences between the letters to Janka and the others. Our further goal is to find the best stylometric method to use for the text classification problems of letters.

2. The corpus

Within the projects of the Digital Humanities Centre of the Petőfi Literary Museum, various computer linguistic researches are also carried out in parallel with the work on digital text editions. As a special project, a stylistometric analysis of the texts forming the basis of the digital critical edition of Zsigmond Móricz's

[1] Special thanks go to my colleagues at Petőfi Literary Museum Digital Humanities Centre, especially to Eszter Simon, and the so-called Móricz task force: Csilla T. Makkai, Emese Varga, Bence Vétek, Mátyás S. Török.

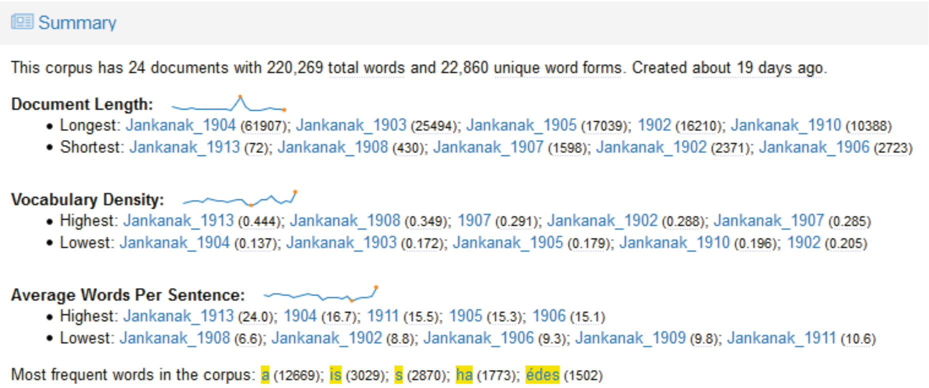
https://doi.org/10.46437/ActaUnivEszterhazyLinguistica.2021.137

correspondence has been started, which aims to confirm previous assumptions and to develop new research aspects and to analyse their results.

A unique achievement in the field of basic research on critical editions is the publication of 1500 letters from Zsigmond Móricz's previously unexplored correspondence from the period 1892–1913. The digital form in which this publication was produced makes it possible for the grammatical and syntactical features of his language to be the subject of a multi-level scientific investigation. One possibility is to observe the changes in the fictional narrative identity of the letter-writer, which can be revealed by comparative linguistic-textological analyses of the correspondence corpus, grouped by period and addressee. What is new is that it takes the correspondence corpus out of the realm of source history and offers the possibility of exploring intratextual layers.

The corpus is based on the above mentioned digital scholarly editions of letters from the Móricz special collection of the Petőfi Literary Museum (Kómár et al. 2018). The complete (already annotated) digital corpus of letters written by Zsigmond Móricz between 1892 and 1913 contains 498 items, of which he wrote 259 to Janka. The latter date from 1902. Therefore, we have selected letters written to other recipients after 1902 to compare them with those written to Janka, so our corpus used in the present study contains 478 letters (259 written to Janka, 219 written to other recipients). It was easy to separate the letters, as the recipient is always encoded in TEI XML.

The digital scholarly editions contain TEI XML versions of the letters, however, for the stylometric analysis, we cleaned the tags off using a Python script.[2] The result is plain text, which is the input format of the Stylo package. We compiled all the Janka letters in one file year by year (from 1902 to 1913), and did the same with the other subcorpus. The corpus thus consists of 24 files with 220 269 words (tokens) and 22 860 unique word forms (types). The Janka subcorpus contains 135 528 tokens and 15 338 types, the other one has 84 741 tokens and 11 619 types.[3]



Summary

This corpus has 24 documents with 220,269 total words and 22,860 unique word forms. Created about 19 days ago.

Document Length:
- Longest: Jankanak_1904 (61907); Jankanak_1903 (25494); Jankanak_1905 (17039); 1902 (16210); Jankanak_1910 (10388)
- Shortest: Jankanak_1913 (72); Jankanak_1908 (430); Jankanak_1907 (1598); Jankanak_1902 (2371); Jankanak_1906 (2723)

Vocabulary Density:
- Highest: Jankanak_1913 (0.444); Jankanak_1908 (0.349); 1907 (0.291); Jankanak_1902 (0.288); Jankanak_1907 (0.285)
- Lowest: Jankanak_1904 (0.137); Jankanak_1903 (0.172); Jankanak_1905 (0.179); Jankanak_1910 (0.196); 1902 (0.205)

Average Words Per Sentence:
- Highest: Jankanak_1913 (24.0); 1904 (16.7); 1911 (15.5); 1905 (15.3); 1906 (15.1)
- Lowest: Jankanak_1908 (6.6); Jankanak_1902 (8.8); Jankanak_1906 (9.3); Jankanak_1909 (9.8); Jankanak_1911 (10.6)

Most frequent words in the corpus: a (12669); is (3029); s (2870); ha (1773); édes (1502)

Fig 1.: Basic statistic data of the corpus in Voyant Tools (W1)

[2] For the Python script special thanks to Iván Mittelholcz.

[3] The corpus can be downloaded from https://pim.hu/hu/node/2193867.

3. Methods and tools

3.1. Distance measuring methods and the Vector Space Model.

Most distance measuring methods are based on the Vector Space Model (VSM). Gomez-Adorno et al. (2018) list the textual and stylometric features used in VSM to study the similarity between texts. “In authorship analysis, typical features used for text representation in the Vector Space Model (VSM) are words, Bag of Words (BoW) model, word n-grams, character n-grams, and syntactic n-grams. The values of these features can be Boolean, tf-idf (term frequency-inverse document frequency), weights, or values based on probabilistic models. Another popular statistical-based text representation are the stylometric features, such as length of sentences, complexity of sentences, frequent words, spelling errors, etc.” (Gomez-Adorno et al. 2018: 47.) The basic idea of the model is to compute multidimensional vector coordinates in vector space based on the matrix of values computed from the textual features above, and then compare their distances. Multidimensional vector space values are usually projected into two dimensions. The resulting feature vectors form the basis for quantifying the overall similarity of texts using a variety of methods.

The following figure illustrates how three different types of distance measures of the VSM between texts A and B in terms of a two-dimensional VS of the words *and* and *the* are calculated.

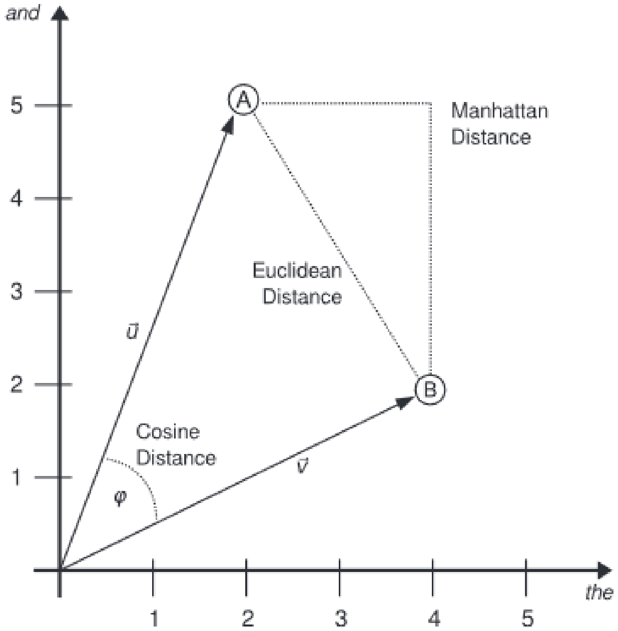


Fig 2.: Different vector distances between two example documents A and B (in terms of the words *and* and *the*) illustrated in two-dimensional space (Source: Evert et al. 2017: ii7)

3.2. Delta distances

A concise summary of the metrics used in stylometry with their formulas and an experimental comparative study are given in (Stanikknas et al. 2017). Evert et al. 2017 also list the historically evolved types of the most prominent calculation method, the so-called delta distance, from Burrows classic distance measurement (Burrows 2002) to Argamon's (Argamon 2008) and Eder's (simple) delta (Eder et al. 2016). All types of delta distance measuring methods rely on the relative frequencies of the most frequent words. This is the best performing method for authorship attribution. As noted by Evert et al. 2017, "The information particularly relevant to the identification of the author of a text lies in the profile of deviation across the most frequent words rather than in the extent of the deviation or in the deviation of specific words only" (Evert et al. 2017: ii4). When we use it to separate the works of an author written in different styles, we have to do the opposite of authorship attribution, we have to prove that the texts are classified in different clusters. The use of the delta distance measure in the study of Móricz's letters is based on the assumption that the most frequent word bigrams in letters to Janka are different from those in letters to others.

3.3. Text Lab in PLM DHC

The Petőfi Literary Museum's Digital Humanities Centre is developing a tool for text mining and text analysis, called Szöveglabor (Text Lab), which includes a stylometric toolkit. This toolkit is based on the Stylo package and contains several kinds of metrics used in stylometric research. It is suitable for carrying out delta distance measurement experiments and visualising the results. Experimental investigations described in this article were performed with this tool.

4. Experimental results

4.1. Stylo settings

When the Stylo package was set up, some values were given, but in some cases, multiple values were tested to ensure the best results. The text format was plain text and the language was Hungarian in all cases of course (Input & language settings of Stylo). Features and Statistics settings were some of the most important ones.

4.1.1. A given value in Features was to study word bigrams, because studying word constructions (word groups and phrases) instead of unigrams (words) or character n-grams are more common and effective in literary investigations. To choose a value for n in an n-gram model, it is necessary to find the right compromise between the stability of the estimate against its appropriateness. Trigrams are a common choice with large training corpora (millions of words), whereas bigrams are often used with smaller ones (like ours).

4.1.2. We kept Stylo's default setting for lower case in all our investigations. Although it would have been interesting to keep capital letters for proper names, we decided against it because of the distorting effect of capital letters at the beginning of sentences (Stylo cannot differentiate between the two cases).

4.1.3. The Most Frequent Words (MFW) settings are particularly important in stylometry. According to Stanikknas et al. 2017, when using the delta distance (see above), an MFW value between 1000 and 5000 seems to be the most appropriate (Stanikknas et al. 2017: 3–4., esp. Figure 1–2). Therefore, we decided to set the min. value to 1000. Finding the upper value was more difficult, but after several attempts, a number around 3000 seemed to give the most reliable result. An increment between 1000 and 3000 was set to 50 to obtain a finer dataset by several analyses. A cut was made in the MFW settings, the 14 most common word bigrams were omitted from the test (start at frequency rank value was set to 15), because the more frequent multi-word phrases were not as stylistically distinctive. Moving away from this (for lower or higher values), the disorder in the results gradually increased.

4.1.4. Due to the nature of our study (we wanted to find differences, not similarities), we decided not to set culling values (technically it is 0), which means that none of the word bigrams were filtered out from texts. If this value is set to 20, a given feature (word bigram in this case) has to appear in at least 20% of the texts, in the case of 40 in 40%, etc. so as not to be filtered out (see Eder et al. 2016: 111). We wanted to keep all the word bigrams so that the differences would come out better.

4.1.5. As mentioned above, we decided to use the delta distance in this experiment. Two types of it were applied, the so-called classic delta (Burrows' delta) and Eder's Simple delta. As reported by Stanikknas et al. 2017, both performed well on quite similar size corpora as ours.

4.2. Cluster analysis

Cluster analysis (dendrogram) was used first on the dataset in order to visualize the main clusters of the letters. The results are shown in the two figures (Fig. 3–4) below.

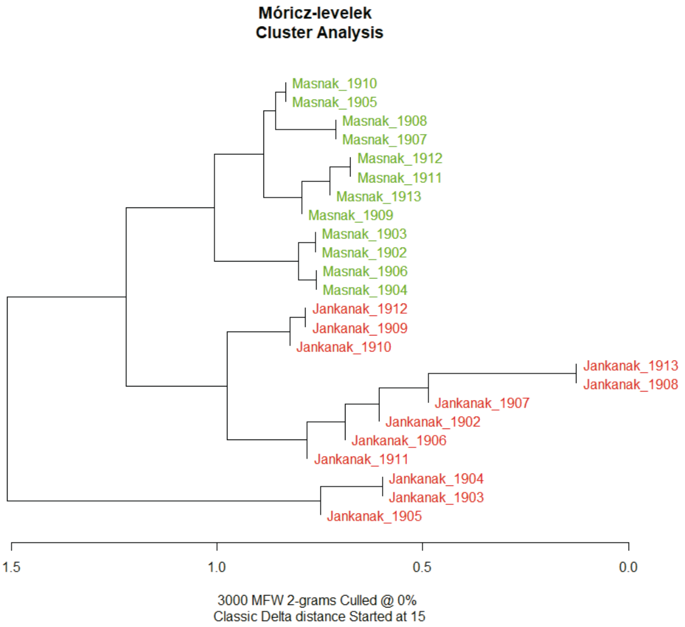


Fig 3.: Dendrogram of Móricz letters, based on classic delta (Burrows' delta) distance calculated on the 3000 most frequent word bigrams

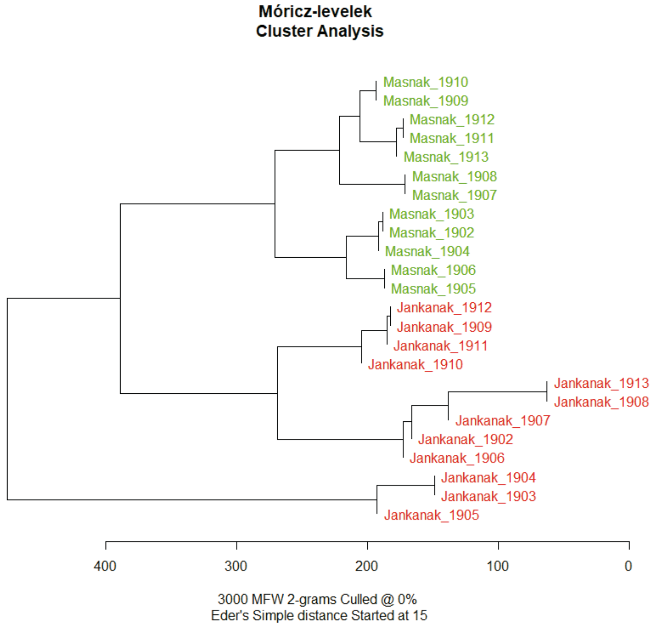


Fig 3.: Dendrogram of Móricz letters, based on Eder's simple delta distance calculated on the 3000 most frequent word bigrams

As can be seen from the figures, the application of the two distance measures resulted in differences only for some groups of letters. Two large groups emerge, but it is not simply the letters to Janka that are separated from those not written to her, but one group of letters (written between 1903 and 1905) to Janka from the others. It is worth noting that 1905 was the year of their marriage. In both tests, however, the letters written to Janka on the other branch were separate from the others. This seems to indicate that there is a stylometrically measurable difference between the letters to Janka and the others.

4.2.1. In the case of the letters to Janka, the difference between the two deltas was limited to a single group of texts, the letters written in 1911. According to the classical delta, they are closer to those written in 1906, while according to Eder's simple delta, they are closer to those written in 1910 (and 1909, 1912). The letters written before and after their marriage are not separated clearly, because those written at the beginning of their correspondence in 1902 are clustered with those written after their marriage.

4.2.2. In the case of letters to others, those written in 1905, measured by Eder's simple delta, were classified with those written between 1902 and 1906, while measured by classic delta, they were classified with different ones. Those written in 1909 are also classified in a chronologically more correct place when measured by Eder's delta.

4.3. Principal Components Analysis

Another visualization method to classify the Móricz letters was Principal Components Analysis (PCA), which is a popular stylometric identification technique. PCA's ability to capture essential variance across large amounts of features in a reduced dimensionality makes it attractive for text analysis problems, which typically involve larger feature sets (so we probably need to increase the size of our corpus to obtain more reliable results in the future, e.g., by transcribing more Móricz letters from the period after 1913). The essence of PCA can be described as follows: given a feature matrix with each column representing a feature and instance vector rows for the various texts, project the matrix into a lower-dimensional space by plotting principal component scores (which are the product of the component weights and instance feature vectors). The similarity between texts can be compared based on visual proximity of patterns (Kjell et al. 1994) or computation of average distance (Abbasi and Chen 2006, 2007).

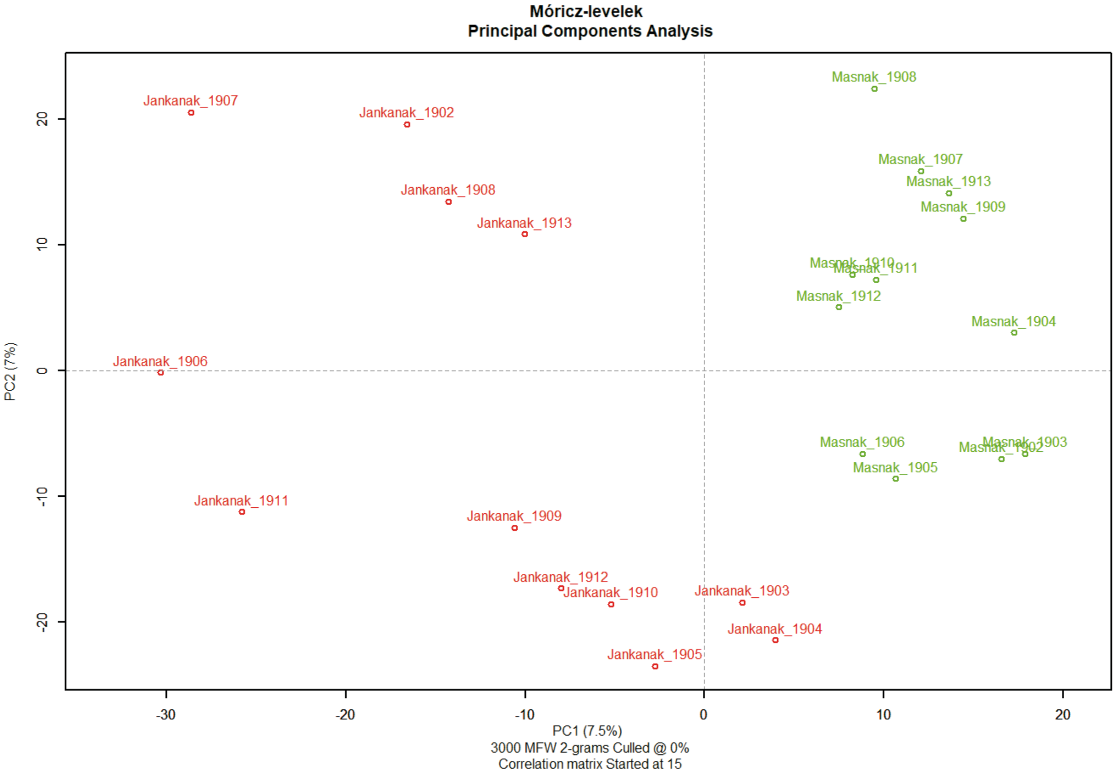


Fig 4.: Principal Components Analysis of Móricz letters, based on classic delta (Burrow's delta) distance calculated on the correlation matrix of 3000 most frequent word bigrams

Two types of PCA can be applied by Stylo, one based on the correlation matrix, and one based on the covariance matrix (we used the former). There is no need to present the PCA of both deltas, because the two are actually identical. In these figures, the clusters can be separated somewhat differently, but it is clear to see why the cluster analysis of the two types of delta resulted in differences. The letters to other recipients are clearly able to be distinguished from the others. The letters written to Janka in 1906 and 1911 are somewhat separate from the others. This is why the ones written in 1911 may have been classified in different places by the two types of delta when cluster analysis was applied. For letters written to other recipients in 1905, however, the situation is different, with the PCA clearly showing them as close to those written between 1902 and 1906. The PCA also shows a much smaller distance for letters written in 1909 compared to the other letter groups, probably causing the uncertainties in the classification by the deltas.

5. Conclusions and future work

5.1. Conclusions

Our main conclusion is that Eder's simple delta seems to be more suitable than the classical delta for stylistic text classification problems. Eder's simple delta for bigrams gives more accurate results in cluster analysis. Experiments have also been carried out with cosine distance and Manhattan distance, but with far worse results than the classic and Eder's Simple deltas.

Another important conclusion is that the visualisation method of distance measurement is a key factor in the evaluation of the results. The differences between the dendrograms plotted on the basis of the two deltas were not visible until PCA was performed. However, it is also clear that the separation of the branches of the dendrograms is based on different principles than the plotting of the PCA coordinate system, and therefore leads to slightly different results. For most classification problems, a combination of visualisation methods will likely be appropriate.

5.2. Future works

In most classification studies, the size of the available corpus is crucial. It is quite clear that if we increase the size of our corpus we can get even better results, although the results supporting our hypothesis are already apparent from the corpus we have so far. A good way of expanding the corpus might be to include other works by Móricz (e.g., his novels).

It also seems to be a good idea to analyse other textual and stylometric features, e.g. syntactic n-grams, which requires syntactic analysis of the corpus though. It would also be worth examining the letters using some kind of machine learning algorithm (e.g., automatic text classification methods), but it also requires an increase in the size of the corpus. However, these topics should be the subject of another paper.

WEB SOURCES

W1 = https://voyant-tools.org/?corpus=7ca70f930f575a021120b1cfcbfa3cdc&view=Summary (2021. 08.11.)

REFERENCES

Abbasi, Ahmed – Hsinchun, Chen 2006. Visualizing authorship for identification. In Sharad Mehrotra – Daniel D. Zeng – Hsinchun Chen – Bhavani Thuraisingham – Fei-Yue Wang (szerk.): *Intelligence and Security Informatics*: *Lecture Notes in Computer Science (3975)*. Berlin – Heidelberg: Springer. 60–71. https://doi.org/10.1007/11760146_6

Abbasi, Ahmed – Hsinchun, Chen 2007. A framework for stylometric similarity detection in online settings. In: *AMCIS 2007 Proceedings*. 127. http://aisel.aisnet.org/amcis2007/127 (2021. 07. 14.)

Argamon, Shlomo 2007. Interpreting Burrows's Delta: geometric and probabilistic foundations. *Literary and Linguistic Computing* 23: 131–47. https://doi.org/10.1093/llc/fqn003

Burrows, John 2002. „Delta": A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing* 17: 267–87. https://doi.org/10.1093/llc/17.3.267

Eder, Maciej – Piasecki, Maciej – Walkowiak,Tomasz 2017. An open stylometric system based on multilevel text analysis. *Cognitive Studies | Études cognitives* 17. https://ispan.waw.pl/journals/index.php/cs-ec/article/view/cs.1430 (2021. 07. 14.) https://doi.org/10.11649/cs.1430

Eder, Maciej – Rybicki, Jan –Kestemont, Mike 2016. Stylometry with R: a package for computational text analysis. *The R Journal* 8: 107–21. https://doi.org/10.32614/RJ-2016-007

Fazli, Can – Patton, Jon M. 2004. Change of writing style with time. *Computers and the Humanities* 38: 61–82. https://doi.org/10.1023/B:CHUM.0000009225.28847.77

Gómez-Adorno, Helena Montserrat – Ríos-Toledo, Germán – Posadas-Durán, Juan-Pablo – Sidorov, Grigori – Sierra, Gerardo 2018. Stylometry-based approach for detecting writing style changes in literary texts. *Computación y Sistemas* 22: 47–53. https://doi.org/10.13053/cys-22-1-2882

Grieve, Jack 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22: 251–270. https://doi.org/10.1093/llc/fqm020

Kjell, Bradley – Woods, W. Addison – Frieder, Ophir 1994. Discrimination of authorship using visualization. *Information Processing & Management* 30: 141–150. https://doi.org/10.1016/0306-4573(94)90029-9

Kómár, Éva – Cséve, Anna – Fellegi, Zsófia 2018. Móricz Zsigmond levelezésének (1892–1913) digitális kritikai kiadása. [A digital critical edition of Zsigmond Móricz's correspondence, 1892–1913] *Digitális Bölcsészet* 1: 159–74. https://doi.org/10.31400/dh-hun.2018.1.227

Lancashire, Ian – Hirst, Graeme 2009. Vocabulary changes in Agatha Christie's mysteries as an indication of dementia: a case study. *19Th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*. https://www.cs.toronto.edu/pub/gh/Lancashire+Hirst-2009-poster.pdf (2021. 07. 14.)

Moisl, Hermann 2015. *Cluster analysis for corpus linguistics.* H. n.: De Gruyter Mouton. https://doi.org/10.1515/9783110363814

Pennebaker, James W. – Stone, Lori D. 2003. Words of wisdom: language use over the life span. *Journal of Personality and Social Psychology* 85: 291–301.
https://doi.org/10.1037/0022-3514.85.2.291

Stanikknas, Daumantas – Mandravickaite, Justina – Krilavicius, Tomas 2017. Comparison of distance and similarity measures for stylometric analysis of Lithuanian texts. In: *CEUR Workshop proceedings* [electronic resource]: *ICYRIME 2017: proceedings of the symposium for young researchers in informatics,* *mathematics and engineering.* Kaunas, Lithuania, April 28, 2017. Aachen: CEUR-WS. 1–7. http://ceur-ws.org/Vol-1852/p01.pdf (2021. 07. 14.)

Móricz Zsigmond levelezésének stilometriai elemzése

Jelen cikk egy kutatásról számol be, amelynek keretében számítógépes stilometriai módszerekkel vizsgáltuk meg Móricz Zsigmond feleségéhez és másokhoz 1902 és 1913 között írt leveleinek textuális és stilometriai sajátosságait. Ez a kísérlet a Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központjának az első stilometriai próbálkozása. A korpusz a Petőfi Irodalmi Múzeum Móricz-különgyűjteményének leveleiből készült digitális tudományos kiadásán alapul, 478 levelet (220 268 szót) tartalmaz. Egy R-csomagot, a Stylót, valamint távolságmérési módszereket (klasszikus deltát és Eder egyszerű deltáját) alkalmaztunk a fent említett sajátosságok elemzésére. Az eredményeket kétféleképpen vizualizáltuk: klaszteranalízissel (dendrogramon) és főkomponens-analízissel. A levelek klasszifikációja sikeres volt, bár csak a két vizualizációs módszer együttes alkalmazása vezetett eredményre. Sikerült kimutatnunk, hogy stilometriailag mérhető különbségek vannak a Jankának és másoknak írt Móricz-levelek között.