Complexity and Style of Modern Spanish Literary Texts

Katharina Dziuk Lameira

Abstract This article discusses whether text complexity can be seen as a dimension of authorial style, and if so, which linguistic features are suitable for the description of both complexity and style. First, the notion of text complexity will be defined, followed by a presentation of its evolution from readability studies. In addition, a comparison regarding definitions of text complexity and literary style will be followed by various parameters of text complexity used in a research project on the linguistic complexity of modern Spanish literary texts conducted at the University of Kassel, presented together with the tools and methods used for their analysis. In conclusion, an investigation concerning a selection of parameters and data from the project regarding a stylistic point of view will be offered, prior to the results being discussed and future research steps being proposed. First analyses show that metaphorical density, sentence length, clause length, the index of subordination, the density of the noun phrase, and logarithmic average (word) frequency and combinations of some of those parameters could be suitable for the description of some of the novels analyzed in this article.

Keywords text complexity, authorial style, readability, Spanish, literary texts, metaphor identification

1. Introduction

Text complexity is defined as the interaction of different textual features that influence text difficulty and can be measured and observed objectively (Dziuk Lameira 2023). Following Dahl's definition of relative complexity (2004), a text can be seen as more complex the more it deviates from typical patterns.[1] Analogously, different definitions

1 *Pattern* refers here to a variety of phenomena such as linguistic patterns like phrasemes or patterns regarding textual genres.

Dziuk Lameira, Katharina: Complexity and Style of Modern Spanish Literary Texts, in: Hesselbach, Robert, et al. (Eds.): Digital Stylistics in Romance Studies and Beyond, Heidelberg: Heidelberg University Publishing, 2024, pp. 279–297. DOI: https://doi.org/10.17885/heiup.1157.c19376

for literary style have been proposed, e.g. as the interaction of textual features or deviation from patterns in texts.[2] These parallel definitions lead to the questions of if and to what extent text complexity can be seen as a dimension of literary style and vice versa. Given that many of the variables investigated in a current research project conducted at the University of Kassel by Dziuk Lameira on linguistic complexity of modern Spanish texts are based on textual features that are often used in stylistics as well, the data from the project will be explored in this article from a stylistic point of view.

2. Text Complexity

2.1 The Notion of *Text Complexity*

Text complexity is investigated here from a linguistic point of view. It can be defined as a property of a text that emerges from the interaction of different text features and levels that can be measured and observed objectively (Dziuk Lameira 2023). *Text complexity* and *text difficulty* are seen here as different concepts. *Text difficulty* is considered a consequence of the interaction of text complexity and extralinguistic parameters, such as reader or task characteristics (e.g. motivation, cognitive capabilities or type of task). The definition offered here regarding linguistic text complexity also excludes the difficulty of content that can also lead to text difficulty.

2.2 From Readability to Text Complexity

Research on text complexity can be traced back to the first readability formulae that have been developed mostly by psychologists and educational researchers since the 1920s in order to predict the comprehensibility of texts for specific groups of readers. Mikk describes the standard way of finding a readability formula as follows:

> To elaborate a readability formula, a sample of texts, representative of the texts in the area of intended formula application, should be taken. The comprehension level of the texts is measured by some experiments and the texts are analysed to establish the values of the hypothesised measures for the text comprehensibility. The comprehension level and the comprehensibility measures are tied in a formula by multiple regression analysis. The most valid compre-

2 See Fix (2009) for an overview on deviation and pattern in rhetoric and stylistics.

> hensibility measures intercorrelations of which are low [sic] are included in the readability formula as the result of the analysis (2005, 913).

Which comprehensibility measures are valid can vary depending on text type and target group, e.g. Mikk found over 200 valid comprehensibility measures for popular scientific texts read by students (2000). However, many studies have shown that word difficulty (mostly measured by word length and/or word frequency) and sentence difficulty (mostly measured by sentence length), account for most of the variance in readability measures, as the addition of other variables could scarcely improve the prediction of comprehensibility (Entin and Klare 1978; Klare 1974; 1984). For this reason, most readability formulae make use of those variables. Readability formulae are often criticized for only relying on superficial textual features neglecting reader and task characteristics (although those can also be integrated into a formula, e.g. Mikk and Elts 1999). The terminological shift from *readability* to *text complexity* illustrates the attempt to integrate quantitative and qualitative text features and reader and task characteristics to a model of text complexity.[3]

2.3 Complexity: A Complex Notion

Complexity itself is a complex notion, as there is no generally accepted definition of complexity and the existing definitions vary depending on the discipline or the object of investigation. The Santa Fe Institute defines complexity as follows: "In general, the complexity of a system emerges from the interactions of its interrelated elements as opposed to the characteristics of those elements in and of themselves" (Santa Fe Institute 2018). Rescher defines it from a philosophical point of view in the following way: "Complexity is first and foremost a matter of the number and variety of an item's constituent elements and of the elaborateness of their interrelational structure, be it organizational or operational" (1998, 1). Both definitions underline the importance of interrelation and interaction in complex systems. According to Rescher (1998)[4], the notion of complexity encompasses various modes (see Table 1).

First, Rescher mentions the EPISTEMIC MODES of complexity: *descriptive complexity*, which refers to the length of a description of a system; *generative complexity* or the number of instructions necessary to generate a system; and *computational complexity* referring to the time and resources needed to solve a problem.

3 See e.g. the Common Core State Standards model of Text Complexity (CCSSO 2010).

4 Rescher's definition and systematization of complexity have been part of the discussion about complexity in linguistics since their reception by Karlsson, Miestamo, and Sinnemäki (2008).

Table 1 Modes of Complexity according to Rescher (1998, 9)

Epistemic modes	
Formulaic complexity	*Descriptive complexity*
	Generative complexity
	Computational complexity
Ontological modes	
Compositional complexity	*Constitutional complexity*
	Taxonomic complexity (or heterogeneity)
Structural complexity	*Organizational complexity*
	Hierarchical complexity
Functional complexity	*Operational complexity*
	Nomic complexity

Secondly, the author lists the ONTOLOGICAL MODES of complexity, which are compositional, structural, and functional complexity.

Compositional complexity is divided into *constitutional complexity*, which refers to the number of elements that constitute an object (relating to texts this could be the number of words or paragraphs); and *taxonomic complexity (*or *heterogeneity)*, which refers to the different types of elements in a system (e.g. modes and tenses used in a text).

Moreover, there are two types of structural complexity: *organizational* and *hierarchical complexity*. *Organizational complexity* alludes to the number of different ways to organize the different elements of an object; whereas *hierarchical complexity* refers to the degree of elaborateness of the hierarchical relations in the system (e.g. levels of syntactic subordination).

The last type that Rescher mentions is functional complexity, which encompasses *operational* and *nomic complexity* (1998, 9). *Operational complexity* relates to the number of different possible functions and states that can arise during a process. The higher the number of possible states, the less predictable the behavior of a system is (which leads to higher complexity) (Rescher 1998, 12–13). *Nomic complexity* finally refers to the structures and laws that govern a system (ibid., 13).

2.4 Absolute and Relative Complexity

In the study of linguistic complexity, the differentiation between *absolute* and *relative complexity* is often made. According to Miestamo, absolute complexity is theory-oriented and objective whereas relative complexity is receiver-oriented and subjective (2006, 2008). The relative notion of complexity can be illustrated by the question:

"Complex to whom?" (Miestamo 2006, 3; Kusters 2003, 6) and Miestamo argues that this form of complexity should be called "cost" or "difficulty" (2008, 25–26).

Dahl uses the term *relative complexity* in a different sense:

> An entity E would have a certain complexity relative to a description or theory T measured by the length of the additional description necessary to characterize E provided that T is already given. […] A theory of a class of entities may specify (or predict) the properties that are common to all the members of the class. However, it may go beyond that and also specify properties that are typical of or 'normal' for the members of the class—what holds in a default or prototypical case. The description of each member may then be considerably simplified, given that only deviations from the normal case have to be specified. An interesting consequence that now appears is that an entity which deviates from the default case in more respects will tend to be more complex, in this sense (2004, 25–26).

According to him, a deviation from the normal case leads to a longer description and thereby to higher complexity. He exemplifies this by saying that when describing a person, we don't have to mention that "she has two legs, two arms and one head" (Dahl 2004, 25), because we know what human beings look like in general. Transferred to text complexity, this could mean that the complexity of a text is higher the more it deviates from the typical text of its class, because the description of its properties is longer.

In a similar way, Merlini Barbaresi (2011) argues that texts can be seen as complex systems that are characterized on different levels by "markedness/naturalness," in which markedness leads to higher degree of complexity (for a criticism on the use of the term *markedness* see Haspelmath 2006).

2.5 Text Complexity and Text Difficulty

Text complexity can be defined as a property of a text that emerges from the interaction of different text features and levels that influence text difficulty and can be measured and observed objectively, whereas text difficulty can be seen as the consequence of the interaction of text complexity and extralinguistic parameters that can be perceived subjectively (Dziuk Lameira 2023). Extralinguistic parameters are reader characteristics such as age, educational background, reading experience, motivation etc.; and task characteristics such as the translation of a text versus the retrieval of information. Thus, *text complexity* can be seen as the cause and *text difficulty* as the effect.

According to Rescher, complexity and cognitive difficulty are two different concepts that are coordinated:

> As an item's complexity increases, so do the cognitive requisites for its adequate comprehension, although, of course, cognitive ineptitude and mismanagement can manage to complicate even simple issues. All the same, our best practical index of an item's complexity is the effort that has to be expended in coming to cognitive terms with it in matters of description and explanation (1998, 1).

Another important issue is how linguistic text complexity can be measured, or if it is even possible to measure text complexity. If we assume that it is possible, there are two possibilities: First, text complexity can be measured directly by means of measurement categories like the type and number of relations between elements etc.[5]

Second, complexity could be measured indirectly by measuring its effect (e.g. the cost caused by the complexity). In the case of texts, the cost of complexity could be the difficulty that a reader experiences reading the text.

However, according to the Santa Fe Institute, which is known for its studies on complex (adaptive) systems, "complex behavior generally cannot be reduced to, or derived from, the sum of the behavior of the system's components" (Santa Fe Institute 2018). Although the definition of complex adaptive systems is not perfectly fitted to texts, many complexity definitions emphasize the important role of interrelations in complex systems. A model of text complexity has to be able to measure or at least describe these interrelations. An interesting approach comes from Stede (2018), who suggests that the interplay of features on different text levels can be analyzed by annotating the structure of texts in a multilevel annotation in order to find correlations between the different text levels. The approach is called *level-oriented text linguistics* ("Ebenen-orientierte Textlinguistik") and presents different tools for the analysis of separated text levels as well as a database that comprises all annotations (Stede 2018). In the case that those multilevel annotations don't exist, the qualitative text analyses (e.g. semantic text analysis according to Gardt 2012) can help to comprehend the nature of interactions in texts.

2.6 Text Complexity and Literary Style

The comparison of definitions of the notions of complexity and literary style show two parallel branches:

1. the definition of text complexity and literary style as ensembles of textual features
2. the definition of text complexity and literary style as deviation

5 See Rescher's modes of complexity (section 2.3)

As mentioned above, text complexity is defined here as a property of a text that emerges from the interaction of different text features and levels that influence text difficulty and can be measured and observed objectively. Another emergent feature of literary texts is their style. The notions of complexity and style of a text are defined in similar ways. According to Herrmann, van Dalen-Oskam, and Schöch "[s]tyle is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively" (2015, 44). The emergent character of *style* is not mentioned by the authors, but should be considered, as well. The main difference between the two notions is the effect, which is difficulty in the case of complexity, and the perception of an (individual) style in the case of style.

The second possible definition of *style* is style as deviation from typical patterns or reader expectations,[6] which connects back to Dahls' definition of complexity as a deviation from the normal case.

As the definitions of *style* and *complexity* are very similar, it is valid to ask if linguistic text complexity can be seen as a dimension of literary style and vice versa.

This can be understood in two ways:

— the overall complexity of a text could be seen as a dimension of style
— the parameters (text features) that are suited to measure or describe complexity could also be suited to identify a style

One difference between *style* and *complexity* is that the description of style is often not made gradually but, in the case of author identification, with the aim being to attribute a text to a specific author. Complexity on the other hand is always a matter of graduality or degree.

As the description of the overall complexity of a text can vary depending on the parameters chosen by the researcher, the only way to see if the overall complexity of a text can be related to its style would be to study this question indirectly by comparing the difficulty of texts judged by readers or experts with a facet of style, e.g. authorship or register. Another possibility would be to study isolated or combined text features for their suitability to discriminate a certain facet of style. It is this second possibility that is being tested in this paper.

6 See Fix (2009) for an overview on deviation and pattern in rhetoric and stylistics.

3. Analyzing Text Complexity

3.1 Project on Linguistic Complexity Profiles of Modern Spanish Texts

In a current research project conducted at the University of Kassel (Germany) by Dziuk Lameira, first the complexity of a text collection consisting of 30 Spanish literary text excerpts from novels published between 2004 and 2017 is analyzed quantitatively and qualitatively regarding lexical, semantic, syntactical, morphological, and textual features to provide a complexity profile for each text. One half of the novels were read in the B1 Spanish course and the other half in the B2 course at the University of Kassel. The choice of the novels was based on didactic considerations. In a second step, six representative texts were chosen from the corpus for an online-questionnaire developed by Friedrich (2017) that was presented to German-speaking students of Spanish philology which were grouped according to their Spanish proficiency level (CEFR[7] levels A2–C1). The students rated the difficulty of the given texts and answered questions concerning the difficulty of those texts. The mean values of these text ratings were used for a statistical analysis to identify text features contributing to text complexity. In this paper, various parameters that are usually used to measure text complexity will be tested for their suitability for distinguishing different authorial styles.

The analyzed text excerpts were taken from the following novels (three excerpts per novel):

— Javier Cercas: *El impostor* (2014)
— Juan Gabriel Vásquez: *El ruido de las cosas al caer* (2010)
— Almudena Grandes: *El lector de Julio Verne* (2012)
— Javier Marías: *Así empieza lo malo* (2014)
— Eduardo Mendoza: *Riña de gatos. Madrid 1936* (2010)
— Juan José Millás: Hay algo que no es como me dicen (2004)
— Edurne Portela: *Mejor la ausencia* (2017)
— Carme Riera: *Naturaleza casi muerta* (2011)
— Andrea Stefanoni: *La abuela civil Española* (2014)
— Andrés Trapiello: *Ayer no más* (2012)

The excerpts were compiled based on the following criteria: They are between 308 and 353 words long and are understandable without context in order to be used in the questionnaire. Furthermore, they contain no direct speech (with or without *inquit*

7 Common European Framework of Reference for Languages.

formulae) and no parts of the original text are omitted or changed. Additionally, they contain at least one metaphor and can be assigned to narrative and/or descriptive text types. The analyzed text compilation contains 9,759 words in total.

3.2 Quantitative Complexity Parameters

The quantitative analysis was carried out by means of the program TRUNAJOD (Véliz and Karelovic) and different readability formulae like Fernández Huerta (1959), Gutiérrez (1972), Szigriszt Pazos (1993), INFLESZ (Barrio-Cantalejo et al. 2008), legibilidad μ (Muñoz Baquedano and Muñoz Urra 2019) etc., which are validated or adapted for Spanish texts and available online.[8]

The program TRUNAJOD (Véliz and Karlovic) can determine the readability of texts in Spanish and is compatible with the tagger Connexor Machinese Syntax. It calculates the following indices:

— LO (*longitud de la oración*): sentence length
— LC (*longitud de la cláusula*): clause length
— IS (*índice de subordinación*): index of subordination
— DeP *(índice de densidad proposicional*): propositional density
— DeL (*índice de densidad léxica*): lexical density
— DiL (*índice de diversidad léxica*): lexical diversity
— DFN (*densidad de la frase nominal*): density of the noun phrase
— FP (*frecuencia promedio de palabras*): average word frequency
— FPL (*frecuencia promedio logarítmica*): logarithmic average frequency

3.3 Semantic Complexity Parameters

Semantic text complexity consists of predictors based on semantic information. It is contested whether semantic text features are suitable parameters for the measurement or description of text complexity and readability. François and Fairon developed a model of readability for learners of French as a foreign language based on support vector machines which allows a better prediction of readability than classic readability formulae (2012, 466). According to the authors "the information carried by semantic predictors is largely correlated with that of lexico-syntactical ones" (ibid., 475). Their explanation is that "semantic and lexical predictors are correlated because the methods used for the parameterization of the semantic factors heavily rely on lexical

8 E.g. https://legible.es or LEXILE.

information. This is the case for the LSA,[9] as well as for the propositional approach of the content density" (ibid.). This observation leads to the question of how the semantic dimension of text complexity can be operationalized.

3.4 Metaphorical Text Complexity

In order to explore possibilities to operationalize the semantic dimension of text complexity, the field of metaphorical text complexity was focused on in the project mentioned above.

The following parameters were thus suggested for the measurement of metaphorical text complexity (see also Dziuk Lameira 2019; Dziuk Lameira 2023):

— METAPHORICAL DENSITY: The number of metaphor related words divided by the number of lexical units[10] multiplied by 100 (MD = MRW[11]/lexical units × 100)
— METAPHORICAL VARIETY: The number of different concept combinations (conceptual metaphors) divided by the number of metaphor related words (MV = Number of different concept combinations/MRW)
— NNMRW: The percentage of non-nominal metaphor related words within the total of metaphor related words (Percentage of non-nominal MRWs)
— EXTENDED METAPHORS: The number of extended metaphors (EM) divided by the number of lexical units multiplied by 100 (EM/lexical units × 100)
— DEGREE OF CONVENTIONALIZATION: The percentage of lexicalized and new metaphors per text

Other parameters that are analyzed qualitatively include: CO-TEXTUALIZATION (Skirl 2009), the interaction with similes,[12] and the revitalization of metaphors (Goatly 1997). Metaphors were counted as proposed by MIPVU (Metaphor Identification

9 Latent Semantic Analysis (LSA), also known as latent semantic indexing, is a method developed by T. K. Landauer for the semantic analysis of document collections. The objective is to extract latent concepts, which are relevant terms and co-occurrences, from the documents. In LSA analysis, documents are represented in the form of a term-document matrix (vector space model), common stop words are filtered out, frequently occurring terms are specifically weighted, and a singular value decomposition is performed (a method of linear algebra to reduce the number of dimensions per document) (Glück and Rödel 2016, 390).
10 Unit of analysis when applying the Metaphor Identification Procedure Vrije Universiteit (MIPVU) (Steen et al. 2010).
11 Metaphor related words (see Steen et al. 2010).
12 This parameter could be as well seen as a subparameter of co-textualization.

Procedure Vrije Universiteit, Steen et al. 2010) in order to increase the objectivity of metaphor identification.

The procedure was developed by Pragglejaz Group (2007) as MIP (Metaphor Identification Procedure) and further developed by Steen et al. (2010) under the name of MIPVU. The method consists of the following steps (ibid., 25–26):

1. Find metaphor related words (MRWs) by examining the text on a word-by-word basis.
2. When a word is used indirectly and that use may potentially be explained by some form of cross-domain mapping from a more basic meaning of that word, mark the word as metaphorically used (MRW).
3. When a word is used directly and its use may potentially be explained by some form of cross-domain mapping to a more basic referent or topic in the text, mark the word as direct metaphor (MRW, direct).
4. When words are used for the purpose of lexico-grammatical substitution, such as third person personal pronouns, or when ellipsis occurs where words may be seen as missing, as in some forms of coordination, and when a direct or indirect meaning is conveyed by those substitutions or ellipses that may potentially be explained by some form of cross-domain mapping from a more basic meaning, referent, or topic, insert a code for implicit metaphor (MRW, implicit).
5. When a word functions as a signal that a cross-domain mapping may be at play, mark it as a metaphor flag (MFlag).
6. When a word is a new-formation coined, examine the distinct words that are its independent parts according to steps 2 through 5.

Although first attempts to automatize metaphor identification exist (e.g. Berber Sardinha 2010; Rai et al. 2016; Rai and Chakraverty 2017), they are still not ready for application and therefore the metaphor identification has to be carried out manually.

4. Analysis

4.1 Preliminary Results

In the following, the complexity parameters calculated by the software TRUNAJOD and one parameter for metaphorical complexity (MD: metaphorical density) have been tested for their ability to discriminate different authorial styles. As the text compilation used for the analyses contains 30 excerpts from ten different novels written by different authors, *authorial style* was here defined as the affiliation of a text excerpt to a novel. Since the text compilation includes three excerpts per novel, the question is whether

Table 2 Kruskal Wallis Test

	MD	LO	LC	IS	DeP	DeL	DiL	DFN	FP	FPL
Kruskal Wallis H	19,916	19,957	17,434	17,730	9,713	10,729	9,959	18,528	12,742	18,841
df	9	9	9	9	9	9	9	9	9	9
Asymptotic significance	*0.018*	*0.018*	*0.042*	*0.038*	0.374	0.295	0.354	*0.030*	0.175	*0.027*

the chosen quantitative variables (MD: metaphorical density, LO: sentence length, LC: clause length, IS: index of subordination, DeP: propositional density, DeL: lexical density, DiL: lexical diversity, DFN: density of the noun phrase, FP: average frequency and FPL: logarithmic average frequency) show a significant difference between novels.

Table 2 shows the results of the Kruskal Wallis Test. The null hypothesis that the distribution of a variable is the same across the novels is rejected for the variables MD, LO, LC, IS, DFN and FPL. Hence, there are significant differences between novels for these variables.

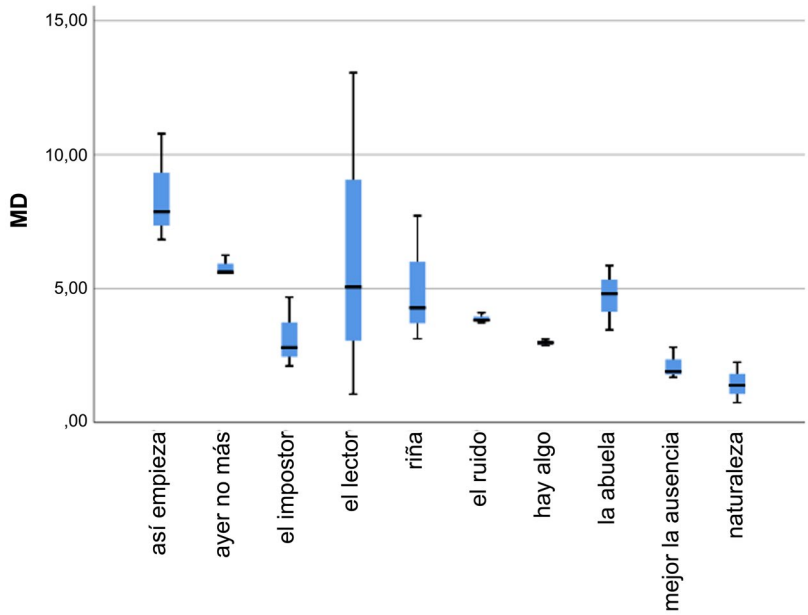

Fig. 1 Boxplot comparison for variable MD (Dziuk Lameira, CC BY).

The boxplot diagrams in Figure 1 visualize the spread within novels as well as differences between medians for the variable MD (metaphorical density). Although tendencies can be recognized, the pairwise comparison of novels for MD using t-test only shows a significant difference between the novels *Naturaleza casi muerta* and *Así empieza lo malo* (see Table 3). This shows that metaphorical density can be a valid criterion for the differentiation of novels and therefore should be further investigated.

Table 3 Pairwise comparison of novels for the variable MD

Sample 1-Sample 2	N	Test Statistic	Std. error	Std. Test Statistic	Sig.	Adj. Sig.[a]
naturaleza-así empieza	3	24,000	7,188	3,339	0.001	0.038

a Significance values have been adjusted by the Bonferroni correction for multiple tests.

When using the Bonferroni-Holm correction for multiple tests the only significant difference remains between the novels *Naturaleza casi muerta* and *Así empieza lo malo*.

At the same time, Figure 1, as well as Figure 2, shows the high spread of the MD and LO values within some of the novels, e. g. the three excerpts from *El lector de Julio Verne* written by Almudena Grandes show a high spread for the variable MD (metaphorical density) whereas the three excerpts from the novel *El impostor* by Javier Cercas show a high spread for the variable LO (sentence length). (the same could be observed for the other analyzed variables). The high deviation paired with the small sample size per novel influences the suitability of the variables.

The results for the Pearson test show a very high correlation for the variables LO (sentence length) and IS (index of subordination) ($r = 0.95$, $p < 0.001$) as well as for the variables FP (average frequency) and FPL (logarithmic average frequency) ($r = 0.875$, $p < 0.001$). The reason for the correlation between LO and IS could be that longer sentences have a higher probability of containing more clauses. FP and FPL could be correlated because both variables measure word frequency. Because of the correlation, one of each correlating variables was discarded from the following cluster analysis. In the case of FP and FPL, FPL, the logarithmic average frequency, was maintained, because high frequency words have a lower impact when calculating the index. In the case of the correlating variables LO and IS, LO, the sentence length, was maintained, because many studies have shown its importance as a predictor of text comprehensibility (Entin and Klare 1978; Klare 1974; 1984).

The dendrogram in Figure 3 shows the result of the cluster analysis using the variables MD (metaphorical density), LO (sentence length), LC (clause length), DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) and FPL (logarithmic average frequency). The cluster analysis was performed with squared Euclidean distance, z-score standardization, and Ward linkage. Looking at the distribution of the text excerpts when 10 clusters are defined (see

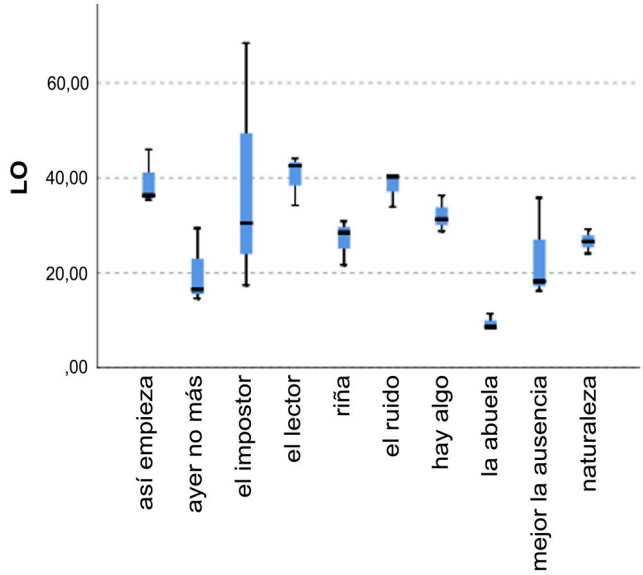

Fig. 2 Boxplot comparison for variable LO (Dziuk Lameira, CC BY).

Table 4 Overview of clusters

Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Así empieza 1	Ayer no más 1	Ayer no más 2	Ayer no más 3	Impostor 3
Así empieza 2	Impostor 1	La abuela 2	Mejor la ausencia 2	
Así empieza 3	Impostor 2	La abuela 3		
Cluster 6	**Cluster 7**	**Cluster 8**	**Cluster 9**	**Cluster 10**
El lector 1	El lector 2	El lector 3	Rina 1	La abuela 1
	Hay algo 2	El ruido 3	Rina 2	Mejor la ausencia 1
	Naturaleza 1		Rina 3	
	Naturaleza 2		El ruido 1	
	Naturaleza 3		El ruido 2	
			Hay algo 1	
			Hay algo 3	
			Mejor la ausencia 3	

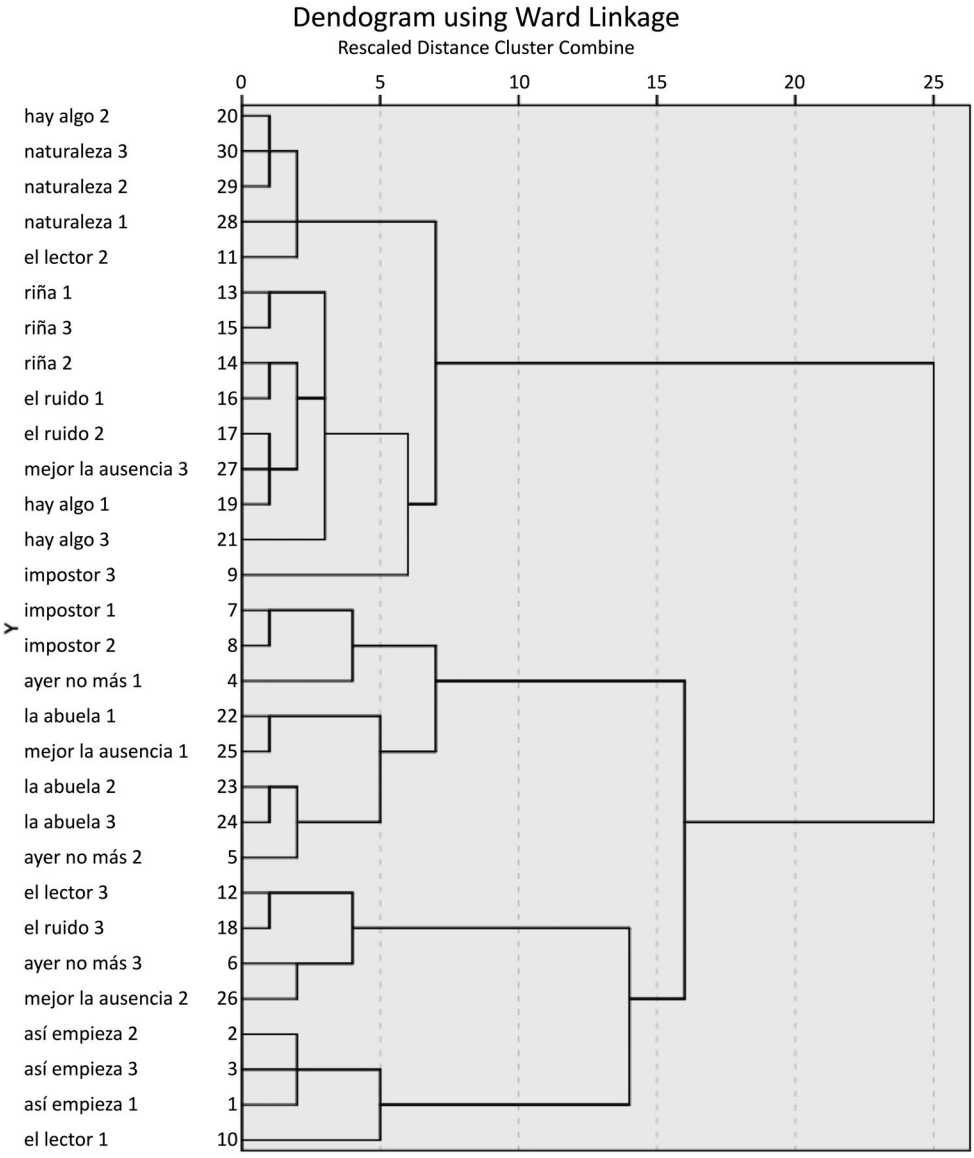

Fig. 3 Dendrogram (Dziuk Lameira, CC BY).

Table 5 Clusters 1, 7 and 9

	Cluster 1 (high MD and high FPL)			Cluster 7 (low MD and low FPL)			Cluster 9 (high LC and low FPL)		
	N	Mean	Standard deviation	N	Mean	Standard deviation	N	Mean	Standard deviation
MD	3.00	8.49	1.18	5.00	1.71	0.43	8.00	3.83	0.63
LO	3.00	39.23	3.39	5.00	32.06	3.64	8.00	31.40	1.96
LC	3.00	9.00	0.49	5.00	9.48	0.30	8.00	10.68	0.35
DeP	3.00	43.33	1.20	5.00	42.20	0.58	8.00	39.25	0.62
DeL	3.00	48.33	0.67	5.00	50.40	0.40	8.00	50.88	0.58
DiL	3.00	64.00	0.58	5.00	66.60	0.87	8.00	64.50	0.71
DFN	3.00	6.93	0.90	5.00	4.06	0.62	8.00	3.21	0.15
FPL	3.00	269.53	32.97	5.00	122.54	5.30	8.00	127.88	10.12

Table 4), the excerpts from the novels *Así empieza lo malo* (Cluster 1), *Naturaleza casi muerta* (Cluster 7), and *Riña de gatos* (Cluster 9) are grouped together. The excerpts from *Así empieza lo malo* were the only ones grouped on their own (see Figure 3).

In the following, the clusters 1, 7 and 9 (see table 5) will be described further in order to identify parameters or combinations of parameters which are potentially typical of the authors whose text excerpts were grouped together in the same cluster.

Cluster 1, which includes only the three text excerpts from the novel *Así empieza lo malo*, as well as two excerpts from other novels, is characterized by a high metaphorical density (MD) and a high logarithmic average frequency (FPL) compared to the other clusters. It should be further investigated if the combination of a high metaphorical density and a relatively high logarithmic average word frequency is characteristic for the novel *Así empieza lo malo*, or Javier María's authorial style in general.

In contrast to cluster 1, cluster 7, which includes all three text excerpts from the novel *Naturaleza casi muerta* by Carme Riera as well as two other novel excerpts, has low values for the variables MD (metaphorical density) and FPL (logarithmic average frequency). The combination of a low metaphorical density and a low logarithmic average frequency could therefore be potentially typical of Carme Rieras novel.

Cluster 9 includes all three excerpts from the novel *Riña de gatos. Madrid 1936* by Eduardo Mendoza as well as five other excerpts. It is characterized by a high clause length (LC) and a low FPL (logarithmic average frequency), which could be typical for the style of the novel *Riña de gatos. Madrid 1936*.

The analysis of the clusters showed that the variables DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) showed little deviation throughout the text excerpts. When the cluster analysis was performed without the variables DeP (propositional density), DeL (lexical density), DiL (lexical diversity), DFN (density of the noun phrase) only using the variables MD (metaphorical density), LO (sentence length), LC (clause length), and FPL (logarithmic average frequency), the same text excerpts were grouped together. Thus, the analysis shows that the parameters MD, LO, LC and FPL are more suitable for the discrimination of different styles than the other parameters.

4.2 Discussion

This analysis has shown that some parameters that are used to measure text complexity are significantly different between some of the novels. Therefore, those parameters can potentially be used for stylistic analysis. This is the case for the variables MD (metaphorical density), LO (sentence length), LC (clause length), IS (index of subordination), DFN (density of the noun phrase), and FPL (logarithmic average frequency). Also, the combination of some parameters could be typical of the style of some authors (e.g. a high metaphorical density and a high logarithmic average word frequency for Javier Marías). This should be further investigated using more text excerpts from the same novels as well as from other novels by the same authors.

The relatively short length of the text excerpts of approximately 300 words could explain cases of high standard deviation for the variable LO ('sentence length'). Given that a large amount of sentences included in the text excerpts exceed 100 words, whereas a similar number of sentences consist of less than five words, the average sentence length can vary greatly depending on the selected sample. Therefore, the variable LO is not suitable for the analysis and comparison of relatively short text excerpts.

The cluster analysis did not group the text excerpts consistently according to their novel. Future investigation should include more text excerpts per novel to perform this kind of analysis.

Additionally, the spread of the values for the different variables varies depending on the novel. More samples should be analyzed in order to determine if certain variables are more consistent within particular novels or novels by the same author.

References

Barrio-Cantalejo, Inés María, Pablo Simón-Lorda, M.C. Puerta Melguizo, Isabel Escalona, María Isabel Marijuán, and Pablo Hernando. 2008. "Validation of the INFLESZ scale to evaluate readability of texts aimed at the patient." *Anales del sistema sanitario de Navarra* 31 (2): 135–52.

Berber Sardinha, Tony. 2010. "A Program for Finding Metaphor Candidates in Corpora." *ESPecialist* 31 (1): 49–67.

Council of Chief State School Officers (CCSSO). 2010. *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects. Appendix A*, Washington, DC. https://achievethecore.org/content/upload/corestandards_appendix_a_text_complexity_ela.pdf (Accessed July 16, 2023)

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity.* Studies in Language Companion Series. Amsterdam: John Benjamins.

Dziuk Lameira, Katharina. 2019. "Complejidad Semántica: El Ejemplo de la Metáfora." In *Competencia textual y complejidad textual. Perspectivas transversales entre didáctica y lingüística*, edited by Angela Schrott and Bernd Tesch, 125–45. Berlin: Peter Lang.

Dziuk Lameira, Katharina. 2023. *Textkomplexität und Textverständlichkeit: Studien zur Komplexität spanischer Prosatexte.* Berlin, Boston: De Gruyter.

Entin, Eileen B., and George R. Klare. 1978. "Factor Analyses of Three Correlation Matrices of Readability Variables." *Journal of Reading Behavior* 10 (3): 279–90.

Fernández Huerta, José. 1959. "Medidas sencillas de lecturabilidad." *Consigna* 214: 29–32.

Fix, Ulla. 2009. "Muster und Abweichung in Rhetorik und Stilistik." In *Rhetorik und Stilistik. Ein internationales Handbuch historischer und systematischer Forschung*, vol. 2, edited by Ulla Fix, Andreas Gardt, and Joachim Knape, 1300–15. Berlin, Boston: De Gruyter.

François, Thomas, and Cédrick Fairon. 2012. "An 'AI Readability' Formula for French as a Foreign Language." In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 466–77. Stroudsburg, PA: Association for Computational Linguistics.

Friedrich, Marcus. 2017. *Textverständlichkeit und ihre Messung: Entwicklung und Erprobung eines Fragebogens zur Textverständlichkeit*. Münster, New York: Waxmann.

Gardt, Andreas. 2012. "Textsemantik. Methoden der Bedeutungserschließung." In *Geschichte der Sprache und Sprache der Geschichte. Probleme und Perspektiven der historischen Sprachwissenschaft des Deutschen. Oskar Reichmann zum 75. Geburtstag*, edited by Jochen A. Bär and Marcus Müller, 61–82. Berlin: Akademie-Verlag.

Glück, Helmut, and Michael Rödel (edd.). 2016. *Metzler Lexikon Sprache*. Stuttgart: Metzler.

Goatly, Andrew. 1997. *The Language of Metaphors*. London, New York: Routledge.

Gutiérrez de Polini, Luisa Elena. 1972. "Investigaciones sobre lectura en Venezuela. Informe presentado a las *Primeras Jornadas de Educación Primaria*." Caracas: Ministerio de Educación.

Haspelmath, Martin. 2006. "Against Markedness (and What to Replace It With)." *Journal of Linguistics* 42 (1): 25–70.

Herrmann, Berenike, Karina van Dalen-Oskam, and Christof Schöch. 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52.

Karlsson, Fred, Matti Miestamo, and Kaius Sinnemäki. 2008. "Introduction: The Problem of

Language Complexity." In *Language Complexity: Typology, Contact, Change*, edited by Fred Karlsson, Matti Miestamo, and Kaius Sinnemäki, vii–xiv. Amsterdam: John Benjamins.

Klare, George R. 1974. "Assessing Readability." *Reading Research Quarterly* 10 (1): 62–102.

Klare, George R. 1984. "Readability." In *Handbook of Reading Research,* vol. 1, edited by P. David Pearson. 681–744. London: Routledge.

Kusters, Wouter. 2003. *Linguistic Complexity*. Utrecht: LOT.

Merlini Barbaresi, Lavinia. 2011. "A 'Natural' Approach to Text Complexity." *Poznań Studies in Contemporary Linguistics* 47 (2): 203–36.

Miestamo, Matti. 2006. "On the feasibility of complexity metrics." In *FinEst Linguistics, Proceedings of the Annual Finnish and Estonian Conference of Linguistics*, edited by Krista Kerge and Maria-Maren Sepper, 11–26. Tallinn: Tallinn University Press.

Miestamo, Matti. 2008. "Grammatical Complexity in a Cross-Linguistic Perspective." In *Language Complexity: Typology, Contact, Change*, edited by Fred Karlsson, Matti Miestamo, and Kaius Sinnemäki, 23–41. Amsterdam: John Benjamins.

Mikk, Jaan. 2000. *Textbook: Research and Writing*. Frankfurt am Main, Berlin: Peter Lang.

Mikk, Jaan. 2005. "Text Comprehensibility." In *Quantitative Linguistics: An International Handbook*, edited by Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, 909–21. Berlin, Boston: De Gruyter.

Mikk, Jaan, and Jaanus Elts. 1999. "A Reading Comprehension Formula of Reader and Text Characteristics." *Journal of Quantitative Linguistics* 6 (3): 214–21.

Muñoz Baquedano, Miguel, and José Muñoz Urra. 2019. *Legibilidad Mu*. Viña del Mar, Chile. http://www.legibilidadmu.cl (Accessed April 15, 2019).

Pragglejaz Group. 2007. "MIP: A Method for Identifying Metaphorically Used Words in Discourse." *Metaphor and Symbol* 22 (1): 1–39.

Rai, Sunny, and Shampa Chakraverty. 2017. "Metaphor Detection Using Fuzzy Rough Sets." In *International Joint Conference on Rough Sets*, 271–79. Cham: Springer International Publishing.

Rai, Sunny, Shampa Chakraverty, and Devendra K. Tayal. 2016. "Supervised Metaphor Detection Using Conditional Random Fields." In *Proceedings of the Fourth Workshop on Metaphor in NLP*, 18–27. San Diego, CA: Association for Computational Linguistics.

Rescher, Nicholas. 1998. *Complexity: A Philosophical Overview.* New Brunswick, NJ: Transaction Publishers.

Santa Fe Institute. 2018. *Complexity*. https://www.complexityexplorer.org/explore/glossary/11-complexit (Accessed July 16, 2023)

Skirl, Helge. 2009. *Emergenz als Phänomen der Semantik am Beispiel des Metaphernverstehens. Emergente konzeptuelle Merkmale an der Schnittstelle von Semantik und Pragmatik.* Tübingen: Narr.

Stede, Manfred. 2018. *Korpusgestützte Textanalyse: Grundzüge der Ebenen-orientierten Textlinguistik*. Tübingen: Narr Francke Attempto.

Steen, Gerard J., Aletta G. Dorst, Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A Method for Linguistic Metaphor Identification.* Amsterdam: John Benjamins.

Szigriszt Pazos, Francisco 1993. *Sistemas predictivos de legibilidad del mensaje escrito: Fórmula de Perspicuidad* (Doctoral dissertation) Madrid: Universidad Complutense de Madrid.

Véliz, Mónica, and Bruno Karelovic. "Trunajod." Universidad de Concepción, Chile. http://www.udec.cl/~trunajod/.