Chaos, Solitons and Fractals 163 (2022) 112604

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

journal homepage: www.elsevier.com/locate/chaos

Derivative of a hypergraph as a tool for linguistic pattern analysis

Ángeles Criado-Alonso [a,b], David Aleja [c,e,f], Miguel Romance [c,d,e,f], Regino Criado [c,d,e,f,*]

[a] Grupo de Investigación LIyNMEDIA, Universidad Rey Juan Carlos, c/Tulipán s/n, 28933, Móstoles (Madrid), Spain
[b] Departamento de Filología Extranjera, Traducción e Interpretación, Universidad Rey Juan Carlos, C/Tulipán s/n, 28933, Móstoles (Madrid), Spain
[c] Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, C/Tulipán s/n, 28933 Móstoles (Madrid), Spain
[d] Center for Computational Simulation, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón (Madrid), Spain
[e] Data, Complex Networks and Cybersecurity Sciences Technological Institute, Univ. Rey Juan Carlos, Pza. Manuel Becerra 14, 28028 Madrid, Spain
[f] Laboratory of Mathematical Computation on Complex Networks and their Applications, Universidad Rey Juan Carlos, Calle Tulipán s/n, Móstoles, 28933 Madrid, Spain

ARTICLE INFO

Keywords:
Linguistic patterns
Hypergraph
Derivative of a hypergraph
Higher order network
Dual hypergraph
PageRank
Linguistic stylometry

ABSTRACT

The search for linguistic patterns together with stylometry and forensic linguistics has in the theory of complex networks, its structures and its associated mathematical tools essential resources for representing and analyzing texts. In this paper we introduce a new model able to analyze the mesoscopic relationships between sentences, paragraphs, chapters and texts. This model is supported by several mathematical structures such as the hypergraphs or the concept of derivative graph. The methodology raised from this perspective focuses not only in a quantitative index but also in two peculiar mathematical structures named derivative graph and homogeneity graph. These structures are of singular help to both: detecting the style of an author and determining the linguistic level of a text and, eventually, also for detecting similarities and dissimilarities in texts and even plagiarism.

1. Introduction

In the last decades the emergence of new structures and models in the field of complex networks and the successive advances in the study and development of their associated tools have made it possible to model the different types of interactions between the diverse parts of a complex system in an efficient and remarkably successful way in practically every area of knowledge [1–6]. Complex networks have become an essential and indispensable element in the representation of systems for simulating the interactions and relationships between the components of a complex system in fields as diverse as biology, technology, and human social organization [1,5,7–14].

It is notorious that Network Science can be traced back to the analysis of heterogeneity in real-world complex systems, both in terms of their nature and function. Thus, the role played by some nodes in these systems is very different from the one obtained by the classical Erdős-Rényi model of random networks, which was a first fundamental milestone in the modeling of real-world complex systems and in the assumption in these models as a first level of heterogeneity [15]. The famous scale-free model made it possible to successfully modeling real-world complex systems by highlighting the relevant role of nodes with heterogeneous connectivity [15]. A second milestone consisted in the emergence of multilayer network models having in mind that links could also be heterogeneous in nature [7]. The third milestone is currently being developed under the consideration that the heterogeneity of complex systems may affect not only the function of links, but also their nature, since links may be formed by subsets of nodes of different cardinality [16]. Many complex systems are produced by considering interactions between more than two nodes simultaneously. Thus, from collaborative networks to linguistic networks, including collective social interaction networks, trophic networks and biochemical regulation networks, make the classical network theory a model that is certainly insufficient. Therefore, the new challenge for the network community is to find new mathematical models that fit multiparty interactions to model complex systems with relationships of heterogeneous nature.

The emergence of new tools allowing to automatically handle and analyze large datasets has led to the development of new approaches in many areas of knowledge including text analysis [10,17,18].

Classical approaches for linguistic analysis of texts were based on simple statistical studies that relied on word frequency [10,19]. However, it should be noted that in recent decades modern linguistics

* Corresponding author.
E-mail addresses: angeles.criado@urjc.es (Á. Criado-Alonso), david.aleja@urjc.es (D. Aleja), miguel.romance@urjc.es (M. Romance), regino.criado@urjc.es (R. Criado).

https://doi.org/10.1016/j.chaos.2022.112604
Received 19 July 2022; Received in revised form 20 August 2022; Accepted 22 August 2022
Available online 2 September 2022
0960-0779/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

has received a major breakthrough derived from the treatment of a language as a complex system or network, having at its disposal in this representation all the tools, measures and procedures to obtain a new, efficient and effective approach to the study of languages through complex network that includes qualitative and quantitative aspects [11,12,20–27].

Therefore, the analysis of linguistic theories based in the study of specialized corpora and the new approach provided by complex networks makes it possible to obtain certain stylistic and typological characteristics together with some intrinsic properties of languages. The perspective provided by complex networks goes beyond the use of word adjacency or co-occurrence methods. These classical methods successfully capture the syntactic elements of the texts [28], but do not have the capacity to represent certain characteristics that develop at the mesoscopic level related to the semantic relationships between the different sentences and paragraphs.

The linguistic network model we are working with in this manuscript emerges from the need to work with sentences or paragraphs as a group or collection of certain words in contrast to the type of links considered in previous works where directed and weighted links are devoted to represent the relationships between linguistic units as in [14,21,25]. In this work, as in [29,30], instead of considering the co-occurrence relationship between two adjacent words or linguistic units within a sentence, we will study not only the relationship between sentences (those that share lexical words) but also the relationship between paragraphs or even articles, seeking to characterize, by using network theory parameters, the style of an author of a text as well as the level of language and/or specialization in a text. This approach leads us to a completely different perspective from the one used, for example, in [10], where, among many other differences, words are transformed and reduced to their canonical forms and the text is organized in consecutive sets of paragraphs.

In order to apply the tools described in this work, and perform a computer processing on a linguistic corpus understood as a collection of texts collected electronically as a representative sample of texts selected according to a determined linguistic criteria [31]. Following this, a corpus of texts composed of 86 extended abstracts (volumes 1–6 of the International Journal of Complex Systems in Science (IJCSS), published between April 2011 and November 2016 (http://www.ij-css.org)) has been considered. This corpus provides us with a total amount of 147637 words as well as 25210 sentences, considered in this study. It should be noted that the unit of analysis from which we start in this work is the sentence, i.e., the words enclosed between two periods [32]. In addition, it is important to note that commas and other punctuation marks within the sentence have not been considered for this analysis.

The questions we addressed when we started to write this paper were "How can the competence level of language used in a text be characterized?" or "Can the style of an author be determined using specific parameters in the linguistic network under consideration?". We considered other issues such as: "What is the most frequently used combination of words in a corpus beyond locating the most relevant individual lexical words?", or even "How can the most representative words of a text (not necessarily the most frequent) be determined?".

Taking into account that the English language has four major word classes: nouns, adjectives, verbs, and adverbs, and that the main remaining word classes are prepositions, conjunctions, determiners, interjections, or pronouns, we established in [29] a four-layer network in order to study a specialty language. In this paper, we will focus on words belonging to the lexical layer, i.e., those significant words (mainly nouns and adjectives) with a specific meaning in the specialty language under study [29,30].

Therefore, in this paper we use the tools and methodology derived from some complex network structures to describe interactions between groups of words. Each of these groups is formed by the lexical words belonging to a specific sentence in the analyzed corpus (syntagmatic approach, from the Greek "$\sigma\iota\nu\tau\alpha\gamma\mu\alpha$", syntagma: "assembled group"). It is important to note that the syntagmatic approach, which corresponds to the analysis presented in this paper, is different from the paradigmatic approach used in other works of computational linguistics [21].

Since the syntagmatic relationship is based on the interrelationships of words in a linguistic structure [29,30,33], it makes sense to consider the relationships between sets of two, three or more significant words that appear in the same sentence, paragraph, abstract or article and that in some way characterize a text as belonging to an author, or discriminate the level of language used in it, as well as those other words and relations that allow distinguishing it from texts belonging to other authors or that use a different level of language.

The methodology presented here makes it possible to determine the language level in a text as well as the style of an author. It can be also useful for analyzing and ranking sentences, abstracts, paragraphs and texts (sets of words) according to their importance, having mind their interrelationships in the context of the multilayer network structure defined in [29,30] as well as for extracting new features of a text from the relationships between significant sets of words in the text.

High-order networks or hypergraphs are the natural generalization of networks that takes into account the fact that a link can connect more than two nodes. Interest in this type of network is growing due to the inability of classical graphical representations to describe group interactions. Their applicability goes beyond the field of social sciences [34–36] and the study of group interactions, public cooperation or opinion formation. In our case, we will consider its applicability in the field of linguistics and specialty languages beyond other approaches based on classical complex networks, multiplex networks or multilayer networks [12,14,20–24,37–39].

As it can be easily understood, a property referring to a finite set of objects (in our case, the nodes of a network), is completely characterized by the subset of elements that satisfy it, which in this case will be represented by the hyperedge formed by these elements, making it possible to compare and relate properties of the nodes and the network studying and analyzing the corresponding hypergraph.

Thus, the study of the relationships between the properties of the nodes consists of mathematically analyzing the properties and typical parameters of the associated hypergraph. Therefore, the applications of this methodology to the field of linguistics range from the characterization of an author's style to the detection of plagiarism, including the detection and identification of the same concept expressed in a different way. To this end, starting, in the first instance, from the identification of a sentence of our corpus with the hyperedge formed by the set of lexical words of that sentence, the hypergraph will be constructed in which the nodes will be all the lexical words of the corpus and the hyperedges all the sentences of the corpus, defining the concept of derivative of two words with respect to a set of hyperedges and the degree of independence of two words of a text with respect to that set of hyperedges. This study can be extended considering as hyperedges, successively, the sets of nodes formed by the lexical words of a paragraph, an abstract or even a chapter, taking the corresponding sequence of parameters as a feature of the text and pointing to new applications of this structure.

The structure of the paper is as follows. After this introduction, in Section 2 some basic concepts and a summary of the most important relationships between the line graph the dual hypergraph, the bipartite graph associated to a certain hypergraph and its corresponding matrices are introduced. Section 3 is devoted to introduce the concept of derivative of a hypergraph with respect to a set of nodes and to establish the definition of the homogeneity graph of a hypergraph obtaining some remarkable results related to this new structure. In Section 4 we apply the mathematical concepts and the structures defined in the previous sections to obtain tools able to characterize the style and level of a text belonging to the linguistic hypergraph considered. In Section 5 the lexical density of the set of texts that make up the analyzed corpus

is studied, and some numerical experiments and computational results are presented. Thus, using three different algorithms we illustrate the diverse types of relationships that can be established between sentences within a text and their relative importance. Section 6 is devoted to apply the instruments and tools developed in order to obtain distinctive characteristics that allow us to distinguish the styles of the different authors and linguistic competence levels of the written texts included in the corpus considered. Finally in Section 7 we present some conclusions of this work.

2. Basic concepts and some preliminary results

A network (or graph) $G = (X, E)$ is just a finite set of vertices (or nodes) $X = \{1, \dots, N\}$ connected by a set of edges (or links between certain pairs of nodes) $E = \{e_1, \dots, e_m\}$. If the edges have a direction, we will say that G is a directed network (or digraph). In the sequel, we will denote by $e_{ij} \in E$ the link between the nodes i and j, although sometimes we will also denote the edge e_{ij} by $\{i, j\}$ or, if G is a directed network, by $i \to j$. Finally, a weighted network is a graph in which each edge e_{ij} has an associated numerical value $w(e_{ij}) = w_{ij}$ called its weight. In the same way, following [40], a hypergraph $\mathcal{H} = (X, \varepsilon)$ is a finite set of vertices (or nodes) $X = \{1, \dots, N\}$ and a collection $\varepsilon = \{h_1, h_2, \dots, h_n\}$ of subsets of X such that $h_i \neq \emptyset \quad (i = 1, 2, \dots, n)$ and $X = \bigcup_{i=1}^{n} h_i$. Each of these subsets is called a hyperedge. In this way, hypergraphs appeared as the natural extensions of graphs to describe group interactions. In the following sections, the study is developed with undirected graphs and hypergraphs, though some of the definitions can be easily extended to the directed case.

In order to carry out our study it is necessary to introduce the concepts of linegraph and dual hypergraph of a hypergraph. In this regard it should be noted that the concept of linegraph $L(G)$ associated to a graph $G = (X, E)$ was introduced by H. Whitney in 1932 [41] and extended for higher order networks by J.C. Bermond et al. in 1977 [42, 43]. It is important to point out that the study of these structures, as well as the relationships between them and their applications, has been increasing steadily in recent years (see, for example, [34,44–49]).

So, if $\mathcal{H} = (X, \varepsilon)$ is a hypergraph, the linegraph associated to $\mathcal{H}$ is the graph $L(\mathcal{H}) = (\varepsilon, E')$, where if $h_i, h_j \in \varepsilon$, then

$$\{h_i, h_j\} \in E' \Leftrightarrow h_i \cap h_j \neq \emptyset.$$

It is also notorious that the linegraph $L(\mathcal{H})$ of a hypergraph $\mathcal{H}$ is a graph even though $\mathcal{H}$ is a hypergraph. Note that this concept is a particular case of the concept of intersection graph [49]. On the other hand, it is also possible to consider the dual hypergraph of a hypergraph: if $\mathcal{H} = (X, \varepsilon)$ is a hypergraph, the dual hypergraph associated with $\mathcal{H}$ is the hypergraph $\mathcal{H}^* = (\varepsilon, X')$ in such a way that if $X = \{1, \dots, N\}$, then $X' = \{v_1, \dots, v_N\}$ where $v_i = \{h_j | i \in h_j\}$, $i = 1, \dots, N$. It is not difficult to verify that $(\mathcal{H}^*)^* = \mathcal{H}$. Moreover, if I is the incidence matrix of $\mathcal{H}$, then its transpose matrix I^t is the incidence matrix of $\mathcal{H}^*$. In this context, to concretize the relationship between $L(\mathcal{H})$ and $\mathcal{H}^*$, we consider the function Π_2 that turns a hypergraph $\mathcal{H} = (X, \varepsilon)$ into a graph $\Pi_2(\mathcal{H}) = (X, E')$ as follows:

$$\{i, j\} \in E' \Leftrightarrow \exists h \in \varepsilon \mid \ i, j \in h.$$

So, for any hypergraph $\mathcal{H}$ we have that $\Pi_2(\mathcal{H}^*) = L(\mathcal{H})$. Furthermore, if $G = (X, E)$ is a graph, with $X = \{1, \dots, N\}$, we can also consider the dual hypergraph $G^* = (E, \varepsilon)$ of G where $\varepsilon = \{h_1, \dots h_n\}$ and $\forall i \in \{1, \dots, n\}$ we consider the corresponding hyperedge $h_i = \{e_j \in E | \ i \in e_j\}$, and also $\Pi_2(G^*) = L(G)$.

Now, if we denote by $I(\mathcal{H})$ the incidence matrix of $\mathcal{H}$, then it is not difficult to verify that

$$I(\mathcal{H})^t \cdot I(\mathcal{H}) = \widetilde{A(\mathcal{H})} = (\widetilde{a_{ij}}) \in \mathbb{R}^{|\varepsilon| \times |\varepsilon|}$$

and

$$I(\mathcal{H}) \cdot I(\mathcal{H})^t = A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N},$$

where

$$\widetilde{a_{ij}} = \begin{cases} |h_i| & \text{if} \quad i = j, \\ |h_i \cap h_j| & \text{if} \quad i \neq j, \end{cases}$$

and

$$a_{ij} = \begin{cases} |\{h \in \varepsilon \mid \ i \in h\}| & \text{if} \quad i = j, \\ |\{h \in \varepsilon \mid \ i, j \in h\}| & \text{if} \quad i \neq j. \end{cases} \tag{2.1}$$

In fact, if we consider in addition the bipartite network $B(\mathcal{H})$ associated to the hypergraph $\mathcal{H} = (X, \varepsilon)$ defined by $B(\mathcal{H}) = (X \cup \varepsilon, E(\mathcal{H}))$ then its adjacency matrix is given by

$$A_{B(\mathcal{H})} = \left(\begin{array}{c|c} 0 & I(\mathcal{H}) \\ \hline I(\mathcal{H})^t & 0 \end{array} \right)$$

and

$$(A_{B(\mathcal{H})})^2 = \left(\begin{array}{c|c} A(\mathcal{H}) & 0 \\ \hline 0 & \widetilde{A(\mathcal{H})} \end{array} \right).$$

The matrix $A(\mathcal{H}) = (a_{ij})$ is known as the frequency matrix of relations between the elements (nodes) of the hypergraph $\mathcal{H}$.

3. Hypergraphs and derivative graph

Quantifying the similarity between two models or structures is one of the most important aspects that has contributed to the development of theories and models in science and technology. There are multiple works whose objective is to model generic data sets in the field of complex networks in order to, by using the constructed model, study the level of similarity or coincidence of such data [50–52]. Thus, since the introduction of Jaccard's index in 1901 [53], through different adaptations and generalizations of this concept [52,54], several types of indexes and generalizations have been established with the aim of quantifying the similarity between two sets or mathematical structures [50–52,54–56].

The basic Jaccard index to compare the degree of coincidence or similarity between two sets A and B can be obtained from the formula

$$\mathcal{J} = \frac{|A \cap B|}{|A \cup B|}.$$

The different applications of the Jaccard index along time made possible the development of new indexes, improving the accuracy of the original results. So, the overlap index and the coincidence similarity [51,52,56,57] are examples of additional indexes that allow to establish similarity between certain types of models and structures, including approaches aimed at quantifying similarity between paragraph contents using the concept of multisets [57].

In our case, we are going to introduce a methodology to analyze and quantify the similarity between two nodes i, j of a hypergraph, applying it to the study of the linguistic network built through the corpus under study.

In this section we are going to introduce the concept of derivative graph of a hypergraph with the idea of associating not only a numerical index that allows us to quantify the heterogeneity and absence of similarity between the corresponding hyperedges, but also to associate a structure (in this case a graph) to characterize the heterogeneity and dissimilarity of the elements of the hypergraph under consideration. Now, we are in a good position to present the concept of derivative graph of a hypergraph over a pair of nodes, bearing in mind this concept is related to some of the ideas partially and briefly sketched in [58]:

Definition 3.1. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, with $A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N}$, we will call the derivative hypergraph of $\mathcal{H}$ with respect to the pair of nodes $i, j \in X$ as the numerical value $\frac{\partial \mathcal{H}}{\partial \{i,j\}}$ obtained by applying the following formula

$$\frac{\partial \mathcal{H}}{\partial \{i, j\}} = \frac{a_i - a_{ij} + a_j - a_{ij}}{a_{ij}} = \frac{a_i - 2a_{ij} + a_j}{a_{ij}}. \tag{3.2}$$

Obviously, if there is not a hyperedge $h \in \varepsilon$ such that $i, j \in h$, we will have $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = \infty$, and if $\forall h \in \varepsilon \ \ (i \in h \Leftrightarrow j \in h)$ then we will have $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = 0$. Note that $\forall i, j \in X$ we have that $\frac{\partial \mathcal{H}}{\partial\{i,j\}} \geq 0$.

It is important to point out that the above definitions can be extended without difficulty to the context of a collection of sets (which would play the role of the hyperedges) and of the elements (respectively the nodes) of the sets of that collection.

If we now consider each hyperedge $h \in \varepsilon$ as a property or a feature that a node may or may not have, or even as an event or affair in which a particular node may or may not participate, so that the entire hypergraph is a set of features or events, the value of $\frac{\partial \mathcal{H}}{\partial\{i,j\}}$ characterizes the (relative) heterogeneity of the properties ε satisfied simultaneously by nodes i and j, or the intensity of participation of the nodes i and j in the set of events ε. Moreover, the smaller the value of the derivative of the network with respect to the set of events over the pair of nodes i, j is, the greater identification and similarity between the corresponding nodes i, j with respect to the considered set of events (in fact, if $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = 0$, these nodes, which participate in exactly the same hyperedges, are so similar that they are, from the point of view of $\mathcal{H}$ indistinguishable). In other words, the higher the value of the derivative is, the greater the degree of unequal participation of the nodes in the hyperedges. Thus, it makes sense to give the following definition:

Definition 3.3. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$ and $i, j \in X$, we will call degree of independence of i and j with respect to $\mathcal{H}$ the numerical value of $\frac{\partial \mathcal{H}}{\partial\{i,j\}}$.

Definition 3.4. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, the derivative graph $\partial\mathcal{H}$ of $\mathcal{H}$ is the weighted graph obtained by considering the derivative of $\mathcal{H}$ with respect all the pairs of nodes $i, j \in X$, and by setting $\forall i, j \in X$ the corresponding numerical value of $\frac{\partial \mathcal{H}}{\partial\{i,j\}}$ on the edge $\{i, j\}$, in such a way that if $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = 0$, then the nodes i and j collapse into a single node (ij), and having in mind that if $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = \infty$, then the edge $\{i, j\}$ does not exist in the derivative graph.

Globally, it can be said that the derivative graph $\partial\mathcal{H}$ gives us a representation of the degree of heterogeneity of participation of nodes on the different hyperedges of $\mathcal{H}$.

Assuming that if k is any positive number then $\frac{k}{0} = +\infty$ and $\frac{k}{\infty} = 0$, for continuity and consistency sake of the established concepts, we are interested in defining the homogeneity matrix and homogeneity graph of a hypergraph:

Definition 3.5. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, we will call homogeneity matrix of $\mathcal{H}$, to the matrix $H(\mathcal{H}) = (h_{ij}) \in \mathbb{R}^{N\times N}$ defined by

$$h_{ij} = \begin{cases} 0 & \text{if} \quad i = j, \\ \dfrac{1}{\frac{\partial \mathcal{H}}{\partial\{i,j\}}} & \text{if} \quad i \neq j. \end{cases}$$

Definition 3.6. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, the homogeneity graph $HG(\mathcal{H})$ of $\mathcal{H}$ is the weighted graph with the same nodes and edges as $\partial\mathcal{H}$, but considering as the weight of each edge the inverse value of the weight corresponding to the derived graph $\partial\mathcal{H}$.

At this point it is remarkable that the application of the PageRank algorithm on the homogeneity graph $HG(\mathcal{H})$ will allow us to extract the most representative nodes of the hypergraph, in the sense that the nodes located in the first places of the ranking obtained will be the "most similar" (in the sense that underlies the definition of homogeneity graph) to each other and to the rest of the nodes of the hypergraph as it will be shown in Section 5.

To clarify the concepts and ideas introduced, let us examine the following example:

Example 3.7. Consider the hypergraph $\mathcal{H} = (X, \varepsilon)$, where $X = \{1, 2, 3, 4, 5\}$, $\varepsilon = \{h_1, h_2, h_3\}$, and $h_1 = \{1, 2, 3, 5\}$, $h_2 = \{2, 4\}$, $h_3 = \{3, 4\}$, represented in panel (a) of Fig. 1. We have that

$$I(\mathcal{H})^t = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix},$$

$$I(\mathcal{H}) \cdot I(\mathcal{H})^t = A(\mathcal{H}) = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 0 & 1 & 1 & 2 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

The values of the derivatives of $\mathcal{H}$ with respect to all the pair of nodes of G are, respectively:

$$\frac{\partial \mathcal{H}}{\partial\{1,5\}} = 0, \ \frac{\partial \mathcal{H}}{\partial\{1,2\}} = 1, \ \frac{\partial \mathcal{H}}{\partial\{1,3\}} = 1, \ \frac{\partial \mathcal{H}}{\partial\{1,4\}} = +\infty, \ \frac{\partial \mathcal{H}}{\partial\{2,3\}} = 2,$$
$$\frac{\partial \mathcal{H}}{\partial\{2,4\}} = 2, \ \frac{\partial \mathcal{H}}{\partial\{2,5\}} = 1, \ \frac{\partial \mathcal{H}}{\partial\{3,4\}} = 2, \ \frac{\partial \mathcal{H}}{\partial\{3,5\}} = 1, \ \frac{\partial \mathcal{H}}{\partial\{4,5\}} = +\infty.$$

so that the derivative graph $\partial\mathcal{H}$ is the one represented in part (b) of Fig. 1 and the homogeneity matrix of $\mathcal{H}$ is:

$$H(\mathcal{H}) = \begin{pmatrix} 0 & 1 & 1 & 0 & \infty \\ 1 & 0 & 1/2 & 1/2 & 1 \\ 1 & 1/2 & 0 & 1/2 & 1 \\ 0 & 1/2 & 1/2 & 0 & 0 \\ \infty & 1 & 1 & 0 & 0 \end{pmatrix}.$$

Note that the edge $\{1, 4\} \in E$ has been removed in the derivative network $\partial\mathcal{H}$ and that nodes 1 and 5 have collapsed into a single node in the obtained network. So, the adjacency matrix of the homogeneity graph $HG(\mathcal{H})$ is:

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1/2 & 1/2 \\ 1 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix},$$

where the set of nodes of $HG(\mathcal{H})$ is $(\{(1, 5), 2, 3, 4\})$ ordered as they appear (panel (c) of Fig. 1).

Thus, in panel (a) of Fig. 1 it can be observed the original hypergraph $\mathcal{H}$, in part (b) its derivative graph $\partial\mathcal{H}$ and in panel (c) its corresponding homogeneity graph $HG(\mathcal{H})$.

It is worth noting that, in a similar way as it has been done in Definition 3.1, it is possible to establish the derivative of a hypergraph with respect to a set of three or more nodes as follows:

$$\frac{\partial \mathcal{H}}{\partial\{i, j, k\}} = \frac{1}{a_{ijk}} \cdot \left(\sum_{r\in\{i,j,k\}} a_r - 2 \sum_{\substack{r,s\in\{i,j,k\} \\ r\neq s}} a_{rs} + 3a_{ijk} \right),$$

where $a_{ijk} = |\{h \in \varepsilon | \, i, j, k \in h\}|$, and the same type of formula can be obtained for sets of nodes of higher cardinality.

Note that the same idea can be extended to the definition of degree of independence of several nodes as follows: Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, and $i_1, \ldots, i_n \in X$, the degree of independence of $i_1, \ldots, i_n$ in $\mathcal{H}$ is the numerical value $\frac{\partial \mathcal{H}}{\partial\{i_1,\ldots,i_n\}}$.

Finally, it is remarkable that the use of the PageRank algorithm on the homogeneity graph will allow us to extract a ranking of the most representative individuals (or nodes) of either the hypergraph or the network under consideration.

To conclude this section, it must be noted that when both graphs and hypergraphs are used simultaneously to model certain complex systems, it is sometimes very useful to analyze how these structures interact and overlap using the tools introduced in this section. In this regard, it should be noted that the tools introduced in this section can be used to capture intrinsic and mesoscopic characteristics of a graph and to define new invariants of graphs and isomorphic networks. For

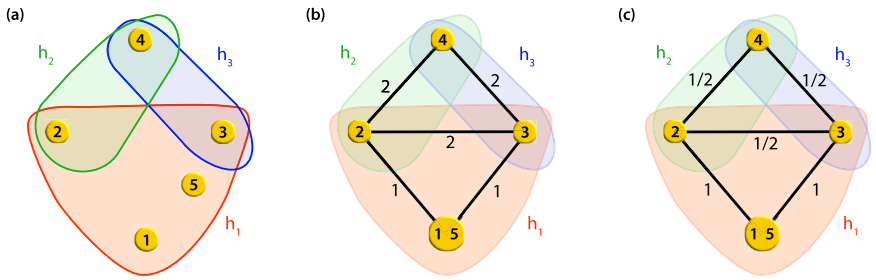

Fig. 1. Hypergraph $\mathcal{H}$ (panel (a)), its derivative graph $\partial\mathcal{H}$ (panel (b)) and its homogeneity graph $HG(\mathcal{H})$ (panel (c)).

example, given a graph $G = (X, E)$, we can consider the hypergraph $\mathcal{H} = (X, \varepsilon)$ such that each of its hyperedges is formed by all the nodes that are part of a cycle, or by all the nodes that are part of a spanning tree of G. The most accurate framework to work with the overlapping of these structures is the use of hyperstructures.

In [59] we can find a first definition of the concept of hyperstructure as follows:

Definition 3.8 (*[59]*)**.** Given a graph $G = (X, E)$ with N vertices and m edges and a hypergraph $\mathcal{H} = (X, \varepsilon)$, a *hyperstructure* $S = (X, E, \mathcal{H})$ is a triple formed by the vertex set X, the edge set E and the hyperedge set $\mathcal{H}$. The hyperstructure S is said to be compatible if for every edge $e = \{v, w\} \in E$ there exists a hyperedge $h \in \varepsilon$ such that $v, w \in h$.

It is not difficult to prove the following result:

Theorem 3.9. *Let $S = (X, E, \mathcal{H})$ be a hyperstructure, $L(G) = (E, E')$ the linegraph of $G = (X, E)$ and $\Pi_2(\mathcal{H}) = (X, E'')$. If S is compatible, then $S' = (E, E', \mathcal{H})$ and $S'' = (X, E'', \mathcal{H})$ are also hyperstructures.*

It is important to highlight that by using the idea of derivative we have introduced in this paper we can examine and determine the uniformity of participation of two, three or more nodes in the considered structure or hyperstructure, or even the binary relationships (edges) between participants of a certain event by simply considering a suitable hyperstructure in which the nodes be the edges of the original graph under consideration.

Now, we can define the derivative graph of a weighted hyperstructure:

Definition 3.10. Given a hyperstructure $S = (X, E, \mathcal{H})$, where $G = (X, E, W)$ is a weighted graph and $\mathcal{H} = (X, \varepsilon)$, if w_{ij} denotes the weight of the edge $e = \{i, j\} \in E$, then we will call the derivative of e with respect to the hyperstructure S the numerical value obtained by applying the following formula

$$\frac{\partial e}{\partial S} = w_{ij} \cdot \left(\frac{a_i - 2a_{ij} + a_j}{a_{ij}} \right).$$

Obviously, if there is not a hyperedge $h \in \mathcal{H}$ such that $e = \{i, j\} \in h$, we will have $\frac{\partial e}{\partial S} = \infty$. On the other hand, it is evident that if a hyperstructure is compatible, the derivative of any edge with respect to S cannot be equal to $+\infty$.

Definition 3.11. Given a hyperstructure $S = (X, E, \mathcal{H})$, where $G = (X, E, W)$ is a weighted graph and $\mathcal{H} = (X, \varepsilon)$, if w_{ij} denotes the weight of the edge $e = \{i, j\}$, then the derivative graph of G with respect to S is the weighted graph $\frac{\partial G}{\partial S}$ obtained by setting $\forall e \in E$ the corresponding numerical value of $\frac{\partial e}{\partial S}$ on the edge $e = \{i, j\}$, in such a way that if $\frac{\partial e}{\partial S} = 0$, then the nodes i and j collapse into a single node (ij).

As a direct application of the definition, note that if we consider the graphs $G = (X, E)$ (panel (a) of Fig. 2) and $G' = (X, E')$ (panel (b) of Fig. 2) and the hyperstructures $S = (X, E, \mathcal{H})$ and $S = (X, E', \mathcal{H}')$ such that each of their hyperedges is composed by all the nodes belonging to a cycle formed by three or more nodes of G and G' respectively, then the derived graphs $\frac{\partial G}{\partial S}$ and $\frac{\partial G'}{\partial S'}$ are completely different since, for example,

$$\frac{\partial\{1,2\}}{\partial S'} = \frac{20}{8} = \frac{5}{2}.$$

On the other hand, as can be seen,

$$\frac{\partial\{1,2\}}{\partial S} = \frac{62}{20} = \frac{31}{10},$$

and, obviously,

$$\frac{5}{2} \neq \frac{31}{10}.$$

Note that Definition 3.11 allows us to iterate the derivatives with respect to a hyperstructure, because if the graph derived from the hyperstructure is $\frac{\partial G}{\partial S} = (X', E', W')$ y $S' = (X', E', \mathcal{K})$, then we can consider the mixed derivatives of a graph G with respect to two different hyperstructures (which may eventually be the same) S and S' (in this order) as

$$\frac{\partial^2 G}{\partial S' \partial S} = \frac{\partial}{\partial S'} \left(\frac{\partial G}{\partial S} \right).$$

It is obvious that the successive derivative graphs obtained by deriving respect a suitable chain of two or more hyperstructures allow to obtain characteristics and properties of the system or model under study related to the absence of similarity between the nodes.

4. A linguistic hyperstructure based on the lexical layer within a multilayer linguistic network model

We are now ready to show the potential applications of the defined mathematical structures and tools to the linguistic analysis of texts, looking for the identification of signs and specific features of a style or competence level of language considering the most significant words and their relationships. It can be said that the English language has four major word grammar categories: nouns, adjectives, verbs, and adverbs. Other word classes are prepositions, conjunctions, determiners, interjections or pronouns [60]. On this basis described in [29,30] we have built a methodology close to supervised machine learning consisting of dividing the words of the corpus under study into a multilayer network [7] composed by four layers: lexical layer, verb layer, linking layer and remaining words layer.

In order to discriminate between the terms (words) and to assign them to one or another layer, a completely lexical linguistic decision was made according to the criteria of several experts. Thus, the terms (words) of the corpus have been distributed in the different layers according to their morphological and lexical properties. Some other linguistic aspects, such as the specific terminology of a specialty language and the different combinations of words referring to new meanings (called "linguistic collocations") have also been successfully studied and modeled in [29,30].

In the model established in [29] interlayer relations are the basic grammatical relations in a sentence, for example, the interaction

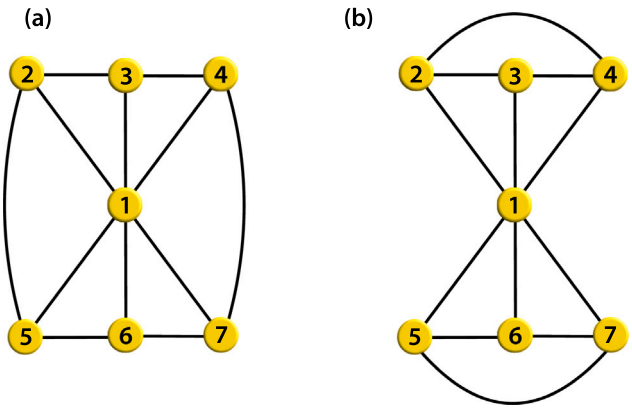

Fig. 2. Two graphs with the same degree distribution: Graph G (panel (a)) and graph G' (panel (b)).

between layers that facilitates the formation and description of specialty verbs (e.g. "cluster together"). On the other hand, throughout the present work, we will consider the sentences as the unit under study, identifying each sentence in the corpus (set of words located between two periods) with the subset of lexical words appearing in that sentence.

For this reason, throughout this work we are going to focus on the words (nodes) located in the lexical layer. At this point, it is remarkable that in the lexical layer many words can act as verbs when we analyze texts written by authors with higher language skills. For example, within the sentence "model a network", the word "model" is a verb, but in the expression "network model" the term "model" is a noun.

In order to set our approach, the model of the corpus analyzed is considered as a set of texts formed by sentences (set of lexical words between two periods). In fact, from a practical and computational point of view, each sentence is identified with the set of lexical words that compose it. This way, let us consider the hyperstructure in which the nodes are the lexical words, the edges between these nodes are established when these words appear in the same sentence, and the set of hyperedges is the set of sentences that constitute the corpus.

It is important to point out that the linguistic hyperstructure considered is a compatible hyperstructure, since the edges are established between words that appear in the same sentence. Therefore, from Theorem 3.9 it is possible to study both the hyperstructure in which the nodes are the words and the hyperedges are the sentences and, in a complementary way, the hyperstructure in which the nodes are the edges between words (dual graph of the original graph) and the hyperedges are also the sentences.

On the other hand, by considering paragraphs as a set of sentences, and the extended abstracts of our corpus as a set of paragraphs, we can add to this model new linguistic hyperstructures that undoubtedly allow us to characterize a text or set of texts from the derivatives of the corresponding graphs and hypergraphs respectively.

In order to illustrate how useful are the tools presented in the context of the linguistic analysis of texts, let us consider a text in which the same sentence is repeated over and over again. In that case, by deriving the linguistic hypergraph formed by the set of all the repeated sentences with respect to the lexical words of the sentence repeated over and over in all those sentences, the derivative graph will collapse to a single node.

So, by calculating the derivative graph from the linguistic hypergraph composed by all the sentences of a corpus or a text, we will obtain the degree of similarity between the sentences of that text, and also the greater or lesser degree of difference between all the sentences forming such text (or corpus), with the peculiarity that these quantitative measures are represented in the corresponding derivative graph.

Consequently, the derivative graph of a text or a set of texts is a quantitative and qualitative structure of such text that is a specific feature of that text (or set of texts) for real, which may be considered, in certain cases, like a signature or specific characteristic of the style of an author.

When analyzing the hypergraph $\mathcal{H}$ formed by all the sentences of the corpus under study, we obtained 127 pairs of words that appear in exactly the same sentences. Thus, for example

$$\frac{\partial \mathcal{H}}{\partial\{monte, carlo\}} = \frac{\partial \mathcal{H}}{\partial\{differential, rungekutta\}} = \frac{\partial \mathcal{H}}{\partial\{oscillatory, asynchronous\}} = 0.$$

It is important to note at this point that, if three or more words in the corpus analyzed appear in exactly the same sentences, these words have collapsed into a single node. This has happened in 13 cases. Finally, and by way of illustrative example, we will point out that

$$\frac{\partial \mathcal{H}}{\partial\{network, system\}} = 16.33,$$
$$\frac{\partial \mathcal{H}}{\partial\{language, formal\}} = \frac{\partial \mathcal{H}}{\partial\{connectance, asymmetry\}} = 2,$$
$$\frac{\partial \mathcal{H}}{\partial\{process, graph\}} = 28.14,$$
$$\frac{\partial \mathcal{H}}{\partial\{placing, model\}} = +\infty.$$

Fig. 3 shows the homogeneity graph corresponding to the corpus considered, in which the thickness of each edge is proportional to its weight. On the other hand, as it can be seen in the right part of Fig. 3, there is no link between "features" and "properties" because $\frac{\partial \mathcal{H}}{\partial\{feature, properties\}} = +\infty$, and the edge joining "networks" and "complex" is thicker than the rest.

Also, as it can be seen in the histogram of Fig. 4, there are more than 10^3 pairs of words $\{i, j\}$ such that $0 \leq \frac{\partial \mathcal{H}}{\partial\{i,j\}} \leq 10$ and more than 10^6 pairs of words $\{i, j\}$ whose derivative is $+\infty$ (note that in Fig. 4, the length of the intervals of the horizontal axis is 10).

To conclude this section, we would like to point out that the automatic extraction of the linguistic level of a corpus, the search for lexical patterns in sentences of a given author or writer of a particular specialty language, the search for similarities and differences in a set of texts and the automatic classification of texts according to these differences or similarities are some of the possible applications of the methodology underlying this model.

5. On lexical density and three different rankings of sentences: Computational results

As far as it is known, the personalized PageRank of a individual term (node) i is the i-component of the stationary state $\pi_0 \in \mathbb{R}^n$ ($\|\pi_0\| = 1$) of the random walker with transition matrix [61–64]

$$P = \alpha P_B^T + (1 - \alpha)\mathbf{v}\mathbf{e}^T,$$

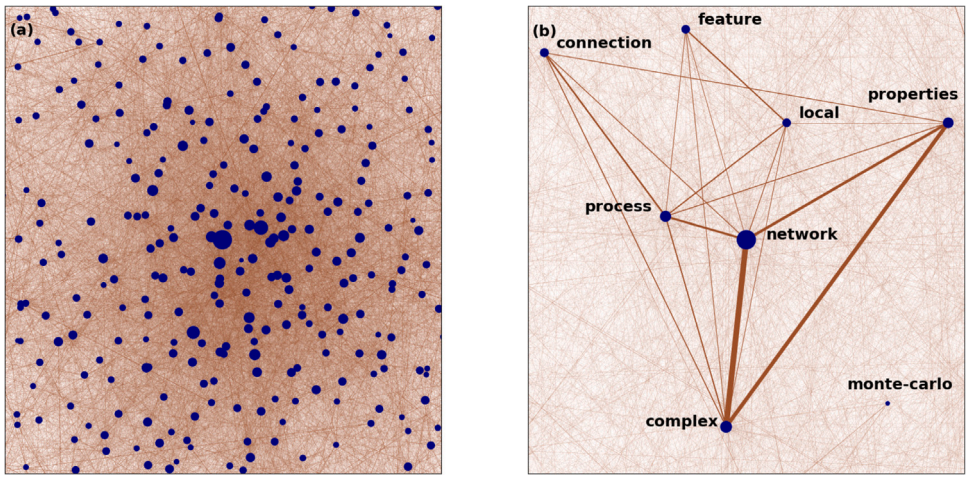

Fig. 3. Homogeneity graph $HG(\mathcal{H})$ of the corpus $\mathcal{H}$ under analysis. On the right side a zoom of the subgraph of neighbors of the word "network" is shown.

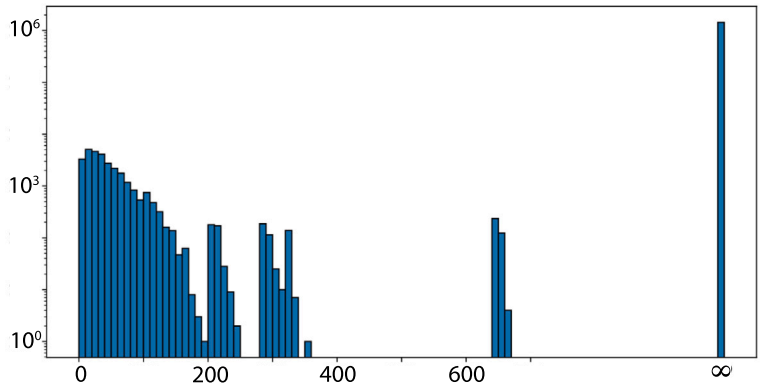

Fig. 4. Histogram clustering the number of pairs of words $\{i, j\}$ by the value of $\frac{\partial \mathcal{H}}{\partial\{i,j\}}$.

where $\alpha \in (0, 1)$, $B = (b_{ij})$ is the adjacency matrix of the network under consideration, $\mathbf{e}^T = (1, \ldots, 1)$, $\mathbf{v} \in \mathbb{R}^n$ ($\|\mathbf{v}\| = 1$) is the personalization vector and

$$P_B = (p_{ij}) = \left(\frac{b_{ij}}{\sum_k b_{ik}}\right).$$

To carry out our study on the hypergraph $\mathcal{H}$ in which the vertices are the lexical words of the corpus, and the hyperedges are the phrases (sets of lexical words of the corpus located between two periods), we will use the same methodology as in [29,30] to associate its corresponding PageRank to each node, with the idea of ranking the lexical words according to their importance [61–63,65,66]. For this purpose, taking into account that for the PageRank calculation used throughout this work we have considered the algorithm described in [67], we will apply this algorithm on three different structures obtained from the application of three different criteria:

1. **Ranking 1**. To calculate this ranking, we first have built a graph on which to apply the PageRank algorithm. In order to do that, we convert each hyperedge of $\mathcal{H}$ into a clique to obtain the projection graph $\Pi_2(\mathcal{H})$. After this, taking into account that the average number of words of a sentence within the corpus under study is 5.809 and that, therefore, the local lexical density is 5.809, we can deduce that the damping factor corresponding to this configuration is 0.853, since

$$\begin{aligned} 5.809 =& \mathbb{E}(\ell) = \sum_{k=0}^{\infty} k \cdot \mathbb{P}(\ell = k) = \sum_{k=1}^{\infty} k \cdot (1-q) \cdot q^k \\ =& (1-q) \cdot q \sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{q}{1-q}. \end{aligned}$$

2. **Ranking 2.** To calculate this ranking, we will apply the PageRank algorithm considered on the network $\Pi_2(\mathcal{H}^*) = L(\mathcal{H})$ so that, once the numerical value attributed to each phrase has been obtained, this value is distributed proportionally among the words that make up that sentence. It is important to note that, in this case, the network considered is a directed network, and that, if $s_1, s_2 \in L(\mathcal{H})$, these sentences will be connected if they have at least one lexical word in common, so that the edge weight $w(s_1 \to s_2)$ is the number of words shared by both sentences multiplied by the number of times that sentence s_2 appears repeated in the corpus. Obviously, the edge weight $w(s_1 \to s_2)$ may be different from $w(s_2 \to s_1)$. Now, using the same reasoning as in the previous case, and having in mind that the average number of sentences of a paper included in the corpus under study is 27.12, in this context, the damping factor corresponding to this configuration is 0.96.
3. **Ranking 3.** To calculate this ranking, we will apply the PageRank algorithm considered on the weighted graph $HG(\mathcal{H})$. Taking into account that the average number of words of a sentence is 5.756 (since, after collapsing words pairs $\{i, j\}$ such that $\frac{\partial \mathcal{H}}{\partial\{i,j\}} = 0$, the average length of sentences decreases, albeit slightly), the damping factor corresponding to this configuration is 0.852. Fig. 3 shows the homogeneity graph corresponding to the corpus considered. The size of the nodes is proportional to the component of the PageRank vector corresponding to that node, and the thickness of each edge is proportional to its weight.

In all of the described cases, the corresponding value of q is the probability that a random walker will not vary its trajectory by moving

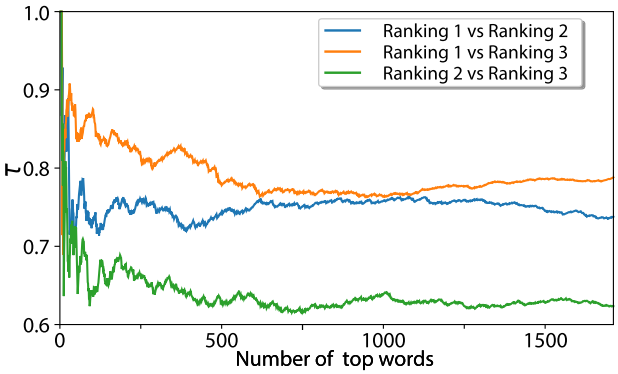

Fig. 5. Kendall's tau coefficient variation by comparing the rankings in pairs among them, depending on the number of top lexical words considered in each ranking.

Table 1
Rankings of lexical words.

	Ranking 1	Ranking 2	Ranking 3
1st	network	network	network
2nd	system	system	system
3rd	model	model	model
4th	complex	complex	complex
5th	process	number	graph
6th	number	process	process
7th	information	structure	structure
8th	graph	new	information
9th	new	information	number
10th	structure	distribution	new
11th	properties	properties	properties
12th	distribution	graph	distribution
13th	study	study	dynamics
14th	dynamics	dynamics	study
15th	case	interaction	analysis

to a node directly connected by an edge to the current node instead of jumping to another node in this network not necessarily connected to the previous one. In our situation, this jump can be understood as the end of the current sentence and the starting point of a new sentence for Ranking 1 and Ranking 3, and as the end of the current paper and the starting point of a new paper for Ranking 2. To complete the necessary elements to apply the algorithm, we will point out that for Ranking 1 and Ranking 3 the personalization vector considered is the (relative) frequency of lexical words, and for Ranking 2 the personalization vector considered is the (relative) frequency of each sentence included in the corpus under study.

As it can be seen in Table 1, there is hardly any difference at the top of the three rankings. As expected, Ranking 3 gives us the most representative words of the corpus in the sense that they are the words at the heart of the corpus linking the largest number of sentences together. In any case, the three rankings should not be very different from each other, as it is actually the case (since the first four positions are occupied by the same words in all three cases) and, as it happens in the case under study, Ranking 1 and Ranking 3 are more similar to each other than to Ranking 2. However, as the number of words considered at the top of each ranking increases, the differences between the three rankings become much more evident, as it can be seen in Fig. 5, where we plot the differences between these rankings by visualizing the variation of the Kendall's tau coefficient (τ) [68] regarding the number of lexical words considered in the three rankings.

6. Seeking for distinctive characteristics that allow distinguishing the styles of different authors and language levels

By considering several types and models of hypergraphs and hyperstructures for a given text or corpus, we can associate to that written text or corpus various features that allow us to identify it as if it were some sort of mathematical signature associated with them. For example, for a given text it is possible to consider a hypergraph in which the nodes are the words and the hyperedges are the sentences, another in which the nodes are the words and the hyperedges are the paragraphs, another in which the nodes are the sentences and the hyperedges are the paragraphs, just to mention some of the possibilities. This succession of mathematical structures and the different parameters (such as diameter, degree distribution, centrality, efficiency, among others, that characterize them) are, without a doubt, elements that configure and allow us to characterize and compare different texts, making it clear the characteristics that constitute their seal of identity in terms of style.

7. Conclusions

We introduce and study the derivative of a hypergraph and the homogeneity graph of a hypergraph as two new and useful structures that can be applied to study the degree of independence of the nodes of a hypergraph. Also, these structures may be employed to obtain a ranking of the most representative nodes of the hypergraph in the sense that the lexical words represented by these nodes link the most significant ideas and concepts of the text without necessarily being those terms usually considered as keywords.

These concepts allow us to associate not only a numerical index useful for quantifying the heterogeneity and lack of similarity between the nodes of the hypergraph, but also a graph aiming to characterize the heterogeneity and dissimilarity of the different elements of the considered hypergraph.

Moreover, these concepts also let us obtain technical characteristics related to the styles of different authors and the language competence level of any text written in English. Also, a possible application to text classification, text summarization, automated translation, stylometry and authorship detection is found.

Undoubtedly, the tools derived from the linguistic analysis obtained by using this new tool will provide us with new models and better instruments to typify and locate the characteristics of the style of different authors together with the style and intrinsic linguistic characteristics found in specialized texts in terms of collocations, word sense disambiguation and syntagmatic structures.

Finally, it is important to mention that the construction of tools to find lexical patterns of the style of an author or a text belonging to a specialty language, the automatic classification of texts according to their style and the automatic labeling and identification/verification of lexical patterns are some possible additional applications of these new tools.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

Authors would like to thanks Karin Alfaro-Bittner for some inspiring discussions. This work has been partially supported by projects PGC2018-101625-B-I00 (Spanish Ministry, AEI/FEDER, UE) and M1993 Grant (Rey Juan Carlos University, Spain). Authors acknowledge the usage of the resources, technical expertise and assistance provided by the supercomputer facility CRESCO of ENEA in Portici (Italy).

References

[1] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U. Complex networks: Structure and dynamics. Phys Rep 2006;424:75–308.
[2] Criado R, Flores J, García del Amo A, Gómez-Gardeñes J, Romance M. A mathematical model for networks with structures in the mesoscale. Int J Comput Math 2012;89(3):291–309.
[3] Estrada E. Networks science. New York: Springer; 2010.
[4] Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. J Complex Netw 2014;2(3):203–71.
[5] Newman M. Networks: An introduction. Oxford University Press; 2010.
[6] Wasserman S, Faust K. Social network analysis. Cambridge: Cambridge University Press; 1994.
[7] Boccaletti S, Bianconi G, Criado R, Del Genio CI, Gómez-Gardeñes J, Romance M, et al. The structure and dynamics of multilayer networks. Phys Rep 2014;544(1):1–122.
[8] Chapela V, Criado R, Moral S, Romance M. Intentional risk management through complex networks analysis. Heidelberg New York Dordrecht London: Springer International Publishing; 2015.
[9] L.d.F. Costa, Oliveira ON, Travieso G, Rodrigues FA, Villas Boas PR, Antiqueira L, et al. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. Adv Phys 2011;60(3):329–412.
[10] de Arruda HF, Nascimento S, Marinho VQ, Amancio DR, Costa LdF. Representation of texts as complex networks: A mesoscopic approach. J Complex Netw 2018;6(1):125–44.
[11] Dogorovtsev SN, Mendes JFF. Language as an evolving word web. Proc R Soc Lond B 2001;268:2603–6.
[12] Ferrer i Cancho R, Solé RV. The small world of human language. Proc R Soc Lond B 2001;286:2261–6.
[13] Latora V, Nicosia V, Russo G. Complex networks: Principles, methods and applications. Cambridge University Press; 2017.
[14] Masucci A, Rodgers G. Network properties of written human language. Physica A 2006;457:117–28.
[15] Albert R, Barabasi AL. Statistical mechanics of complex networks. Rev Modern Phys 2002;74:47–97.
[16] Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, et al. Networks beyond pairwise interactions: Structure and dynamics. Phys Rep 2020;87492.
[17] de Arruda HF, Costa LdF, Amancio DR. Using complex networks for text classification: Discriminating informative and imaginative documents. Europhys Lett 2016;113(2):28007.
[18] Kalimeri M, Constantoudis V, Papadimitriou C, Karamanos K, Diakonos FK, Papageorgiou H. Wordlength entropies and correlations of natural language written texts. J Quant Linguist 2015;22(2):101–18.
[19] Altmann EG, Pierrehumbert JB, Motter AE. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. PLoS One 2009;4(11):e7678.
[20] Borge-Holthoefer J, Arenas A. Semantic networks: Structure and dynamics. Entropy 2010;12:1264–302, X.
[21] Cong J, Liu H. Approaching human language with complex networks. Phys Life Rev 2014;11(4).
[22] Ferrer i Cancho R, Riordan O, Bollobás B. The consequences of Zipf's law for syntax and symbolic reference. Proc Biol Sci/ R Soc 2005;272(1562):561–5.
[23] Liu H, Hu F. What role does syntax play in a language network? Europhys Lett 2008;83:18002.
[24] Liu H, Xu C, Liang J. Dependency distance: A new perspective on syntactic patterns in natural languages. Phys Life Rev 2017;21:171–93.
[25] Martincic S, Margan D, Mestrovic A. Multilayer network of language: A unified framework for structural analysis of linguistic subsystems. Phys Rev E 2016;74:026102.
[26] Mehler A, Lűcking A, Banisch S, Blanchard P, Frank-Job B, editors. Towards a theoretical framework for analyzing complex linguistics networks. Springer-Verlag; 2016.
[27] Solé RV, Corominas-Murtra B, Valverde S, Steels L. Language networks: Their structure, function, and evolution. Complexity 2010;15(6):20–6.
[28] Ferrer i Cancho R, Solé RV, Köhler R. Patterns in syntactic dependency networks. Phys Rev E 2004;69:051915.
[29] Criado-Alonso A, Battaner-Moro E, Aleja D, Romance M, Criado R. Using complex networks to identify patterns in specialty mathematical language: A new approach. Soc Netw Anal Min 2020;10(1):1–10.
[30] Criado-Alonso A, Battaner-Moro E, Aleja D, Romance M, Criado R. Enriched line graph: A new structure for searching language collocations. Chaos Solitons Fractals 2021;142:110509.
[31] McEnery T, Hardie A. Corpus linguistics: Method, theory and practice (Cambridge textbooks in linguistics). Cambridge: Cambridge University Press; 2011.
[32] Halliday MAK, Matthiessen CMIM. Introduction to functional grammar. 3rd ed.. London and New York: Routledge, Taylor & Francis Group; 2004.
[33] Sinclair J. Corpus, concordance, collocation. Describing English language. Oxford University Press; 1991.
[34] Benson A. Three hypergraph eigenvector centralities. SIAM J Math Data Sci 2019;1(2):293–312.
[35] Lambiotte R, Rosvall M, Scholtes I. From networks to optimal higher-order models of complex systems. Nat Phys 2019;15:313–20.
[36] Torres L, Blevins AS, Bassett D, Eliassi-Rad T. The why, how, and when of representations for complex systems. SIAM Rev 2021;63:435.
[37] Liu H, Cong J. Empirical characterization of modern Chinese as a multi-level system from the complex network approach. J Chin Linguist 2014;42:1–38.
[38] Solé R. Syntax for free? Nature 2005;434:289.
[39] Zipf GL. Human behavior and the principle of least effort. Hafner; 1965.
[40] Berge C. Hypergraphs. Combinatorics of finite sets. North-Holland; 1989.
[41] Whitney H. Congruent graphs and the connectivity of graphs. Amer J Math 1932;54(1):150–68.
[42] Bermond JC, Heydemann MC, Sotteau D. Line graphs of hypergraphs. I. Discrete Math 1977;18(3):235–41.
[43] Tyshkevich R, Zverovich VE. Line hypergraphs: A survey. Acta Appl Math 1998;52:209–22.
[44] Bagga J. Old and new generalizations of line graphs. IJMMS 2004;29:1509–21.
[45] Criado R, Flores J, García del Amo A, Romance M. Centralities of a network and its line graph: An analytical comparison by means of their irregularity. Int J Comput Math 2014;91(2):304–14.
[46] Criado R, Flores J, García del Amo A, Romance M, Barrena E, Mesa JA. Line graphs for a multiplex network. Chaos 2016;26(6):065309.
[47] Evans TS, Lambiotte R. Line graphs, link partitions, and overlapping communities. Phys Rev E 2009;80:016105.
[48] Evans TS, Lambiotte R. Line graphs of weighted networks for overlapping communities. Eur Phys J B 2010;77:265–72.
[49] Naik RJ. Intersection graphs of graphs and hypergraphs: A survey. 2018, http://dx.doi.org/10.48550/arXiv.1809.08472, arXiv:1809.08472.
[50] Brusco M, Cradit JD, Steinley D. A comparison of 71 binary similarity coefficients: The effect of base rates. PLoS One 2021;16(4):e0247751.
[51] Costa LdF. Coincidence complex networks. J Phys: Complex 2022;(3):015012.
[52] Vijaymeena MK, Kavitha K. A survey on similarity measures in text mining. Mach Learn Appl 2016;3(1):1–28.
[53] Jaccard P. Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines. Bull de la Soc Vaudoise Des Sci Nat 1901;37:241–72.
[54] L.d.F. Costa. Further generalizations of the jaccard index. 2021, https://www.researchgate.net/publication/355381945. [Accessed 21 August 2021].
[55] Hamers L, Hemeryck Y, Herweyers G, Janssen M, Ketters H, Rousseau H, Vanhoutte A. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Inf Process Manage 1989;25(3):315–8.
[56] Costa LdF. On similarity. Physica A 2022;599:127456.
[57] Costa LdF. On the effects of text preprocessing on paragraph similarity networks. 2022, https://www.researchgate.net/publication/361553289. [Accessed 20 June 2022].
[58] Gorbatóv VA. Fundamentos de la matemática discreta (in Spanish). Moscow: Mir; 1988.
[59] Criado R, Romance M, Vela-Pérez M. Hyperstructures, a new approach to complex systems. IJBC 2010;20(3):877–83.
[60] Huddleston RD. The Cambridge grammar of the English language. Cambridge, UK, New York: Cambridge University Press; 2002.
[61] Boldi P, Santini M, Vigna S. Pagerank: Functional dependencies. ACM Trans Inf Syst 2009;27(4):19–23.
[62] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Comput Netw 1998;30:107.
[63] Brin S, Page L, Motwani R, Winograd T. The pagerank citation ranking: Bringing order to the web. Technical report, Standford InfoLab; 1999.
[64] García E, Pedroche F, Romance M. On the localization of the personalized PageRank of complex networks. Linear Algebra Appl 2013;439:640–52.
[65] Langville AN, Meyer CD. Google's pagerank and beyond: The science of search engine ranks. Princeton Univ Press; 2006.
[66] Pedroche F, Romance M, Criado R. A biplex approach to PageRank centrality: From classic to multiplex networks. Chaos 2016;26(6):065301.
[67] Aleja D, Criado R, García del Amo A, Pérez A, Romance M. Non-backtracking PageRank: From the classic model to Hashimoto matrices. Chaos Solitons Fractals 2019;126:283–918.
[68] Kendall M. A new measure of rank correlation. Biometrika 1938;30(1–2):81–93.