Discourses of Economic Crisis in Romanian Media

An Automated Analysis Using the *ReaderBench* Framework

SIMINA-MARIA TERIAN, TEODOR MIHAI COTEŢ,
MARIA-DORINELA SÎRBU, MIHAI DASCĂLU,
ŞTEFAN TRĂUŞAN-MATU

1. Introduction

1.1. State of the Art

"ECONOMICS IS everywhere. Law, management, politics, personal relations—we draw from it all."[1] These half-warning, half-advertising sentences are borrowed from a practical handbook on economics authored by two of the most popular economists of our time. Perhaps they are true and perhaps economics truly is everywhere. Certainly, however, not everybody agrees that this is the case; or, at any rate, not at all times. Even when they encounter financial difficulties, most people relate to them in terms of "tight corners" or "rashness" rather than "economics."

Nevertheless, there are times when economics is felt as intervening at its most brutal in people's private lives. In such cases, "tight corners" are seen as a direct consequence of economics and specialists tend to describe them by using words such as "crisis," "depression" and "recession." It is in such situations that not only common people, but also the representatives of other fields of study strive to understand optimally the mysterious, yet omnipotent mechanism that economics is.

This phenomenon is also witnessed in linguistics—specifically, in the linguistic approaches centered on economic and financial discourses. In fact, since its establishment as a science, linguistics has been in close contact with economics, as "[Ferdinand de] Saussure's

This work was supported by the following research grants: 1) INTELLIT, project number PN-III-P1-1.2-PCCDI-2017-0821, funded by the Romanian National Authority for Scientific Research and Innovation, CNCS–UEFISCDI, within PNCDI III; 2) the project LBUS-IRG-2018-04, financed from the Lucian Blaga University of Sibiu research grants; 3) the README project "Interactive and Innovative Application for Evaluating the Readability of Texts in Romanian Language and for Improving Users' Writing Styles," contract no. 114/15.09.2017, MySMIS 2014 code 119286.

notion of linguistic value is imbued with the formalist abstraction of the capitalist economy, indeed with some of its most fetishized appearances."[2] The subsequent century merely strengthened this connection, although the historical proximity between linguistics and economics seems to have undergone significant changes over recent decades. Following the world economic crisis of 2008–2012, we witness a departure from the works that used economic models for a better understanding of language (e.g., of authors such as Jacob Marshak,[3] François Grin,[4] Barry R. Chiswick and Paul W. Miller[5]) to works deploying various linguistic models in an attempt to better comprehend the functioning of the economy (e.g., of authors such as Honesto Herrera-Soler and Michael White,[6] Zubin Jelveh, Bruce Kogut, and Suresh Naidu[7]).

The situation is somewhat similar in Romanian linguistic research. Prior to the Great Recession, the number of Romanian works on the language of economics was very small and they were written by economists rather than by linguists.[8] Yet, following the beginning of the global crisis, the number of studies has increased and they can be placed in several categories. First, there are works such as Doina Butiurcă's, which deal with the language of economics in general.[9] Second, there are the studies that focus on the specific concept of (economic) crisis from a lexical-semantic perspective, primary among which are Elena Museanu's works.[10] Third, there are the comparative approaches that study the discourse of the crisis in Romanian by drawing parallels to its counterparts in other languages such as Italian and French.[11] Last but not least, in the category of most interest to us, i.e. the computational analyses of the crisis discourse in Romanian, we have been able to identify a single article, namely the study conducted by Daniela Gîfu and Dan Cristea.[12] The two authors do not aim, however, to investigate the manifestations of crisis discourse; they use Natural Language Processing techniques as a starting point for the development of a predictive method that can foresee the risks of a crisis outbreak.

1.2. The Issue

Unlike prior approaches, we seek to draw descriptive conclusions rather than put forward predictive strategies. Specifically, the present study aims to analyze the discourse of the latest global economic crisis against the *ReaderBench* Framework.

However, from the very beginning, some complications arise: when did, in fact, the "latest global economic crisis" occur? According to various data and analyses, a specialist in world economics would most likely argue that it spanned the period between 2008 and 2012. Yet, in today's Romania, not a week goes by without a TV channel, a magazine or a debate platform discussing the "current" crisis, which they refer to in terms of a current, inevitable, or merely probable phenomenon.

This article does not, however, intend to delineate the exact span of the latest global economic crisis, nor does it attempt to answer the question as to whether Romania is or is not experiencing an economic crisis. To our end, this information is irrelevant, because, for a society to have *a discourse of economic crisis,* suffering an economic crisis is not a must. A discourse of crisis may appear irrespective of whether the crisis is a matter of the past, present, or future, a current, inevitable, or probable phenomenon.

Moreover, it should be noted that there is no unique and homogeneous discourse of crisis; instead, there are multiple *discourses*, contradictory at times. The latter vary according to their subject matter—how "real" the crisis is—, attitudes toward it—why the crisis did/will (not) occur—, manner of depiction, style and to the profile of the one to which they are addressed. For these reasons, it would be more pertinent to reformulate our main objective and to say that this article seeks to analyze the Romanian discourses between 2008 and 2018 on the economic crisis against the *ReaderBench* Framework.

We have classified these types of discourses according to formal criteria (medium of communication, language used, and addressee type), thus delineating four categories:

- discourses in *specialized academic publications* (in general, ample articles built on specific argumentative structures, of a technical nature, written by economics specialists, and addressed to economics specialists);
- discourses in *specialized non-academic publications* (articles of various sizes, targeting a specialized audience, but also accessible to a wider public, which use a less technical language, and are authored by economists, but also by journalists with considerable experience in the subfield of economic journalism);
- discourses in *non-specialized non-academic publications* (smaller articles addressed to a non-specialized audience; in general, the authors of these articles are journalists with no special competence in the field of economics, who want to attract readers not only through their use of a more accessible language, but also by speculating sensationalistic, "breaking news" information);
- discourses on *informal debate platforms* (the texts in this category make up the most heterogeneous class, precisely because this environment is the most constraint-free; contributors to such platforms may be "opinion leaders" without special competences in economics, but also reputed specialists in this field, which leads to widely varying inputs).

The aim of our article is to identify and compare linguistic features specific to these four discourses.

2. Method

Around 200 complexity indices tailored for Romanian language were generated using the *ReaderBench* framework[13] and were used to explore differences between four writing styles, all addressing the economic crisis. The textual complexity indices covering lexical, semantic, and cohesive features were used in statistical analyses to highlight differences in writing style between the selected documents.

2.1. Corpus

Our corpus includes 200 texts, equally covering—in terms of amount (50), not necessarily length-wise—all the above-mentioned categories. More precisely, the samples were extracted from the following publications and websites:

- specialized academic publications: *Amfiteatru economic* (The Amphitheatre of Economics); *Buletinul AGIR* (The Bulletin of the General Association of Engineers in Romania–AGIR); *Revista română de economie* (The Romanian Journal of Economics); *Economie teoretică şi aplicată* (Theoretical and Applied Economics); *The Romanian Economic Journal*; *Urbanism. Arhitectură. Construcţii* (Urbanism.

Architecture. Construction Industry); *Management intercultural* (Intercultural Management); *Revista română de statistică* (The Romanian Journal of Statistics); *Economia. Seria Management* (Economics. Management Series);

- specialized non-academic publications: *Ziarul financiar* (The Financial Paper); *Capital*; *Bursa* (The Stock); *Săptămâna financiară* (Financial Weekly); *Money Express*; *Economistul* (The Economist); *Piaţa financiară* (The Financial Market);
- non-specialized non-academic publications: *Adevărul* (The Truth); *Evenimentul zilei* (The Event of the Day); *România liberă* (Free Romania); *Libertatea* (The Freedom); *Jurnalul naţional* (The National Journal);
- informal debate platforms: http://www.criticatac.ro; http://www.contributors.ro; http://inliniedreapta.net; http://www.cogitus.ro; https://voxpublica.ro; https://republica.ro; http://www.romaniacurata.ro.

All selected texts were published between 1 January 2008 and 31 December 2018. We aimed to cover all the eleven calendar years included in this timeframe, but it was not our intention to obtain an equal distribution of the texts by year, as such a homogenization would be artificial. Instead, as far as the last two categories are concerned, we sought to include the sources from which we could single out texts that cover a wide range of ideological perspectives (at least from the viewpoint of the Left *vs.* Right opposition).

Section headings from the texts were cleaned by removing all lines which contained fewer than 7 words. *ReaderBench* requires that every analyzed document has at least a few paragraphs in order to explore all textual complexity indices—three is generally a minimum in order to be capable to compute global cohesion indices. However, only 4 texts had less than 4 paragraphs, and we decided not to disregard them. General statistics on our corpus are presented in Table 1.

TABLE 1. GENERAL CORPUS STATISTICS

Period	N	Paragraphs M (SD)	Sentences M (SD)	Words M (SD)
Specialized academic publications	50	51.66 (19.01)	119.40 (52.97)	2,238.86 (1,220.11)
Specialized non-academic publications	50	14.30 (9.85)	39.04 (39.52)	889.24 (797.29)
Non-specialized non-academic publications	50	13.47 (9.89)	37.47 (24.27)	799.47 (607.27)
Informal debates	50	22.88 (23.54)	82.88 (89.71)	2,098.98 (2,348.51)

2.2. Textual Complexity Indices for Romanian Language

The textual complexity indices for Romanian Language were generated using the *ReaderBench* framework.[14] *ReaderBench* is an advanced Natural Language Processing framework providing various services designed to process texts in several languages (English, Romanian, Dutch, French, as well as partial support for Spanish and Italian) including: automated essay scoring, generating personalized feedback for writing, topics extraction, as well as comprehensive analyses of online communities.

ReaderBench is grounded in Cohesion Network Analysis[15] and integrates WordNet semantic distances,[16] as well as three semantic models in order to quantify text cohesion, namely: Latent Semantic Analysis (LSA),[17] Latent Dirichlet Allocation (LDA)[18] and word2vec.[19] *ReaderBench* generates multiple textual complexity indices, which are split into five categories: *surface level* (e.g., word/sentence/paragraph counts, punctuation marks), *syntactic indices* (e.g., statistics on syntactic dependencies and part-of-speech tags), *semantics* (focused on local and global cohesion), *word complexity* spanning across multiple analysis layers, and *discourse-centered indices* centered on specific connectors. Taking into account the dimension of our corpus for Romanian language, which exceeds 800 million words extracted from online public sources, only the word2vec model was used in the current experiments, besides WordNet semantic distances.

2.2.1. Surface Metrics

Surface analysis provides basic measurements considering lexical elements at various granularities—paragraphs, sentences, and words—, for example: the average length of characters in each paragraph; the average number of commas at paragraph/sentence levels; the average length of sentences expressed in character counts; the average number of sentences, at paragraph level; the average number of words at paragraph/sentence levels; the standard deviation of number of words in sentence, at paragraph level, etc.

Entropy[20] provides valuable insights in terms of vocabulary diversity. A more complex text contains more information and requires more memory and more time for a reader to process and understand it. Therefore, the state of disorder modeled by entropy is reflected in the variety of word root forms considered by our model.

2.2.2. Syntax

The analysis of subordinate elements and part-of-speech (POS) tagging play an important role in the syntactic analysis of texts in terms of textual complexity, by providing two possible evaluation vectors: the normalized frequency of each POS tag, and indices related to structure and lexical dependencies derived from the parsing tree. Although the most common parts of speech which create the contextualization of a text are nouns and verbs, our emphasis is also on prepositions, adjectives and adverbs, which dictate a more complex and deeper structure of the text. In addition, pronouns indicate the presence of co-references which require anaphora resolution, together with a more complex structure and additional interconnections within the discourse. Moreover, several indices can be derived by analyzing the structure of the syntactic parsing tree and an increased number of specific syntactic dependencies can indicate a more complex phrase structure, thus a higher textual complexity.[21] Therefore, the indices generated by the *ReaderBench* framework at syntax level are: the average number of all or unique nouns, pronouns, verbs, adverbs, adjectives and prepositions, at paragraph and sentence level; the average number of pronouns at paragraph and sentence level (first, second, third person, indefinite, interrogative).

The lexical dependencies for Romanian Language are computed using models based on neural networks trained on the RoRefTrees corpus,[22] which contains manually annotated texts from various areas, including literature, medicine, and academic writings. The RoRefTrees contains 218,511 tokens. Based on the annotations from the corpus, 50 types of dependencies are learned, out of which 11 are specific to Romanian. The dependency types which are not language dependent are taken from a trans-linguistic standardization initiative of the syntactic annotation methodology, called Universal Dependency (UD).[23]

2.2.3. Semantics, Cohesion and Discourse Structure

In order to understand a text, a reader must first create a well-connected representation of the contained information linked with prior knowledge, called a situation model.[24] A well-structured document should have a clear and logical linkage of ideas, besides a semantic correlation of text segments.[25] Cohesion can be established as an average measure of semantic similarity and lexical proximity between textual segments, which can be words, phrases, paragraphs, or the whole conversation. Semantic similarity is computed using lexical proximity, identified as semantic distances[26] within WordNet ontology,[27] and various semantic models. In addition, specific language processing techniques are applied to improve the system's accuracy: a) reduce word forms to their dictionary inflected root forms (i.e., lemmas) and b) eliminate stop-words (words that have no semantic content and thus do not define the context, for example: prepositions, conjunctions, etc.).

Cohesive links are defined as connections between textual elements that have high cohesion values (i.e., a value exceeding the mean value of all semantic similarities between constituent textual elements) and a cohesion graph is built to model all these links.[28] *ReaderBench* framework uses Cohesion Network Analysis which combines semantic cohesion with Social Network Analysis (SNA)[29] measures applied to the multi-layered cohesion graph. Concerning local and global cohesion, the following indices are computed for Romanian language within the *ReaderBench* framework: the average and standard deviation scores of relevance, at paragraph and sentence level, obtained from CNA; the average cohesion between adjacent paragraphs; the average cohesion between each paragraph and the entire document; the average of local cohesion between sentences and paragraph; etc.

The method proposed by Galley and McKeown (2003)[30] is used to disambiguate the meanings of words and construct lexical chains comprising of related concepts. The algorithm consists in three steps. First, the entire text is processed and a graph representing all the possible interpretations of the text using WordNet ontology is created. Each node in this graph is a word, along with all its meanings taken from WordNet. Although a node is a word, links are defined according to meaning, not to words. Depending on this relationship, each relative edge is associated with a weight, and the resulting graph is called the disambiguation graph. Second, the meaning of each word is selected by maximizing the sum of the weights. Third, all linkages between the unselected meanings or senses are eliminated, thus generating lexical chains of related words. Thus, multiple metrics are generated at the level of lexical chains, such us: the average coverage of lexical chains; the maximum coverage of lexical chains; the average number of lexical chains identified at paragraph level; the percentage of words included in lexical chains.

2.2.4. Discourse Structure

Regular expressions are used to recognize the use of different discourse connectors, thus providing more insights into the degree of discourse elaboration by computing the average number of connectors at both paragraph and sentence levels of the following categories: addition (e.g., "şi," "iar," "sau"), concessions (e.g., "deşi," "cu toate că"), conditions (e.g., "dacă," "în cazul că"), temporal connectors (e.g., "primul," "întâi"), logical connectors (e.g., "şi," "iar," "sau"), quasi-coordinators (e.g., "precum şi," "ca şi"), semi-coordinators (e.g., "nici," "aşa"), conjunctions (e.g., "sau," "dar"), coordinating conjunctions (e.g., "încă," "totuşi"), contrasts (e.g., "dar," "ci," "cu toate că").

2.2.5. Word Complexity

The complexity of words is a combination of various analysis levels, and it can be reflected using the following indices: the average length of words expressed in character count; the standard deviation in the number of characters, at word level; the average distance between the lemmas and the roots of the words; the average distance between words and their roots; the maximum depth/the average depth of words in the hyponymy tree. The indices associated to words are computed in a simple manner by averaging the values of all content words from the text (i.e., dictionary word forms, lemmatized, not included in the stop-words list, and having as part-of-speech: noun, verb, adverb, or adjective). Our assumption is that words with multiple senses are more difficult to correctly disambiguate and include within the situational model. Therefore, simpler texts will likely contain less ambiguous words. In addition, the distance between the position of a word in the hyponymy tree and its relation to the root of this tree can be perceived as a measure of word specificity. In other words, more specific words tend to have a longer path to the root of the ontology.

3. Results

The *ReaderBench* framework facilitates a thorough statistical analysis of various properties including lexicon, syntax, semantics, text cohesion, and discourse structure. The analysis started from approximately 200 textual complexity indices for the Romanian language which were further filtered along different criteria, as described below.

The first criterion was linguistic coverage, namely that an index is relevant and can be computed for at least 20% of the documents; indices with low linguistic coverage were disregarded. Most of those indices were related to the number of specific word lists. Second, the normality of the distribution of each index was checked; more precisely, indices with the absolute Skewness or Kurtosis value greater than 2 were eliminated. Third, multicollinearity tests based on pair-wise comparisons were performed ($r > .90$) and only the most predictive indices were retained. The fourth criterion consisted of Levene's test of equality of error variances and disregarding indices whose resulting p-values are significant ($p < .05$).

TABLE 2. TESTS OF BETWEEN-SUBJECT EFFECTS FOR SIGNIFICANTLY DIFFERENT INDICES (M: MEAN; SD: STANDARD DEVIATION)

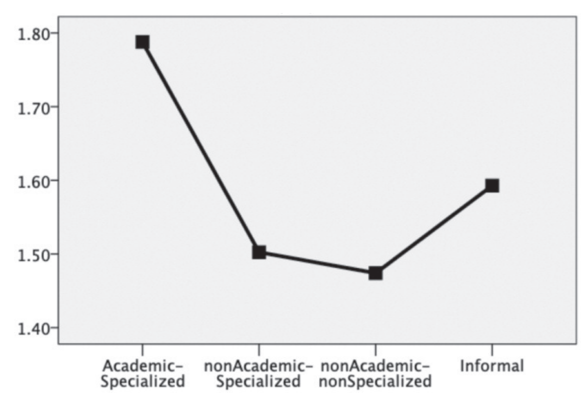
Textual complexity index	Specialized Academic M (SD)	Specialized non-Academic M (SD)	non-Specialized non-Academic M (SD)	Informal	F	*P*	Partial η^2
Word length difference between stem and original form (M)	1.79 (0.19)	1.50 (0.17)	1.47 (0.2)	1.59 (0.17)	30.183	<.001	.316
Word length in characters (M)	7.42 (0.33)	6.95 (0.29)	6.89 (0.41)	7.07 (0.28)	25.277	<.001	.279
Local cohesion – Wu-Palmer distance between sentences and paragraphs (M)	0.90 (0.04)	0.88 (0.06)	0.86 (0.06)	0.83 (0.05)	17.152	<.001	.208
Global cohesion – word2vec similarity between paragraphs and document (M)	0.84 (0.04)	0.89 (0.04)	0.88 (0.05)	0.9 (0.04)	16.701	<.001	.204
Word entropy	5.38 (0.25)	5.15 (0.40)	5.17 (0.34)	5.56 (0.35)	16.182	<.001	.199
Word letters (SD)	2.55 (0.12)	2.44 (0.15)	2.39 (0.13)	2.48 (0.1)	13.892	<.001	.175
Unique content words per sentence (M)	8.41 (1.87)	9.75 (2.45)	8.58 (1.9)	10.51 (2.47)	10.253	<.001	.136
Prepositions per sentence (M)	3.32 (0.84)	4.34 (1.14)	3.89 (1.06)	4.33 (1.21)	10.179	<.001	.135
Coordinating conjunctions per sentence (M)	2.72 (0.57)	2.33 (0.80)	2.02 (0.67)	2.38 (0.68)	8.739	<.001	.118
Paragraph CNA score (SD)	14.45 (7.87)	9.41 (5.95)	9.96 (9.05)	16.47 (9.76)	8.63	<.001	.117
Global cohesion – inter-paragraph cohesion using word2vec (M)	0.82 (0.05)	0.83 (0.06)	0.80 (0.07)	0.86 (0.05)	7.4	<.001	.102
Nouns per sentence (M)	6.52 (1.83)	7.88 (2.05)	6.74 (1.7)	7.79 (1.94)	6.916	<.001	.096
Percentage of words included in lexical chains	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.02 (0.01)	3.199	.025	.047

After the four filters were sequentially applied, 17 indices passed, and were subject to a Multivariate Analysis of Variance (MANOVA)[31] in order to identify significant differences in writing styles across the 4 documents types. The MANOVA exhibited significant difference, Wilks' $\lambda = .116$, $F(17, 182) = 11.172$, $p < .001$, and partial $\eta^2 = .512$. The 13 significant indices are presented in descending order of effect size in Table 2, whereas the most representative 8 indices in terms of fluctuations are displayed in Figure 1.

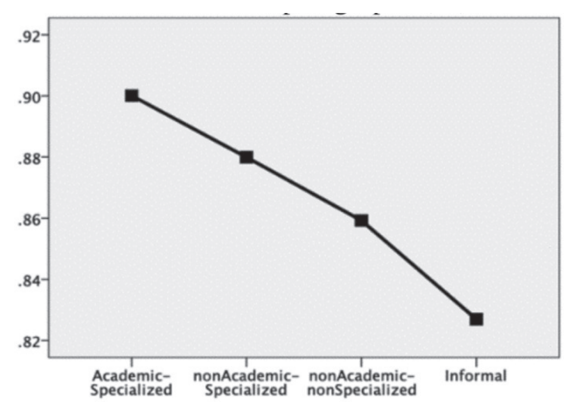
The first index (1.a) denotes the average length between words and their corresponding stems, a higher difference denoting a more sophisticated concept in terms of additional suffixes, prefixes, or complex inflections. As expected, the trend is descending as academic publications tend to use a more specific language, which often implies more sophisticated words. Indices related to cohesion using different semantic models show that specialized publications exhibit high local cohesion (1.b)—i.e., well defined, self-contained idea within each paragraph –, while the trend for global cohesion (1.c) is the opposite as specialized publications contain a more varied vocabulary and have a higher diversity of ideas. Word entropy (1.d) measures the frequency of repeated words, in terms of the logarithm of the probability distributions of word occurrences. Informal debates are the longest and contain the most varied vocabulary compared to the others, followed by academic texts. Figure 1.e exhibits an interesting trait: although the vocabulary is quite diverse for academic and informal texts, academic texts contain considerably fewer unique content words per sentence. Correlated with the previous finding, Figure 1.h shows a great difference between specialized academic texts and all other types of texts, indicating that that the concepts from academic documents are better interlinked with one another, and more lexical chains can be created to group related words.

FIGURE 1.A-H. COMPARISON BETWEEN WRITING STYLES IN TERMS OF TEXTUAL COMPLEXITY INDICES

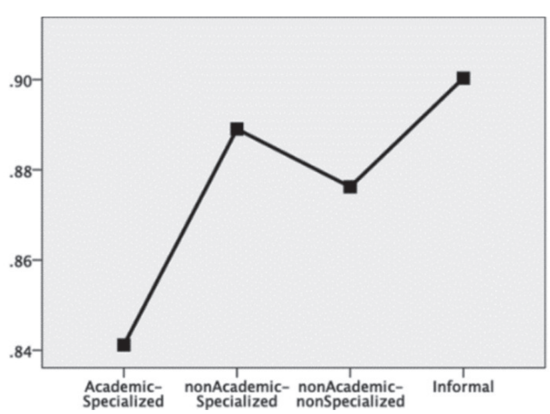
a. Word length difference between stem and original form (M)



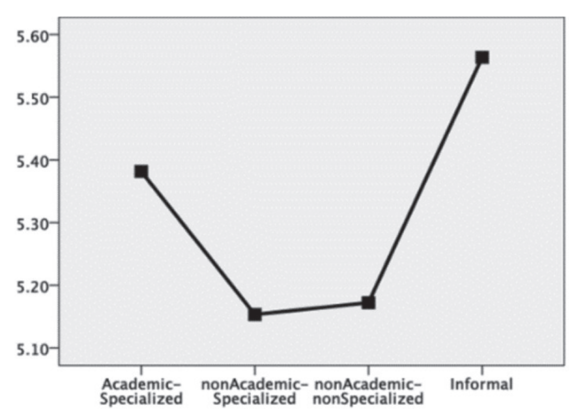
b. *Local cohesion*—Wu-Palmer distance between sentences and paragraphs (M)



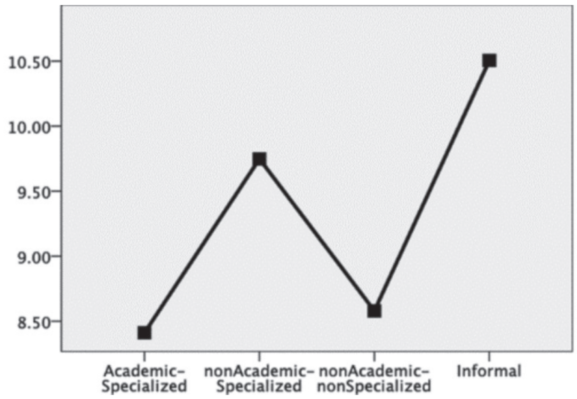
c. *Global cohesion*—word2vec similarity between paragraphs and document (M)



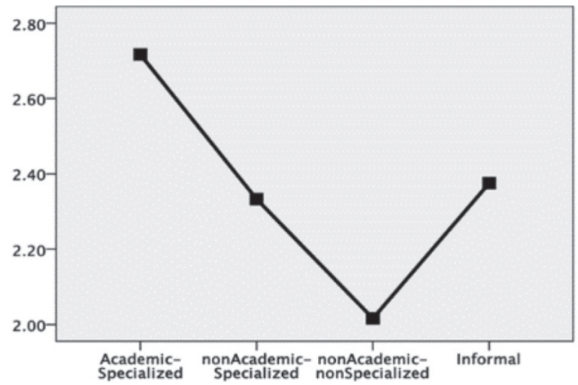
d. Word entropy



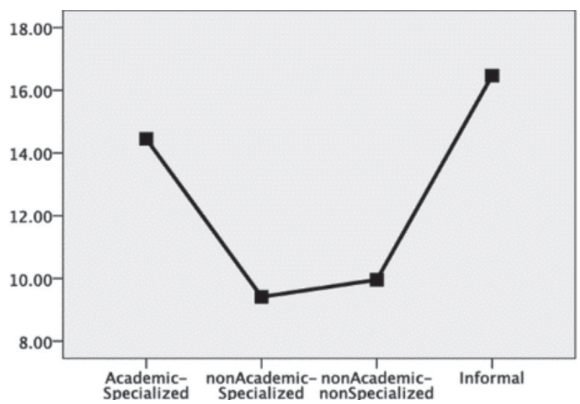
e. Unique content words per sentence
(M)



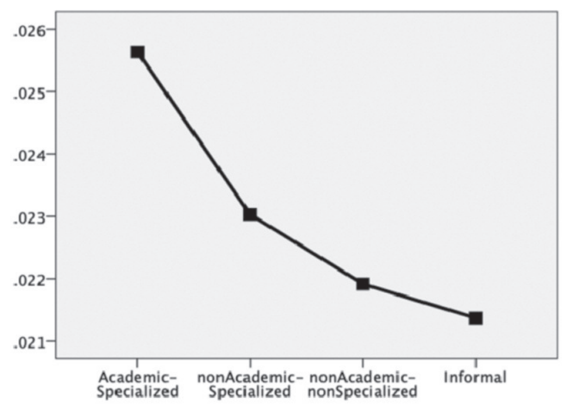
f. Coordinating conjunctions per sentence
(M)



g. Paragraph CNA score
(SD)



h. Percentage of words included in lexical
chains



A stepwise Discriminant Function Analysis (DFA) was performed in order to predict the document type based on its underlying writing style properties. The DFA retained eight variables as significant predictors:

- Word length difference between stem and original form (M);
- Local cohesion—Wu-Palmer distance between sentences and paragraphs (M);
- Global cohesion—word2vec similarity between paragraphs and document (M);
- Unique words per sentence (M);
- Prepositions per sentence (M);
- Coordinating conjunctions per sentence (M);
- Global cohesion—inter-paragraph cohesion using word2vec (M);
- Nouns per sentence (M).

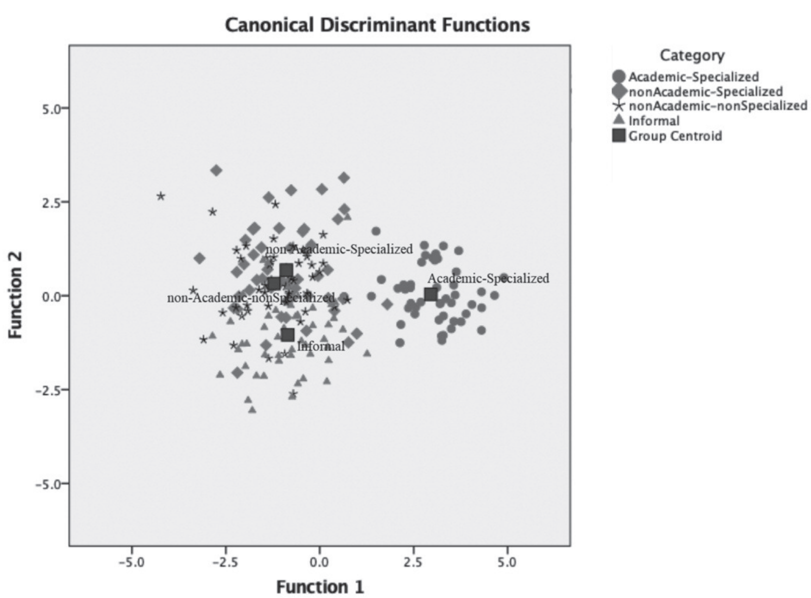
DFA reported an accuracy of 71.05% for correctly categorizing 143 documents (48+27+29+39) from the total of 200. Similar results were obtained for the LOOCV (leave-one-out cross-validation) which reported an accuracy of 64.50% (see the confusion matrix from Table 3). All F1-scores exceed the double of the random baseline, however the differences in F1 scores for the 4 categories are considerable. A clear distinction can be noted between academic specialized documents and all other types

(F1-score of 96%), followed by informal texts (F1 score of 67%), whereas the classifier has the biggest problems in identifying differences between specialized and non-specialized non-academic works, which are similar in writing style. Figure 2 introduces a visual representation of the documents depicted as circles using two canonical discriminant functions obtained from the DFA. In line with the validation results, we can observe a clear separation of academic-specialized documents (right-hand side of the plot), whereas the centroids for specialized and non-specialized non-academic texts are extremely close one to another.

TABLE 3. CONFUSION MATRIX FOR THE DFA

	Document type	Specialized Academic	Specialized non-Academic	non-Specialized non-Academic	Informal	F1 score (%)
Original	Specialized academic	48	2	0	0	96.96
	Specialized non-academic	1	27	15	7	55.10
	Non-specialized non-academic	0	15	29	6	57.42
	Informal	0	4	7	39	76.47
Cross-validated	Specialized academic	48	2	0	0	96.00
	Specialized non-academic	1	22	17	10	46.31
	Non-specialized non-academic	1	16	25	8	48.54
	Informal	0	5	11	34	66.66

FIGURE 2. SEPARATION OF DOCUMENTS TYPES IN TERMS OF WRITING STYLE USING TWO CANONICAL DISCRIMINANT FUNCTIONS OBTAINED FROM THE DFA



4. Discussions and Conclusions

AIMING TO perform an analysis of economic crisis discourses in contemporary Romanian media, this study yielded a series of outcomes, once the discourse categories were outlined. The delineation of the four categories did not seek to reach *specific* results, but a part of these findings seemed predictable. Specifically, we started from the intuitive finding that the four identified types of discourse—specialized academic (SA); specialized non-academic (SN); non-specialized non-academic (NN); and informal debates (ID)—would arrange in a relatively homogeneous progression from a "formal" (academic) pole to an "informal" one (of "free" discussions on platforms). Furthermore, we considered that the distances between the values of two adjacent categories—for example, between SA and SN, or between SN and NN—would be fairly similar or even almost equal, regardless of the values. The only major unknown was, in this context, the informal debates (ID), where the name "informal" may be misleading. In this case, *informal* does not necessarily mean *anti-academic* (opposed to SA), but rather *uncodified*, *insufficiently regulated*—and, for this reason, ID do not oppose SA only, but also SA, as well as SN and NN. Therefore, ID were the major unknowns, which we expected to yield the most surprising findings.

The preliminary exploration of the corpus (described in Table 1) appeared to confirm this theory, since the ID documents included, at the level of three surface indices (their number of paragraphs, sentences, and words), the highest coefficient of variation (CV). If in the case of SA, a heavily codified discourse, the CV (as ratio between SD and M) ranges between 36.8% and 54.5%, this coefficient witnesses a substantial increase in the SN (68.8%–101.2%) and NN discourses (70.4%–76%), while in the case of ID, it exceeds each of the indices by 100% (102.8%–111.8%). Therefore, of the four types of discourse analyzed, the informal debates (ID) feature *the highest dispersion of data.*

This heterogeneous character of ID is also evident from Table 2 and Figures 1.a–h. Indeed, the two representations of the indices once again confirm our initial theory: in general, in the SA, SN, and NN discourses, there is a tendency for arrangement from Complex to Simple, in reverse order, which may appear as "regression", but which, in the context of our research, would more appropriately be called a *downward progression.* This trend characterizes the majority of the 13 textual complexity indices depicted in Table 2; below, we offer comments on several of them:

• Word length difference between stem and original form (M): 1.79 (SA)–1.50 (SN)–1.47 (NN); Word length in characters (M): 7.42 (SA)–6.95 (SN)–6.89 (NN). While the SA discourse tends to use longer words, derivatives sometimes, there is a decrease in this tendency in the NN discourse, which prefers plain and short words, for a briefer and more concentrated transmission of the information.

• Local cohesion—Wu-Palmer distance between sentences and paragraphs (M): 0.90 (SA)–0.88 (SN)–0.86 (NN). Given its technical and more heavily conceptualized character, the SA discourse opts for redundancy and a lower level of progression in the presentation of data; journalistic discourses (SN and NN), on the other hand, exhibit a lower local cohesion in the presentation of data, mostly because they cover a more varied spectrum of information in a relatively narrower area of presentation. It should also be noted that this phenomenon of downward progression characterizes the most important three complexity indices relevant to the comparison of the four types of discourses.

• Coordinating conjunctions per sentence (M): 2.72 (SA)–2.33 (SN)–2.02 (NN). Accountable for this drop is the fact that the SA operate, by and large, with complex sentences, deployed to describe the nuances and relations between elements. Conversely, SN and NN prefer the juxtaposition of sentences at the expense of conjunctions, to increase accessibility and reading speed.

• Word entropy: 5.38 (SA)–5.15 (SN)–5.17 (NN); Paragraph CNA score (SD): 14.45 (SA)–9.41 (SN)–9.96 (NN). Although not operating but partially within the "rule" of downward progression—the SN values are, in these cases, lower than that of the NN—, the two indices nonetheless confirm the principle on which the former is based, i.e. that the journalistic discourses (SN and NN) require a reduction of coherence and complexity in relation to the academic discourse (SA).

The question then arises as to whether informal debates (ID) follow this trend. An obvious finding is that ID tend to occupy an *eccentric position* in relation to the other three types of discourse. Specifically, if some indices of ID follow the trends of the non-specialized non-academic discourse (NN), in the case of other variables, it appears to return to—and even exceed—the values of the specialized academic discourse (SA). The former, a sparser phenomenon, is found in the case of two indices alone: *Local cohesion*—Wu-Palmer distance between sentences and paragraphs (M): 0.90 (SA)–0.88 (SN)–0.86 (NN)–0.83 (ID); and Percentage of words included in lexical chains (see Figure 1.h). The other phenomenon is wider and is witnessed across 6 indices: Word length difference between stem and original form (M)—see Figure 1.a; Word length in characters (M): 7.42 (SA)–6.95 (SN)–6.89 (NN)–7.07 (ID); Word entropy—see Figure 1.d; Word Letters (SD): 2.55 (SA)–2.44 (SN)–2.39 (NN)–2.48 (ID); Coordinating conjunctions per sentence (M)—see Figure 1.f; and Paragraph CNA score (SD)—see Figure 1.g. Therefore: on the one hand, the discourse of informal debates appears to verge toward the concentration and heterogeneity of the information proper to non-specialized non-academic discourse; on the other hand, it appears to simulate the complexity and breadth of tones witnessed in the specialized academic discourse.

But how can these theoretically opposing parameters be reconciled? And, more importantly, which would be the consequences of such reconciliation? The answer to the first question is, perhaps, simpler than expected—by a *non-academic specialized discourse*. After all, this was the reason why economic magazines, from *The Financial Times* to *The Nikkei*, were established around the world: to provide readers with dense and quality information in both a non-technical and non-trivial discourse. And the outcome of the ID discourse following this trend is its foreseeable closeness to the values of the NS discourse. This phenomenon is readily observable in 5 complexity indices: *Global cohesion*—word2vec similarity between paragraphs and document (M)—see Figure 1.c; unique content words per sentence (M) – see Figure 1.e; prepositions per sentence (M): 3.32 (SA)–4.34 (SN)–3.89 (NN)–4.33 (ID); *Global cohesion*—inter-paragraph cohesion using word2vec (M): 0.82 (SA)–0.83 (SN)–0.80 (NN)–0.86 (ID); nouns per sentence (M): 6.52 (SA)–7.88 (SN)–6.74 (NN)–7.79 (ID).

Therefore, is the discourse of informal debates (ID) a kind of *second specialized non-academic discourse*? And if so, in what ways is it useful? The answer to these questions can be found in Table 3 and Figure 2, which point to the DFA modeling success in categorizing efficiently the 200 texts making up the corpus. The surprising element here is not necessarily the high F1 score for academic documents (96.96%), but the fact that, despite their high data dispersion coefficient, informal debate documents obtained a score (76.47%) much higher than the journalistic texts, be these of a specialized nature (55.50%) or not (57.42%). At the same time, a relevant aspect is also the high degree of confusion between specialized non-academic texts—which correspond, by and large, to the journalistic category of *feature writing* and, in most cases, illustrate the ideal of quality journalism—and non-specialized non-academic texts—which, in the field of journalism, are equated with *news story* and are associated with the tabloid press[32]: 30% of the texts in each category tend to be mistaken for the discourse of the neighboring category.

Such values confirm that a process already witnessed in and signaled by Western journalism was set into motion in Romania as well, and is progressing at an alarming speed: that of turning quality journalism into tabloid press. In fact, a cursory look at the titles in the publications selected for the SN category shows that these too have started to abuse *breaking news* or even *fake news* elements. Specifically, the difference in "style" and approach between the articles published in *Ziarul financiar* and those in *Libertatea* is increasingly smaller. Therefore, it appears that the proliferation and development of informal debates platforms aims not to complement, but *to replace specialized non-academic publications* threatened of being devoured by the tabloid press in Romania—and, consequently, to save quality journalism. But is this phenomenon witnessed in the economic discourse only or does it affect all ranges of Romanian media today? We incline toward the latter scenario; yet, only further automated analyses could reveal the whole dimension and spread of this trend.

❑

Notes

1. Tyler Cowen and Alex Tabarrok, *Modern Principles of Economics*, 2nd edition (New York: Worth Publishers, 2013), xxvii.
2. David McNally, *Bodies of Meaning: Studies on Language, Labor and Liberation* (Albany: SUNY Press, 2001), 54.
3. Jacob Marshak, "Economics of Language," *Behavioral Science* 10, 2 (1965): 135–140.
4. François Grin, "The Economics of Language: Survey, Assessment, and Prospects," *International Journal of the Sociology of Language* 121, 1 (1996): 17–44.
5. Barry R. Chiswick and Paul W. Miller, eds., *The Economics of Language: International Analyses* (London–New York: Routledge, 2007).
6. Honesto Herrera-Soler and Michael White, eds., *Metaphor and Mills: Figurative Language in Business and Economics* (Berlin–Boston: Mouton de Gruyter, 2012).
7. Zubin Jelveh, Bruce Kogut, and Suresh Naidu, "Political Language in Economics," *Columbia Business School Research Paper*, 14–57 (2018), http://dx.doi.org/10.2139/ssrn.2535453.

8. See Angela Rogojanu, *Comunicare şi limbaj economic*, foreword by Alexandru Taşnadi (Bucharest: Editura ASE, 2003).
9. Doina Butiurcă, "Fundamentele lingvistice ale terminologiei economice," *Studia Universitatis Petru Maior: Philologia* 16 (2014): 34–45.
10. Elena Museanu, "Criză, recesiune, depresiune, tip de paradigmă?" in *Limba română: controverse, delimitări, noi ipoteze: Actele celui de-al 9-lea Colocviu al Catedrei de Limba Română*, 1, Section *Lexic, semantică, terminologii*, eds. Rodica Zafiu, Adina Dragomirescu, and Alexandru Nicolae (Bucharest: Editura Universităţii, 2010) , 283–288; id., "Extralingvistic şi lingvistic în terminologia economică actuală," in *Limba română: direcţii actuale în cercetarea lingvistică: Actele celui de-al 11-lea Colocviu Internaţional al Departamentului de Lingvistică*, 2, eds. Rodica Zafiu, Adina Dragomirescu, and Alexandru Nicolae (Bucharest: Editura Universităţii, 2012), 143–150.
11. Nicoleta Călina, "Criza economică mondială şi strategiile comunicative sau lipsa acestora în limbajul politic recent italian şi român," *Revista de Ştiinţe Politice* 8, 30–31 (2011): 166–172; Aura Iordache, "Metafore ale crizei în textul de presă: metafora verbală," in *Cercetarea lingvistică în actualitate*, ed. Anca Cosăceanu (Iaşi: Institutul European, 2012), 96–110.
12. Daniela Gîfu and Dan Cristea, "Public Discourse Semantics: A Method of Anticipating Economic Crises," *International Journal of Computer Communications and Control* 7, 5 (2012): 832–839.
13. Mihai Dascalu et al., "Mining Texts, Learner Productions and Strategies with ReaderBench," in *Educational Data Mining: Applications and Trends*, ed. A. Peña-Ayala (Cham, Switzerland: Springer, 2014), 345–377; Mihai Dascalu et al., "Please ReaderBench this Text: A Multi-Dimensional Textual Complexity Assessment Framework," in *Tutoring and Intelligent Tutoring Systems*, ed. Scotty Craig (Hauppauge, NY: Nova Science Publishers, 2018), 251–271.
14. Ibid.
15. Mihai Dascalu, Danielle S. McNamara, Stefan Trausan-Matu and Laura K. Allen, "Cohesion Network Analysis of CSCL Participation," *Behavior Research Methods* 50, 2 (2018): 604–619, doi: 10.3758/s13428-017-0888-4.
16. Alexander Budanitsky and Graeme Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics* 32, 1 (2006): 13–47.
17. Thomas K. Landauer, Peter W. Foltz, and Darrell Laham, "An introduction to Latent Semantic Analysis," *Discourse Processes* 25, 2–3 (1998): 259–284.
18. David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.
19. Thomas Mikolov et al., "Efficient Estimation of Word Representation in Vector Space," Workshop at ICLR, Scottsdale, AZ, 2013.
20. Claude E. Shannon, "Prediction and Entropy of Printed English." *The Bell System Technical Journal* 30, 1 (1951): 50–64.
21. Vincenzo Gervasi and Vincenzo Ambriola, "Quantitative Assessment of Textual Complexity," in *Complexity in Language and Text*, ed. Lavinia Merlini Barbaresi (Pisa: Plus, 2002), 197–228.
22. Verginica Barbu Mititelu et al., "Improving parsing using morpho-syntactic and semantic information," *Romanian Journal of Human-Computer Interaction* 9, 4 (2016): 285.
23. Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský, "Prague Dependency Treebank," in *Handbook of Linguistic Annotation* (Berlin: Springer, 2017), 555–594.
24. Teun A. van Dijk and Walter Kintsch, *Strategies of Discourse Comprehension* (New York: Academic Press, 1983).
25. Stephen T. O'Rourke, Rafael A. Calvo, and Danielle S. McNamara, "Visualizing Topic Flow in Students' Essays," *Educational Technology & Society* 14, 3 (2011): 4–15.

26. Budanitsky and Hirst.
27. George A. Miller, "WordNet: A lexical database for English," *Communications of the ACM* 38, 11 (1995): 39–41.
28. Mihai Dascalu et al., "ReaderBench—Automated Evaluation of Collaboration based on Cohesion and Dialogism," *International Journal of Computer-Supported Collaborative Learning* 10, 4 (2015): 395–423.
29. John Scott, *Social Network Analysis*, 3rd edition (Los Angeles–London–Thousand Oaks, CA: SAGE Publications, 2012).
30. Michel Galley and Kathleen McKeown, "Improving Word Sense Disambiguation in Lexical Chaining," conference paper at the 18th International Joint Conference on Artificial Intelligence (IJCAI'03), Acapulco, Mexico, 2003.
31. G. David Garson, *Multivariate GLM, MANOVA, and MANCOVA* (Asheboro, NC: Statistical Associates Publishing, 2015).
32. For the distinction between feature writing and news story, see Vir Bala Aggarwal, *Essentials of Practical Journalism* (New Delhi: Concept, 2006), 300–302.

Abstract

Discourses of Economic Crisis in Romanian Media: An Automated Analysis Using the *ReaderBench* Framework

Not many people are specialists in economics, but everyone tends to have the impression of knowing something about economic crises—and, above all, they are eager to talk about it. The economic crisis seems to have been the dominant topic of the last decade at least in Romania, generating a plurality of parallel or even contradictory discourses. This paper explores, through an automated analysis using the *ReaderBench* Framework, Romanian media discourses on the economic crisis spanning the years between 2008 and 2018. We start by delineating four categories of discourse relevant for the selected topic: specialized academic publications (SA), specialized non-academic publications (SN), non-specialized non-academic publications (NN), and informal debate platforms (ID). For each of these categories, we selected 50 texts, which were then analyzed using the *ReaderBench* Framework. The 13 statistically significant textual complexity indices that resulted from this automated analysis show that the first three categories of discourse follow a downward progression pattern, while the fourth category (ID) appears to exhibit an eccentric behavior. The central thesis of our article is that this behavior stems from the ID's attempt to replace the specialized non-academic discourse, currently under threat of being devoured by the tabloid press in Romania.

Keywords

discourse of economic crisis, types of discourse, Romanian media, Natural Language Processing, *ReaderBench* Framework, textual complexity indices