What Makes Your Writing Style Unique? Significant Differences Between Two Famous Romanian Orators

Mihai Dascalu[1(✉)], Daniela Gîfu[2], and Stefan Trausan-Matu[1]

[1] Computer Science Department, University Politehnica of Bucharest, 313 Splaiul Independenței, 060042, București, Romania
{mihai.dascalu, stefan.trausan}@cs.pub.ro

[2] Faculty of Computer Science, "Alexandru Ioan Cuza" University, 16 General Berthelot, 700483 Iași, Romania
daniela.gifu@info.uaic.ro

Abstract. This paper introduces a novel, in-depth approach of analyzing the differences in writing style between two famous Romanian orators, based on automated textual complexity indices for Romanian language. The considered authors are: (a) Mihai Eminescu, Romania's national poet and a remarkable journalist of his time, and (b) Ion C. Brătianu, one of the most important Romanian politicians from the middle of the 18th century. Both orators have a common journalistic interest consisting in their desire to spread the word about political issues in Romania via the printing press, the most important public voice at that time. In addition, both authors exhibit writing style particularities, and our aim is to explore these differences through our *ReaderBench* framework that computes a wide range of lexical and semantic textual complexity indices for Romanian and other languages. The used corpus contains two collections of speeches for each orator that cover the period 1857–1880. The results of this study highlight the lexical and cohesive textual complexity indices that reflect very well the differences in writing style, measures relying on Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) semantic models.

Keywords: Writing style · Textual complexity for Romanian language · Comparable corpora · Famous orators

1 Introduction

Automated evaluation of writing styles represents a challenge among linguistics and emphasizes the importance of technology in order to facilitate research on language. In addition, quantifying differences between speeches in different languages and between authors has become a trending topic in the field of Natural Language Processing (NLP). The equivalent manual analysis is an extremely time consuming process that requires highly skilled annotators, especially in linguistics. Prior research has proposed methods for creating sets of comparable corpora [1–5] that contain similar texts across multiple languages, genres and authors, which can later on be used to assess linguistic differences.

The novelty of this study is reflected in its focus to compare emblematic writing styles of two Romanian authors who marked our society and its transition towards a more transparent system. Vianu, a famous Romanian literary critic, highlighted the problem of theoretical stylistics, namely the dual intention within language: to communicate and to reflect [6]. The expressiveness found in different variations of writing styles makes its presence felt in the reflective dimension of communication. According to Coteanu [7], a word's expressiveness is latent and can be inferred from context, thus emphasizing the importance of cohesion in terms of discourse representation. This is the case of the particularly expressive journalistic texts pertaining to the two selected authors: *Mihai Eminescu* and *Ion C. Brătianu*, whose speeches are representative for Romanian oratory.

Mihai Eminescu, known as Romania's national poet, was instructed in Vienna and is one of the most important journalistic voices of his time. In fact, we speak of three different journalistic stages [8]: (a) the first stage, until his entry in the *Junimea* society in 1876; (b) the *Iasi period* that includes his work for the *Curierul de Iaşi* publication; (c) his activity for the *Timpul* newspaper between 1877 and 1883. During the previous stages, there are no concrete differences of structure or writing style, but rather tones, thickenings or blurs in his political ideas. *Ion C. Brătianu* [9, 10] was an important Romanian politician, instructed in Paris, whose main concern was to draw the attention of political circles in France to support the cause of Romanians and their national aspirations. The considered collection of journalistic texts coincides with the period when he returned from exile, after nine years, while being involved as the Minister of several Ministries (Finances, Internal Affairs, or War) within the Romanian United Principalities (Moldavia and Wallachia). Furthermore, in 1875 he laid the foundations of the Liberal National Party. Brătianu's speech could be easily recognized due to his preeminent ideas of national consciousness and democratic values, such as individual freedom and social equality.

In order to explore the oratorical styles of both Romanian personalities, we rely on a previously validated textual complexity model, integrated in our *ReaderBench* framework [11–14], adapted for Romanian language [12], that addresses multiple facets of text difficulty and comprehension [11]: *text features* (e.g., length, structure or use of punctuation) [15], *textual formality* (e.g., vocabulary, slang, phrasal verbs, use of idiomatic language, and so on) [16], and *textual styles* (e.g., simple/complex sentences, stylistic markers, cohesion, etc.) [17]. The selected textual complexity indices, presented in detail later on, are reflective of each author's writing style and address different layers of discourse analysis, namely lexical structure and semantics, with emphasis on cohesion. Analyses of writing styles in Romanian language are not singular as they became constituent parts in the current trends of interpreting language facts [18–21], but this study represent a first automated in-depth comparison of famous Romanian speeches.

This paper is structured as follows: section two provides details on the used corpus and of the automated method employed through the *ReaderBench* framework. Section three presents results and corresponding discussions, while the last section highlights conclusions and future work.

2   Automated Assessment of Writing Style

In this section we present the analyzed corpus structured into 2 collections of texts by Eminescu and Brătianu, as well as the *textual complexity* indices from *ReaderBench* framework used to characterize each author's writing style.

2.1   Corpus Selection

Our linguistic analysis is focused on exploring the differences in writing styles between the two Romanian personalities from the 19th century (more specifically 1857–1880). This was a period in which Romania was becoming a well-defined nation in the European political context; thus, the speeches had a strong nationalist tone and the shared emotional load had a high impact on the population. Many of the texts were preceded by corresponding public speeches as oratory in public spaces was the best communication channel at the time. Our corpus was built starting from newspaper articles and contains around 139,000 lexical tokens (see Table 1). The articles were converted from PDF format into plain text using Optical Character Recognition software, followed by manual corrections on the raw texts.

Table 1. General corpus statistics.

Orators	Period	N docs	N words	Newspaper sources
M. Eminescu	1877–1880	65	80,193	Pressa, România liberă, Românul(u), Timpul
I. C. Brătianu	1857–1875	45	58,237	Românul, Monitorul Oficial
Total		*110*	*138,430*	

2.2   Indices of Writing Style

Three main categories of textual complexity indices computed by the *ReaderBench* framework were adapted for Romanian and are used to reflect specific traits of writing style for each orator [12]. First, at *surface analysis*, *ReaderBench* makes use of the proxes (i.e., computer approximations of text difficulty) initially developed by Page [22, 23]. Our model integrates the most representative and commonly used proxes in automated essay grading systems [23, 24], for example: average word/phrase/paragraph length in characters, average unique/content words (dictionary forms that are not stopwords) per phrase or paragraph, average number of commas per sentence or paragraph. Entropy, derived from Shannon's Information Theory [25, 26], is also a relevant metric for quantifying word or character diversity. While word entropy reflects a more varied vocabulary and is related to an increased working memory as more concepts are introduced to the reader, character entropy is a language specific characteristic [27].

Second, *semantic analysis* is centered on cohesion and represents the core of our model. According to McNamara et al. [28], textual complexity is strongly related to cohesion in terms of comprehension, due to the fact that the reader must create a

coherent mental representation of the underlying information (i.e., the situation model [29]). Thus, the lack of cohesion flow can increase the difficulty of a text [30] as readers can easily loose interest by finding text segments too unrelated one to another. In order to evaluate local and global cohesion, our model uses Cohesion Network Analysis (CNA) [31] to compute cohesion as the average semantic similarity [32, 33] at the following levels: intra-paragraph (between sentences of each paragraph), inter-paragraph (between any pair of paragraphs), or adjacency/transition from one paragraph or sentence to the next one. Cohesion between any two text segments is estimated as the average value of the cosine similarity in Latent Semantic Analysis (LSA) vector spaces [34, 35] and the inverse of the Jensen Shannon dissimilarity (JSD) [36] between Latent Dirichlet Allocation (LDA) topic distributions [37, 38]. Both models are based on the bag-of-words approach and reflect co-occurrence patterns from an initial training text corpora. For this study, LSA and LDA semantic models were trained on a Romanian corpus of more than 2 million content words covering journalistic texts, literature, politics, science and religion.

LSA uses a sparse term-document matrix that contains for each word a normalized number of its occurrences within a given document (for example, log-entropy, term frequency-inverse document frequency). The dimensionality of this matrix is reduced by projecting the resulting matrices from the Singular Value Decomposition (SVD) [39] on the most important k dimensions. Words and documents are compared using a cosine distance between their vector representations in the projected semantic space. LDA is a generative probabilistic model based on topic distributions. A topic is a Dirichlet distribution [40] over the vocabulary in which thematically related concepts are grouped together based on co-occurrence patterns in the training text corpora. CNA also provides a scoring mechanism for quantifying the importance of each analysis element (sentence, paragraph or entire document) based on the relevance of the underlying content words [41]. This is useful for evaluating the impact of individual sentences in relation to the whole document. In addition to the cohesion-centered discourse representation, specific discourse connectors and conjuncts for Romanian language are also identified using cue phrases in order to evaluate the degree of discourse elaboration, based on the following categories: coordinating connectives; logical connectors; semi and quasi coordinators; conjunctions; disjunctions; simple and complex subordinators; addition, contrasts, sentence linking, order, reference, reason and purpose constructs.

Third, *word complexity* is focused on evaluating each word's difficulty from multiple perspectives of discourse analysis: (a) distance in characters between the word stem, the lemma and the inflected form, (in general, multiple prefixes and suffixes increase the difficulty a certain word), (b) distinguishability approximated as the inverse document frequency from the Romanian text corpora, and (c) the word poly-semy count from the Romanian WordNet [42] (words with multiple senses tend to be more difficult to comprehend).

3   Results and Discussions

Statistical analyses were performed to investigate the differences in the writing styles of journalistic texts produced by the two famous Romanian orators. As mentioned in the previous section, our analyses were focused on lexical and semantic properties of the journalistic texts. First, all variable indices reported by *ReaderBench* were checked for normality and those that demonstrated non-normality were removed. Multicollinearity was then assessed as pair-wise correlations ($r > .70$); if writing style properties demonstrated multicollinearity, the index that demonstrated the strongest effect in the model was retained for the final analysis (see Table 2 for final list of indices and their descriptive statistics). As it was expected, character entropy is a language feature and there are no significant differences between authors.

Table 2. General statistics.

Index	M (SD) M. Eminescu ($N = 65$)	M (SD) I.C. Brătianu ($N = 45$)	M (SD) Corpus ($N = 110$)
Average word length	3.88 (0.23)	3.68 (0.2)	3.8 (0.24)
Standard deviation in word letters	2.67 (0.18)	2.56 (0.13)	2.63 (0.17)
Average words per sentence	31.20 (8.54)	33.87 (10.11)	32.29 (9.26)
Standard deviation in unique words per sentence	5.53 (1.22)	5.50 (1.54)	5.52 (1.35)
Word entropy	5.41 (0.22)	5.29 (0.28)	5.36 (0.25)
Character entropy	2.71 (0.02)	2.71 (0.02)	2.71 (0.02)
Average difference between word and stem	1.32 (0.17)	1.33 (0.20)	1.32 (0.18)
Average word polysemy count	5.66 (0.75)	6.24 (0.89)	5.89 (0.86)
Average sentence score	0.55 (0.27)	0.75 (0.27)	0.63 (0.28)
Average sentence-paragraph cohesion (LSA)	0.70 (0.09)	0.65 (0.10)	0.68 (0.10)
Average sentence-paragraph cohesion (LDA)	0.82 (0.10)	0.73 (0.11)	0.79 (0.12)
Average intra-paragraph cohesion (LSA)	0.16 (0.07)	0.22 (0.07)	0.19 (0.08)
Average intra-paragraph cohesion (LDA)	0.41 (0.07)	0.46 (0.05)	0.43 (0.07)

Afterwards, a multivariate analysis of variance (MANOVA) [43, 44] was conducted to examine whether the lexical and semantic properties of the journalistic texts differed between the two famous Romanian orators. Box's M test (104.308) of equality of covariance matrices was not significant, $p(.017) > \alpha(.001)$, indicating that there are no significant differences between the covariance matrices. For all the variables presented in Table 3, Levene's test of equality of error variances is not significant

($p > .05$); therefore, the MANOVA assumption that the variances of each variable are equal across the groups is met. There was a significant difference among the two authors, Wilks' $\lambda = .0.512$, $F(11,98) = 8.498$, $p < .001$ and partial $\eta^2 = .488$. The textual complexity indices from Table 3 presented in descending order of effect size denote the variables that were significantly different between the two orators. Sentence-paragraph cohesion in both LSA and LDA semantic models capture the average resemblance between each constituent phrase and its corresponding paragraph (i.e., local cohesion with the main idea of the paragraph), whereas intra-paragraph cohesion measures the cohesion between each pair of phrases of the same paragraph (i.e., local cohesion in-between phrases). Corroborated with Fig. 1, we can observe that Eminescu uses in general more elaborated words (higher length), fewer, but more diverse words per sentence, as well as more self-contained and cohesive paragraphs.

Table 3. Tests of between-subjects effects for significantly different indices.

Index	df	Mean square	F	p	Partial Eta squared
Average word length	1	1.096	23.705	<.001	.180
Average sentence-paragraph cohesion (LDA)	1	0.239	20.500	<.001	.160
Average intra-paragraph cohesion (LSA)	1	0.094	19.455	<.001	.153
Average sentence score	1	1.064	14.958	<.001	.122
Standard deviation in word letters	1	0.363	14.174	<.001	.116
Average word polysemy count	1	8.935	13.551	<.001	.111
Average sentence-paragraph cohesion (LSA)	1	0.063	7.004	.009	.061

(a)   (b)
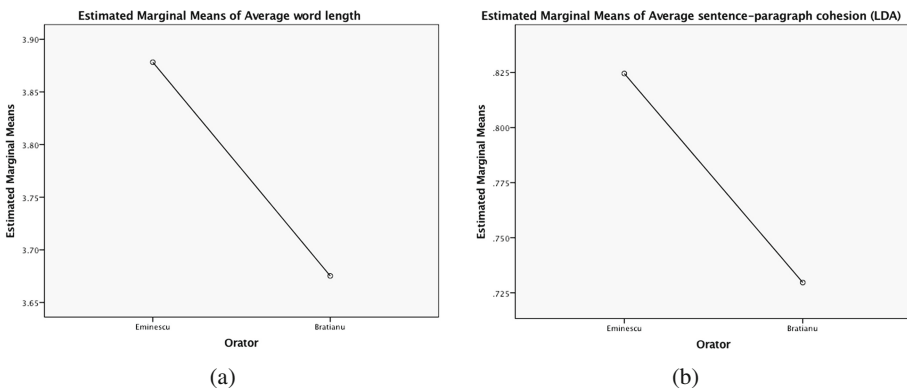
Fig. 1. Comparative views of writing styles reflected in textual complexity indices applied on the journalistic texts of both orators.

A stepwise Discriminant Function Analysis (DFA) was performed to predict the author of a given text based on the underlying writing style properties. The DFA retained five variables as significant predictors (*Average sentence-paragraph cohesion - LDA*, *Average words per sentence*, *Average intra-paragraph cohesion - LSA*, *Average word polysemy count*, *Average word length*) and removed the remaining variables as non-significant predictors.

The results prove that the DFA using these five indices significantly differentiated the texts pertaining to the two authors, Wilks' λ = .609, χ^2(df = 5) = 52.353 p < .001. The DFA correctly allocated 90 (50 + 40) of the 110 documents from the total set, resulting in an accuracy of 81.82 % (the chance level for this analysis is 50 %). For the leave-one-out cross-validation (LOOCV), the discriminant analysis allocated 89 (50 + 39) of the 110 texts for an accuracy of 80.90 % (see the confusion matrix reported in Table 4 for detailed results). The measure of agreement between the actual author and that assigned by the model produced a weighted Cohen's Kappa of 0.636, demonstrating substantial agreement.

Table 4. Confusion matrix for DFA classifying texts pertaining to different orators based on writing style properties.

		M. Eminescu	I.C. Bratianu	Total
Whole set	M. Eminescu	50	15	65
	I.C. Bratianu	5	40	45
		M. Eminescu	I.C. Bratianu	Total
Cross-validated	M. Eminescu	50	15	65
	I.C. Bratianu	6	39	45

4   Conclusions and Future Work

This research presents an in-depth study conducted to compare the work of two Romanian orators in terms of specificities of their writing style. The results reveal significant and interesting differences with regards to the degree of word elaboration (length and polysemy count), word diversity, as well as local cohesion reflected in the intra-paragraph and sentence-paragraph semantic similarity measures. Mihai Eminescu, probably due to the fact that he was also a great poet (considered the most important poet in Romania's literature), used more elaborated, lengthier words. Sentences contained fewer, but more diverse words, and paragraphs were more self-contained and cohesive that in the case of I.C. Brătianu. The journalistic texts of both orators are very complex (more than 30 words per sentence) and the selected features were successfully used to predict the author of a given text based on the underlying writing style properties, thus highlighting a clear demarcation between their work.

As extension of this study, we envision the inclusion of texts pertaining to other representative Romanian authors from different time periods and, potentially, other genres (for example, novels and essays) in order to identify additional individual traces of their writing style. This will also enable us to model trends in the time evolution of the Romanian language.

References

1. de Saussure, F.: Cours de Linguistique Générale. Payot, Paris (1999)
2. Bo, L., Gaussier, E., Morin, E., Hazem, A.: Degré de comparabilité, extraction lexicale bilingue et recherche d'information interlingue. In: Conf´erence sur le Traitement Automatique des Langues Naturelles, vol. 1, pp. 211–222. LIRMM Montpellier, Montpellier (2011)
3. Morin, E., Daille, B.: Comparabilité de corpus et fouille terminologique multilingue. Traitement Automatique des Langues **47**(1), 113–136 (2006)
4. Gîfu, D.: Contrastive diachronic study on romanian language. In: FOI 2015, pp. 296–310. Institute of Mathematics and Computer Science, Academy of Sciences of Moldova (2015)
5. Aijmer, K., Altenberg, B., Johansson, M.: Languages in contrast: papers from a symposium on text-based cross-linguistic studies, Lund 4–5 March 1994, vol. 88. Lund studies in English (1996)
6. Vianu, T.: Arta prozatorilor români. Ed. Contemporană, Bucharest (1941)
7. Coteanu, I.: Stilistica Funcțională a Limbii Române, vol. 81. Editura Academiei, Bucharest (1993)
8. Ibrăileanu, G.: Spiritul Critic în Cultura Românească. Tipografia Moldova, Iași (2001)
9. Brătianu, I.C.: Memoire sur l'Empire d'Autriche dans la question d'Orient, Paris, France (1855)
10. Brătianu, I.C.: Memoire sur la situation de la Moldo–Valachie depuis la Traite de Paris, Paris, France (1857)
11. Dascalu, M.: Analyzing Discourse and Text Complexity for Learning and Collaborating. SCI, vol. 534. Springer, Cham (2014)
12. Dascalu, M., Gifu, D.: Evaluating the complexity of online Romanian press. In: 11th International Conference "Linguistic Resources and Tools for Processing the Romanian Language", Iasi, Romania, pp. 149–162 (2015)
13. Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., Nardy, A.: Mining texts, learner productions and strategies with ReaderBench. In: Peña-Ayala, A. (ed.) Educational Data Mining. SCI, vol. 524, pp. 335–377. Springer, Cham (2014)
14. Dascalu, M., Stavarache, L.L., Dessus, P., Trausan-Matu, S., McNamara, D.S., Bianco, M.: ReaderBench: an integrated cohesion-centered framework. In: Conole, G., Klobucar, T., Rensing, C., Konert, J., Lavoué, E. (eds.) EC-TEL 2015. LNCS, vol. 9307, pp. 505–508. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24258-3_47
15. National Governors Association Center for Best Practices & Council of Chief State School Officers: Common Core State Standards. Authors, Washington D.C. (2010)

16. Eggins, S., Martin, J.R.: Genres and register of discourse. In: van Dijk, T.A. (ed.) Discourse as Structure and Process (Discourse Studies – A Multidisciplinary Introduction), vol. 1, pp. 231–232. Sage Publications, London (1997)

17. Biber, D.: A textual comparison of British and American Writing. Am. Speech **62**, 99–119 (1987)

18. Rosetti, A., Cazacu, B., Onu, L.: Istoria limbii române literare. Editura Minerva, București (1971)

19. Iordan, I.: Stilistica Limbii Române. Editura Științifică, București (1975)

20. Sala, M.: De la latină la română. Limba română, vol. 1. Editura Univers Enciclopedic & Academia Română, București (1998)

21. Guțu-Romalo, V.: Aspecte ale evoluției limbii române, Vol. Repere. Editura Humanitas Educațional, București (2005)

22. Slotnick, H.: Toward a theory of computer essay grading. J. Educ. Meas. **9**(4), 253–263 (1972)

23. Wresch, W.: The imminence of grading essays by computer—25 years later. Comput. Compos. **10**(2), 45–58 (1993)

24. Nelson, J., Perfetti, C., Liben, D., Liben, M.: Measures of text difficulty: Testing their predictive value for grade levels and student performance. Council of Chief State School Officers, Washington, DC (2012)

25. Shannon, C.E.: Prediction and entropy of printed English. Bell Syst. Tech. J. **30**, 50–64 (1951)

26. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J., **27**, 379–423 & 623–656 (1948)

27. Gervasi, V., Ambriola, V.: Quantitative assessment of textual complexity. In: Barbaresi, M.L. (ed.) Complexity in Language and Text, pp. 197–228. Plus, Pisa (2002)

28. McNamara, D.S., Graesser, A.C., Louwerse, M.M.: Sources of text difficulty: Across the ages and genres. In: Sabatini, J.P., Albro, E., O'Reilly, T. (eds.) Measuring up: Advances in how we assess reading ability, pp. 89–116. R&L Education, Lanham (2012)

29. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)

30. Crossley, S.A., Dascalu, M., Trausan-Matu, S., Allen, L., McNamara, D.S.: Document Cohesion Flow: Striving towards Coherence. In: 38th Annual Meeting of the Cognitive Science Society. Cognitive Science Society, Philadelphia (in press)

31. Dascalu, M., Trausan-Matu, S., McNamara, D.S., Dessus, P.: ReaderBench – automated evaluation of collaboration based on cohesion and dialogism. Int. J. Comput.-Support. Collaborative Learn. **10**(4), 395–423 (2015)

32. Dascalu, M., Dessus, P., Trausan-Matu, Ş., Bianco, M., Nardy, A.: *ReaderBench*, an environment for analyzing text complexity and reading strategies. In: Lane, H.C., Yacef, K., Mostow, J., Pavlik, P. (eds.) AIED 2013. LNCS, vol. 7926, pp. 379–388. Springer, Heidelberg (2013)

33. Trausan-Matu, S., Dascalu, M., Dessus, P.: Textual complexity and discourse structure in computer-supported collaborative learning. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) ITS 2012. LNCS, vol. 7315, pp. 352–357. Springer, Heidelberg (2012)

34. Foltz, P.W., Kintsch, W., Landauer, T.K.: An analysis of textual coherence using latent semantic indexing. In: 3rd Annual Conference of the Society for Text and Discourse, Boulder, CO (1993)

35. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: the Latent semantic analysis theory of acquisition, induction and representation of knowledge. Psychol. Rev. **104**(2), 211–240 (1997)

36. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
37. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003)
38. Blei, D.M., Lafferty, J.: Topic models. In: Srivastava, A., Sahami, M. (eds.) Text Mining: Classification, Clustering, and Applications, pp. 71–93. Chapman & Hall/CRC, London (2009)
39. Golub, G.H., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. J. Soc. Ind. Appl. Math.: Ser. B, Numer. Anal. **2**(2), 205–224 (1965)
40. Kotz, S., Balakrishnan, N., Johnson, N.L.: Dirichlet and Inverted Dirichlet Distributions. Continuous Multivariate Distributions, vol. 1, Models and Applications, pp. 485–527. Wiley, New York (2000)
41. Dascalu, M., Trausan-Matu, S., Dessus, P., McNamara, D.S.: Discourse cohesion: a signature of collaboration. In: 5th International Learning Analytics & Knowledge Conference (LAK 2015), pp. 350–354. ACM, Poughkeepsie (2015)
42. Tufiş, D., Barbu Mititelu, V., Bozianu, L., Mihăilă, C.: Romanian wordnet: new developments and applications. In: 3rd Global Wordnet Conference 2006 (GWC 2006), Jeju Island, Korea, pp. 337–344 (2006)
43. Stevens, J.P.: Applied multivariate statistics for the social sciences. Lawrence Erblaum, Mahwah (2002)
44. Garson, G.D.: Multivariate GLM, MANOVA, and MANCOVA. Statistical Associates Publishing, Asheboro (2015)