





RESEARCH ARTICLE

WILEY

A new insight into linguistic pattern analysis based on multilayer hypergraphs for the automatic extraction of text summaries

Ángeles Criado-Alonso¹  | David Aleja^{2,4}  | Miguel Romance^{2,3,4}  |
Regino Criado^{2,3,4} 

¹Departamento de Filología Extranjera, Traducción e Interpretación, Universidad Rey Juan Carlos, Móstoles (Madrid), Spain

²Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, ESCET Universidad Rey Juan Carlos, Móstoles (Madrid), Spain

³Center for Computational Simulation, Pozuelo de Alarcón (Madrid), Spain

⁴Data, Networks and Cybersecurity Research Institute, Universidad Rey Juan Carlos, Madrid, Spain

Correspondence

Regino Criado, Departamento de Matemática Aplicada, Ciencia e Ingeniería de los Materiales y Tecnología Electrónica, Universidad Rey Juan Carlos, Tulipán s/n, 28933 Móstoles (Madrid), Spain.
Email: regino.criado@urjc.es

Communicated by: J. C. Cortés

Funding information

Rey Juan Carlos University, Spain, Grant/Award Number: M1967; Spanish Ministry, AEI/FEDER, UE, Grant/Award Number: PGC2018-101625-B-100

Forensic linguistics and stylometry have in the exploration of linguistic patterns one of their fundamental tools. Mathematical structures such as complex multilayer networks and hypergraphs provide remarkable resources to represent and analyze texts. In this paper, we present a model that includes some specific mesoscopic relations between the different types of words in a corpus (lexical words, verbs, linking words, other words) according to the sentences or paragraphs in which they appear. This model is supported by various mathematical structures such as partial multiline graphs, multilayer hypergraphs, and their derivative graphs. The methodology proposed from this new point of view is of singular help to find meaningful sentences from any text to set up an automatic summary of the text and, eventually, to determine its linguistic level.

KEYWORDS

derivative of a hypergraph, hypergraph, linguistic patterns, multilayer hypergraph, PageRank, partial multiline graph, stylometry

MSC CLASSIFICATION

05C50, 05C81, 05C82, 05C90, 68R10

1 | INTRODUCTION

The science of complex networks is based on the hypothesis that the behavior of many complex systems can be explained by studying the structural and functional relationships between their components through a network representation [1–6]. In the last two decades, new models of interconnected networks have been defined responding to the fact that complex systems include multiple subsystems subject to be organized in layers of connectivity. Thus, interconnected networks have emerged in recent years as a general framework for dealing with hyperconnected systems, either because there exists dependency relationships between objects or systems, or because of the existence of different channels of interaction between them [7–12]. In this respect, the framework provided by multilayer networks and multiplex networks has

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Mathematical Methods in the Applied Sciences* published by John Wiley & Sons, Ltd.

provided a specific formalism whose objective is the analysis of the different ways in which a set of objects or systems can interact. Thus, as such objects or systems are present at the same time in different interaction networks (layers), these layers are interconnected. Moreover, the existence of interactions of different nature and simultaneous interactions between nodes and edges (group collaborations, chemical reactions in which more than two components interact, etc.) have shown that hypergraphs and multilayer networks are very suitable structures for this type of studies [9, 13–16]. Moreover, the latest advances in modern linguistics are based on the treatment of a language as a system or a complex network, to which mathematical tools, statistical measures, and procedures of this branch of science can be applied to obtain a new, efficient, and effective approach to the study of language [3, 17–28]. Additionally, the search for linguistic patterns, stylometry, and forensic linguistics have in the theory of complex networks, their structures, and associated mathematical tools, allies with which to model and analyze texts [18–20, 29, 30]. Some of these analyses have considered semantic similarities [31], automatic summarization [32, 33], stylometry [34], authorship detection [35–37], and even sentiment analysis [38]. Other linguistic aspects and elements such as the specific terminology of a specialty language and the different combinations of words that stand for specific meanings or concepts (called “collocations”) have also been successfully modeled using an appropriate methodology within the scope of this model [18]. A classical and well-known approach for generating complex networks from texts is the word adjacency (or co-occurrence) technique, which is based on connecting pairs of words that are immediately adjacent [22, 26]. However, the networks obtained from co-occurrence do not capture the information of the mesoscopic structure of the text, related to sentences, paragraphs, and chapters. Our idea, to cope with these limitations, is to look for a mesoscopic representation supported by several mathematical structures related to the mathematical structure of hypergraph to analyze the relationships between words, sentences, paragraphs, chapters, and texts. To do so, we will focus on a quantitative concept of dependency between words (homogeneity graph) that will allow us to develop a new methodology that will also be of particular help to detect the style of an author. This methodology may be also useful to detect the level of knowledge of a language of an author, to create new tools to detect plagiarism, and to make automatic summaries of texts. Thus, our motivation is to create and develop tools to identify the main structures and the level of language in written texts and specialized languages, the level of language proficiency of a written text, and the detection of elements to characterize the style of an author. So, the questions we address are the following:

- What is the most used combination of words in a corpus beyond locating the most relevant individual words behind a key concept?
- How to determine the most representative words of a text (not necessarily the most frequent)?
- How to identify the most representative phrase or phrases of a text?
- How can we characterize the level of competence of the language used in a written text?
- Can the style of an author be determined from specific parameters of a linguistic network associated with one of his written texts?
- Is it possible to associate a numerical measurement index and a mathematical structure that characterizes and identifies the author?

In the search for specific characteristics that allow the identification of an author's style, it is important to point out that many questions arise when trying to quantify and to associate a numerical measure index to a text or to a part of such text. In fact, a breakthrough in linguistics was to be able to associate a complex network structure to a text in order to identify patterns, linguistic units, syntactic and semantic structure, and so forth; so in order to solve this question, we will set our sights on some new mathematical structures related to graph theory and higher order networks that will allow us to dissect and analyze this kind of linguistic structures. Therefore, some of our efforts will be focused on associating a mathematical structure to a written text as if it were some sort of seal of identity of that text. Having in mind that the latest advances in modern linguistics are based on the treatment of a language as a system or a complex network, we will extend these ideas to a general context in which the PageRank algorithm, the mathematical structure of hypergraph and a suitable multilayer line graph will be considered. A linguistic corpus, that is, a collection of texts collected electronically according to a set of specific criteria used as a representative sample of a language or subset of that language [39], can be mathematically modeled as a multilayer complex network $G = (X, E)$, in such a way that each node $X = \{1, \dots, N\}$ is a word that appears in any of the texts that make up the linguistic corpus, and a direct link is established between two words that appear consecutively [18]. Within the multilayer network, four layers are distinguished (lexical layer—blue color; verbal layer—green color; linking layer—red color; and a fourth layer formed by the rest of the words not included in the previous layers—brown color). The color indicates the type of node (layer). The unit of analysis considered is the sentence, for this case, the words enclosed between two periods [40]. Also, it is important to note that commas and other punctuation

marks within the sentence have been removed for the analysis done, as well as it is also important to note that there was a previous analysis carried out by linguistic experts to distribute the words into the different layers of the network. In [18, 29], a linguistic corpus composed by 86 extended abstracts and papers (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>]), giving a total amount of 25,210 sentences and 147,637 words (of which 2203 are lexical words) was considered and analyzed.

Throughout this paper, we will adopt a different perspective taking into account that a linguistic corpus can be mathematically modeled as a hypergraph, where the nodes are the words of the corpus, identifying each sentence (set of words between two points) with a hyperedge composed of the specific words that are part of it. More specifically, in our approach, each node (word) is included in all the hyperedges (sentences) of which it is a part, conforming the hypergraph structure associated with the text (or the whole corpus) considered. This new perspective allows us to move from the words identified with the nodes of the hypergraph to consider the mesoscale structures: Sentences, paragraphs, chapters, and so forth.

In a first approach made in [41], we focused only on the lexical words, that is, the words of the lexical layer. For this purpose, we used a model for representing a linguistic corpus as a hypergraph in which each sentence (set of words between two dots or periods) was identified with a hyperedge whose nodes were the lexical words that were part of that sentence. In the new model we are presenting, we will combine the multilayer network structure [18, 29] with the hypergraph structure, resulting in a multilayer hypergraph model in which the hyperedges are formed not only by the lexical words in each sentence but also by all the words that are part of that sentence: Verbs, linking words, lexical words, and the rest of words. As we will see, this has important consequences and sheds light on the characteristics of the derivative graph of a multilayer hypergraph since, as it seems natural to expect, the derivative graph of a multilayer hypergraph will turn out to be a multilayer graph, bringing new light on the potential applications of this concept.

The structure of the paper is as follows. After this introduction, Section 2 introduces some basic concepts and a summary of some of the most important concepts, relationships and results related to multilayer networks, hypergraphs, and the linear and dual graphs of a hypergraph. In Section 3, a direct connection between the Jaccard index and the derivative of a hypergraph with respect to two nodes is shown. Also, several relationships, properties, and bounds related to the derivative graph of a hypergraph are obtained and a first insight related to the application of these concepts and results to the corpus under study is presented. Section 4 is devoted to define the linegraph of a multilayer hypergraph and a partial mathematical structure related to this concept (partial multiline graph), as well as to obtain certain properties and relationships between them. A realistic representation of the homogeneity graph associated to the corpus under study is also shown. In Section 5, the lexical density of the set of texts that make up the corpus analyzed is studied, and some numerical experiments and computational results are presented using three different algorithms to illustrate different ways to obtain meaningful sentences and to obtain tools that facilitate the automatic extraction of summaries from a text. Finally, in Section 6, some conclusions of this work and a proposal for future lines of work are presented.

2 | BASIC CONCEPTS AND SOME PRELIMINARY RESULTS

2.1 | Multilayer networks

A network (or graph) $G = (X, E)$ is simply a finite set of nodes $X = \{1, \dots, N\}$ connected by a set of links (or edges between certain pairs of nodes) $E = \{e_1, \dots, e_m\}$. If the links 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 $i \rightarrow j$. Throughout this paper, we will consider simple networks without loops, that is, for every $i \in X$, we have that $e_{ii} \notin E$, and also without multiple edges. We also consider the (in and out) neighbors of a node $i \in X$: $N^+(i) = \{j \in X | e_{ji} \in E\}$, $N^-(i) = \{j \in X | e_{ij} \in E\}$, and $N(i) = N^-(i) \cup N^+(i)$. If the network is undirected, the set of neighbors of a node $i \in X$ is obviously $N^i = \{j \in X | e_{ij} \in E\}$. Now, if w is a function $w : E \rightarrow (0, +\infty)$, which represents some kind of flow or intensity (data, frequency or similar) that circulates or links both nodes through the edge that joins them, in such a way that for each edge $e_{ij} \in E$, the coefficient $w(e_{ij})$ is called *weight* of e_{ij} , and we will say that $G = (X, E, w)$ is a weighted network. Now, if $G = (X, E, w)$ is a directed and weighted network, the (*weighted*) *adjacency matrix* of G is the matrix $A(G) = A = (a_{ij}) \in \mathbb{R}^{n \times n}$ given by

$$a_{ij} = \begin{cases} w(e_{ij}), & \text{if there exists the edge } e_{ij} \in E, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

For the rest of this paper, a weighted multilayer network \mathcal{M} [9], with m layers is pair $\mathcal{M} = (\mathcal{G}, C)$ where $\mathcal{G} = \{G_\alpha; \alpha \in \{1, \dots, m\}\}$ is a family of weighted graphs $G_\alpha = (X_\alpha, E_\alpha, w_\alpha)$ (called layers of \mathcal{M}), and

$$C = \{E_{\alpha\beta} \subset X_\alpha \times X_\beta \mid \alpha, \beta \in \{1, \dots, m\}, \alpha \neq \beta\}$$

is the set of interconnections between nodes of different layers G_α and G_β with $\alpha \neq \beta$. The elements of C are called *crossed layers* and we assume that they are all weighted. The set of nodes of \mathcal{M} is

$$X = \left(\bigcup_{\alpha=1}^m X_\alpha \right),$$

where $X_\alpha = \{x_1^\alpha, \dots, x_{N_\alpha}^\alpha\}$ and the set of intralayers of \mathcal{M} is $L = \{G_\alpha \mid \alpha \in \{1, \dots, m\}\}$. Each layer of L is a weighted graph $G_\alpha = (X_\alpha, E_\alpha, w_\alpha)$ with a set of nodes $X_\alpha \subset X$ and a set of edges:

$$E_\alpha = \{e_{i,j}^\alpha \mid \alpha \in \{1, \dots, m\}\},$$

where $e_{i,j}^\alpha$ represents the link that connects nodes $i, j \in X_\alpha$ and, as in the case of C , we assume that they are all weighted. In the same way, for any crossed layer

$$E_{\alpha\beta} = \{e_{i,j}^{\alpha\beta} \mid \alpha, \beta \in \{1, \dots, m\}, \alpha \neq \beta\},$$

$e_{i,j}^{\alpha\beta}$ represents the link that connects nodes $i \in X_\alpha$ and $j \in X_\beta$.

A further explanation of this notation for directed and weighted multilayer networks can be found in [9, 42, 43] and [10].

Multilayer networks provide a unified framework that, in particular, allows modeling the structural properties of specialty languages by exploring the interaction between terms and linguistic units. In our model, we consider a multilayer network consisting of four layers (lexical, verbal, linking words, and other words) [18, 29, 41]. This model is built from a specific mathematical corpus with the aim to analyze the specialty mathematical language produced by the scientific community of complex networks from the perspective of the theory and tools of complex networks. In this model, for example, the edges in the lexical layer represent links between words that appear one after the other, so that this “clustering” shows a syntagmatic relationship between words (called collocation). It is important to note that a linguistic collocation is not the result of the addition of two meanings but gives rise to and creates a new meaning that is only possible when these two words appear together (e.g., “black mail”). On the other hand, the interlayer edges may facilitate the formation and description of specialty verbs (e.g., “cluster together”) and other interlayer edges between the verbal layer and the layer of linking words may represent certain phrasal verbs.

2.2 | Hypergraphs, line graphs, and dual of a hypergraph

2.2.1 | Basic definitions

A hypergraph [41, 44–46] $\mathcal{H} = (X, \varepsilon)$ 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 $h_i \neq \emptyset$ ($i = 1, 2, \dots, n$) and $X = \bigcup_{i=1}^n h_i$. Each of these subsets is called a hyperedge (see Figure 1). In this way, hypergraphs appeared as the natural extensions of graphs to describe group interactions. The order of \mathcal{H} is the cardinality of X , that is, $|X|$, and the size of \mathcal{H} is $|\varepsilon|$. Two hyperedges $h_i, h_j \in \varepsilon$ are incident if $h_i \cap h_j \neq \emptyset$. Two nodes are adjacent if a hyperedge contains both nodes. The collection of hyperedges incident on a node i , $\varepsilon(i) = \{h \in \varepsilon \mid i \in h\}$ is called the star of i . Two nodes $i, j \in X$ are said to be similar if $\varepsilon(i) = \varepsilon(j)$. The degree of a node i is the number of hyperedges containing it, that is, $\deg(i) = |\varepsilon(i)|$. The rank of \mathcal{H} is the maximum cardinality of its hyperedges, that is, $\text{rank}(\mathcal{H}) = \max\{|h| \mid h \in \varepsilon\}$. A path P in \mathcal{H} from node i to node j is an alternate node-hyperedge sequence $i = k_1, h_1, k_2, h_2, \dots, k_r, h_r, k_{r+1} = j$ such that $k_1, k_2, \dots, k_r, k_{r+1}$ are distinct nodes (with the possibility that $i = k_1 = k_{r+1} = j$), h_1, h_2, \dots, h_r are distinct hyperedges and $\forall s \in \{1, \dots, r\} k_s, k_{s+1} \in h_s$. The integer r is called the length of the path P . Observe that if there is a path from i to j there is also a path from j to i (in that case P is said to connect i and j). A hypergraph is said to be connected if any pair of nodes is connected by a path. The distance $d(i, j)$ between two nodes i and j is the minimum length of a path connecting i and j . If there is no path between i and j it is assumed that $d(i, j) = +\infty$. Finally, the diameter of $\mathcal{H} = (X, E)$ is defined as the value $d(\mathcal{H}) = \max\{d(i, j) \mid i, j \in X\}$.

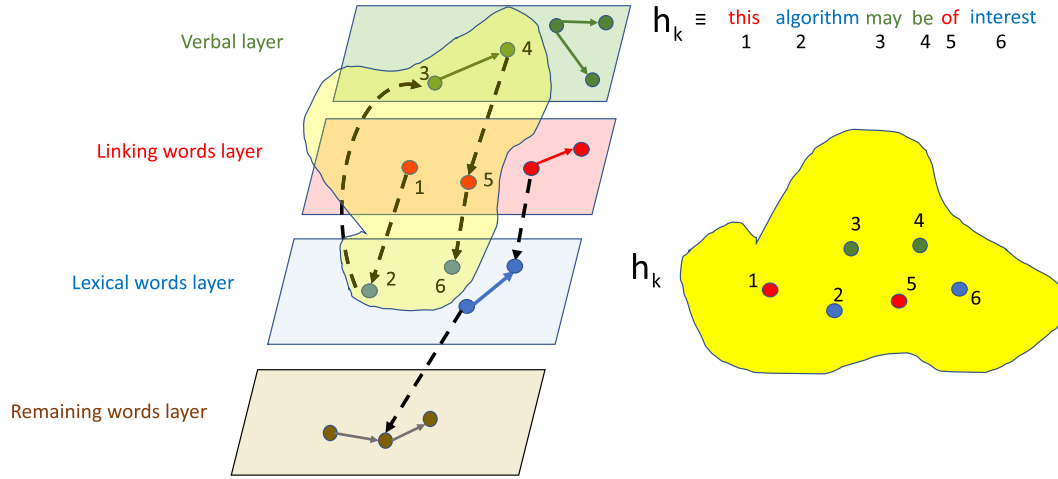


FIGURE 1 A partial representation of our multilayer linguistic hypergraph and an example of a phrase represented as a hyper-edge. Dashed lines represent *interlayer* links (edges). [Colour figure can be viewed at wileyonlinelibrary.com]

2.2.2 | Linegraph and dual graph of a hypergraph

If $\mathcal{H} = (X, \epsilon)$ is a hypergraph, the linegraph associated to \mathcal{H} is the graph $L(\mathcal{H}) = (\epsilon, E')$, where if $h_i, h_j \in \epsilon$, 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. Note that this concept is a particular case of the concept of intersection graph [47]. If $\mathcal{H} = (X, \epsilon)$ is a hypergraph, the dual hypergraph associated with \mathcal{H} is the hypergraph $\mathcal{H}^* = (\epsilon, X')$ 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 $(\mathcal{H}^*)^* = \mathcal{H}$. It naturally makes sense to consider the function Π_2 that converts a hypergraph $\mathcal{H} = (X, \epsilon)$ into a graph $\Pi_2(\mathcal{H}) = (X, E')$ as follows:

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

So, for any hypergraph \mathcal{H} , we have that $\Pi_2(\mathcal{H}^*) = L(\mathcal{H})$. Moreover, if $G = (X, E)$ is a graph, with $X = \{1, \dots, N\}$, we may also consider the dual hypergraph $G^* = (E, \epsilon)$ of G where $\epsilon = \{h_1, \dots, h_n\}$ and $\forall i \in \{1, \dots, n\}$ we consider the corresponding hyperedge $h_i = \{e_j \in E | i \in e_j\}$, and also $\Pi_2(G^*) = L(G)$. On the other hand, if I is the incidence matrix of \mathcal{H} , then its transpose matrix I^t is the incidence matrix of \mathcal{H}^* . In fact,

$$(I(\mathcal{H})^t) \cdot (I(\mathcal{H})) = \widetilde{A}(\mathcal{H}) = (\widetilde{a}_{ij}) \in \mathbb{R}^{|\epsilon| \times |\epsilon|},$$

and

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

where

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

and

$$a_{ij} = \begin{cases} |\{h \in \epsilon | i \in h\}| & \text{if } i = j, \\ |\{h \in \epsilon | i, j \in h\}| & \text{if } i \neq j. \end{cases} \quad (2)$$

It is important to highlight that nowadays the interest in all these structures is increasing more and more.

3 | SIMILARITIES AND DERIVATIVE GRAPH OF A HYPERGRAPH: SOME PROPERTIES AND RELATIONSHIPS

One of the classical methods for comparing the degree of coincidence or similarity between two sets is the Jaccard index [48]. The basic Jaccard index to compare the degree of coincidence between two sets A and B can be obtained from the formula

$$\mathcal{J}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad 0 \leq \mathcal{J}(A, B) \leq 1. \quad (3)$$

Different generalizations of Jaccard index have been introduced in the literature as its use in different applications has become more generalized and widespread [49–55] (including the overlap index $C(A, B)$). Thus,

$$\mathcal{J}_n(A, B) = \frac{|A \cap B|^n}{|A \cup B|^n}, \quad \mathcal{I}(A, B) = \frac{|A \cap B|}{\min\{|A|, |B|\}},$$

and

$$C(A, B) = \mathcal{J}(A, B) \cdot \mathcal{I}(A, B). \quad (4)$$

Observe that $0 \leq C(A, B) \leq 1$. At this point, we can recall the following definition [41]:

Definition 1. Given a hypergraph $\mathcal{H} = (X, \varepsilon)$, with $A(\mathcal{H}) = (a_{ij}) \in \mathbb{R}^{N \times N}$, the derivative hypergraph of \mathcal{H} with respect to the pair of nodes $i, j \in X$ is the numerical value

$$\frac{\partial \mathcal{H}}{\partial \{i, j\}} = \frac{|\varepsilon(i)| - |\varepsilon(i) \cap \varepsilon(j)| + |\varepsilon(j)| - |\varepsilon(i) \cap \varepsilon(j)|}{|\varepsilon(i) \cap \varepsilon(j)|} = \frac{a_i - a_{ij} + a_j - a_{ij}}{a_{ij}} = \frac{a_i - 2a_{ij} + a_j}{a_{ij}}. \quad (5)$$

Obviously, if two nodes $i, j \in X$ are similar, that is, if $\varepsilon(i) = \varepsilon(j)$, then $\frac{\partial \mathcal{H}}{\partial \{i, j\}} = 0$.

So, looking at Figure 2, we get $\frac{\partial \mathcal{H}}{\partial \{1, 2\}} = \frac{2-2+3-2}{2} = \frac{1}{2}$, $\frac{\partial \mathcal{H}}{\partial \{2, 3\}} = 1$, $\frac{\partial \mathcal{H}}{\partial \{5, 6\}} = 0$, $\frac{\partial \mathcal{H}}{\partial \{1, 3\}} = 3$, $\frac{\partial \mathcal{H}}{\partial \{2, 4\}} = 2$, $\frac{\partial \mathcal{H}}{\partial \{1, 4\}} = 1$, $\frac{\partial \mathcal{H}}{\partial \{3, 5\}} = 2$, and $\frac{\partial \mathcal{H}}{\partial \{3, 4\}} = +\infty = \frac{\partial \mathcal{H}}{\partial \{4, 5\}} = \frac{\partial \mathcal{H}}{\partial \{2, 6\}}$.

It is evident that the derivative of a hypergraph $\mathcal{H} = (X, \varepsilon)$ with respect to the nodes $i, j \in X$ provides us with a method to compare the role played by both nodes in the context of such hypergraph, so that the smaller the derivative, the greater the similarity between these nodes and, if the derivative is zero, both nodes have identical behavior as far as the structure of the hypergraph is concerned. The following proposition establishes a direct relationship between the Jaccard index and the concept of the derivative of a hypergraph with respect to two nodes:

Proposition 1. For any hypergraph $\mathcal{M} = (X, \varepsilon)$ and $\forall i, j \in X$, the following identity is satisfied:

$$\mathcal{J}(\varepsilon(i), \varepsilon(j)) = \frac{1}{1 + \frac{\partial \mathcal{H}}{\partial \{i, j\}}}. \quad (6)$$

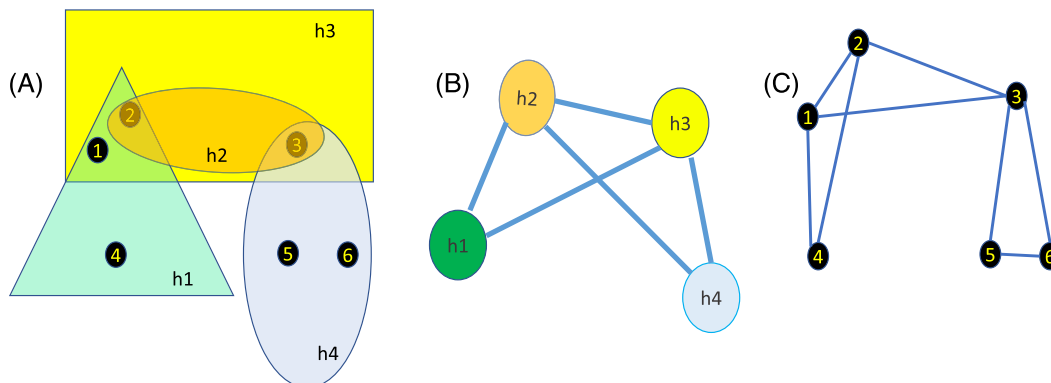


FIGURE 2 An example: $\mathcal{H} = (\{1, 2, 3, 4, 5, 6\}, \{h1, h2, h3, h4\})$ (A), $L(\mathcal{H}) = \Pi_2(\mathcal{H}^*)$ (B), and $\Pi_2(\mathcal{H})$ (C). [Colour figure can be viewed at wileyonlinelibrary.com]

Proof. Given $i, j \in X$, we have that

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial \{i, j\}} &= \frac{|\varepsilon(i)| - |\varepsilon(i) \cap \varepsilon(j)| + |\varepsilon(j)| - |\varepsilon(i) \cap \varepsilon(j)|}{|\varepsilon(i) \cap \varepsilon(j)|} \\ &= \frac{|\varepsilon(i) \cup \varepsilon(j)| - |\varepsilon(i) \cap \varepsilon(j)|}{|\varepsilon(i) \cap \varepsilon(j)|} = \frac{|\varepsilon(i) \cup \varepsilon(j)|}{|\varepsilon(i) \cap \varepsilon(j)|} - 1 = \frac{1}{\mathcal{J}(\varepsilon(i), \varepsilon(j))} - 1. \end{aligned}$$

□

This result establishes the way in which the Jaccard index (which measures the similarity between two sets) is related to the concept of the derivative of a hypergraph with respect to two nodes (which measures the dissimilarity between their corresponding stars). On the other hand, as $\forall i \in X$, we have that $|\varepsilon(i)| \leq \text{rank}(\mathcal{H})$; there is no difficulty in obtaining the following corollaries:

Corollary 1. Given a hypergraph $\mathcal{M} = (X, \varepsilon)$ and $i, j \in X$, if $|\varepsilon(i) \cap \varepsilon(j)| \neq \emptyset$, then $\frac{\partial \mathcal{H}}{\partial \{i, j\}} \leq 2\text{rank}(\mathcal{H}) - 1$.

Corollary 2. If $\mathcal{M} = (X, \varepsilon)$ is a hypergraph and $i, j \in X$ are two nodes such that $|\varepsilon(i) \cap \varepsilon(j)| \neq \emptyset$, then $\mathcal{J}(\varepsilon(i), \varepsilon(j)) \geq \frac{1}{2\text{rank}(\mathcal{H})}$.

Now, from the values $\frac{\partial \mathcal{H}}{\partial \{i, j\}}$ obtained for each pair of nodes $i, j \in X$, the derivative graph $\partial \mathcal{H}$ and the homogeneity graph $HG(\mathcal{H})$ can be constructed [41] (see Figure 3). So, $\partial \mathcal{H}$ 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 when $\frac{\partial \mathcal{H}}{\partial \{i, j\}} = 0$ the nodes i and j collapse into a single node (ij) and, when $\frac{\partial \mathcal{H}}{\partial \{i, j\}} = \infty$, the edge $\{i, j\}$ does not exist in the derivative graph. Once the graph $\partial \mathcal{H}$ has been constructed, 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 derivative graph $\partial \mathcal{H}$.

As it was already indicated in the introduction, in [41] we presented a first vision and approximation of the model in the mesoscale considering each sentence as formed only by its lexical words (i.e., the words of the lexical layer), representing the linguistic corpus under study as a hypergraph in which each sentence was identified with a hyperedge whose nodes were the lexical words that were part of that sentence. In the new model we are presenting, each sentence is formed by the set of all the words that are part of that sentence: Verbs, linking words, lexical words, and the rest of the words.

To give an example extracted from the hypergraph \mathcal{H} formed by all the sentences of the corpus under study

$$\frac{\partial \mathcal{H}}{\partial \{\text{Monte, Carlo}\}} = \frac{\partial \mathcal{H}}{\partial \{\text{differential, Runge-Kutta}\}} = 0.$$

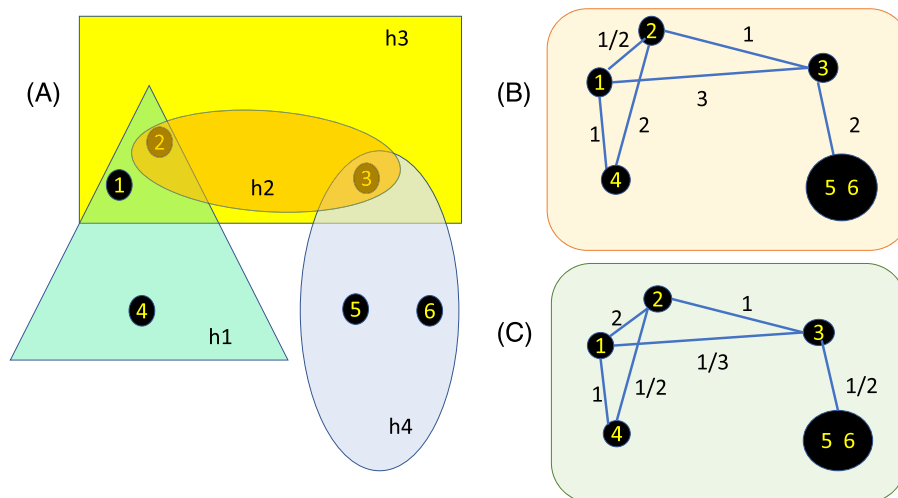


FIGURE 3 Example of hypergraph \mathcal{H} (A), its derivative graph $\partial \mathcal{H}$ (B), and its homogeneity graph $HG(\mathcal{H})$ (C). As it can be seen, nodes 5 and 6 are similar. [Colour figure can be viewed at wileyonlinelibrary.com]

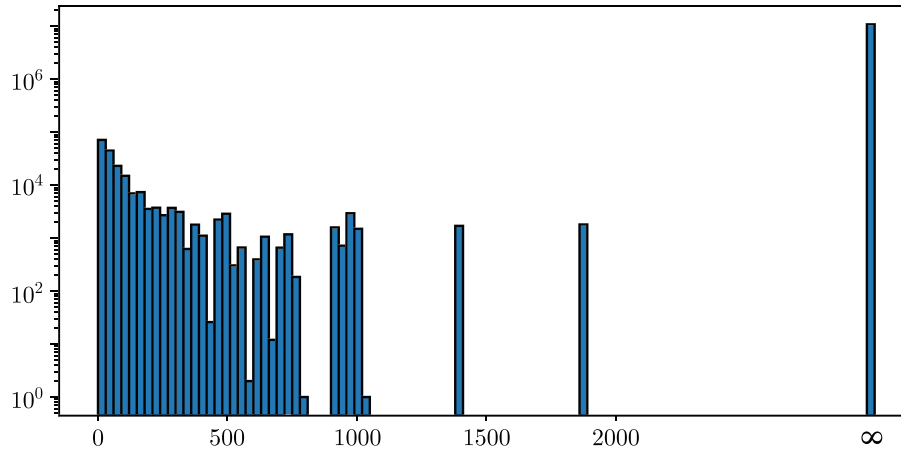


FIGURE 4 Histogram clustering the number of pairs of words $\{i, j\}$ by the value of $\frac{\partial \mathcal{H}}{\partial \{i, j\}}$. [Colour figure can be viewed at wileyonlinelibrary.com]

On the other hand, the histogram of Figure 4 shows that there are nearly 10^5 pairs of words $\{i, j\}$ such that $0 \leq \frac{\partial \mathcal{H}}{\partial \{i, j\}} \leq 10$ and more than 10^7 pairs of words $\{i, j\}$ whose derivative is $+\infty$.

Finally, to conclude this section, it is not difficult to prove the following result:

Proposition 2. *If the hypergraph $\mathcal{H} = (X, \varepsilon)$ is a connected hypergraph, then the derivative graph $\partial \mathcal{H}$ and the homogeneity graph $HG(\mathcal{H})$ are connected graphs. Moreover, we have the following relationship among the corresponding diameters:*

$$d(\mathcal{H}) = d(\partial \mathcal{H}) = d(HG(\mathcal{H})).$$

Sketch of proof: If $k_s, k_{s+1} \in h_{s+1}$ are two consecutive nodes of a path P in \mathcal{H} from node i to node j

$$i = k_1, h_1, k_2, h_2, \dots, k_r, h_r, k_{r+1} = j,$$

obviously $0 \leq \frac{\partial \mathcal{H}}{\partial \{k_s, k_{s+1}\}} < +\infty$ and, if $\frac{\partial \mathcal{H}}{\partial \{k_s, k_{s+1}\}} = 0$, then the hyperedge h_{s+1} can be removed from a path whose length coincides with the distance between i and j , so the elimination of this hyperedge does not affect the length of the considered diameters.

4 | MULTILAYER HYPERGRAPHS

As we have seen, the derivative of a hypergraph $\mathcal{H} = (X, \varepsilon)$ regarding the two nodes $i, j \in X$ provides us with a method to compare the role played by both nodes in the context of that hypergraph, so that the smaller the derivative $\frac{\partial \mathcal{H}}{\partial \{i, j\}}$, the greater the similarity between the roles played by both nodes in that hypergraph context and, if the derivative is zero, both nodes have identical behavior as far as the structure of the hypergraph is concerned.

Definition 2. A multilayer hypergraph with ℓ layers is a pair

$$\mathcal{M} = (\{\mathcal{H}_1, \dots, \mathcal{H}_\ell\}, \Theta),$$

where $\{\mathcal{H}_1, \dots, \mathcal{H}_\ell\}$ is a family of hypergraphs (called layers of \mathcal{M}), $\mathcal{H}_k = (X_k, \varepsilon_k) \forall k \in \{1, \dots, \ell\}$, $X_r \cap X_s = \emptyset$ if $r \neq s$,

$$\Theta \subset \mathcal{P} \left(\bigcup_{k=1}^{\ell} X_k \right),$$

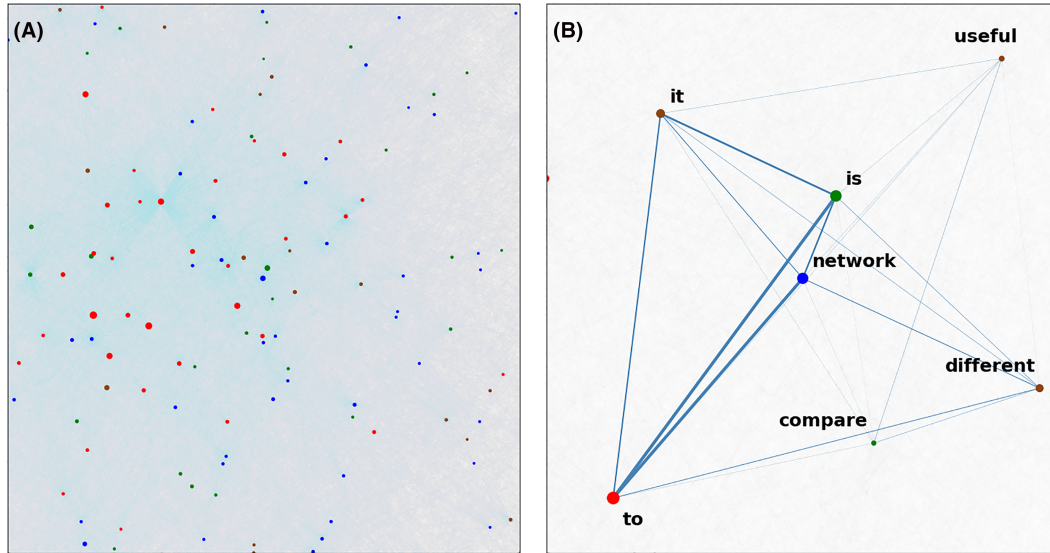


FIGURE 5 Homogeneity graph $HG(\mathcal{H})$ of the corpus \mathcal{H} under analysis. On the right, a zoom of a part of this graph: the thickness of each edge is proportional to its weight. Note that the colors of the nodes indicate the layer to which they belong. [Colour figure can be viewed at wileyonlinelibrary.com]

and $\forall k \in \{1, \dots, \ell\} \epsilon_k \subset \Theta$. Observe that $\mathcal{P}\left(\bigcup_{k=1}^{\ell} X_k\right)$ denotes the powerset of $\bigcup_{k=1}^{\ell} X_k$. Given $h \in \Theta$, if $\exists \epsilon_k$ such that $h \in \epsilon_k$, we will say that h is an intra-hyperedge. On the other hand,

$$C_{\mathcal{M}} = \{h \in \Theta \mid \exists i, j \in \{1, \dots, \ell\}, i \neq j \dots h \cap X_i \neq \emptyset \wedge h \cap X_j \neq \emptyset\}$$

is the set of crossed hyperedges.

Note that if $\ell = 1$, the multilayer hypergraph \mathcal{M} is a classic hypergraph.

If \mathcal{M} is a multilayer hypergraph, the projection hypergraph of \mathcal{M} is the hypergraph $proj(\mathcal{M}) = (X, \Theta)$ where

$$X = \bigcup_{k=1}^{\ell} X_k.$$

Now, it is obvious that if $i \in X_k, j \in X_l, k \neq l$, and $\frac{\partial \mathcal{H}}{\partial \{i,j\}} > 0$, then the corresponding edge in the derivative graph is a cross edge between the corresponding layers. On the other hand, if $\frac{\partial \mathcal{H}}{\partial \{i,j\}} = 0$, that is, when these nodes are similar, it is necessary to indicate the layer to which the new node resulting from collapsing these two nodes into one will belong. A simple way to do this is to establish a priority between layers, so that when two nodes from different layers collapse, the resulting node is incorporated in the layer of higher priority. Bearing this in mind, the following theorem arises directly from the definition of multilayer hypergraphs and the concepts of derivative and homogeneity graph associated to a multilayer hypergraph:

Theorem 1. *If $\mathcal{M} = (\{\mathcal{H}_1, \dots, \mathcal{H}_{\ell}\}, \Theta)$ is a multilayer hypergraph, where $\mathcal{H}_k = (X_k, \epsilon_k)$, and $\exists r, t \in \{1, \dots, \ell\}, r \neq t$ and $\exists i \in X_r, j \in X_t$ such that $\frac{\partial \mathcal{H}}{\partial \{i,j\}} > 0$, then both its derivative graph $\partial \mathcal{M}$ and its homogeneity graph $HG(\mathcal{M})$ are multilayer networks.*

A representation of the homogeneity graph $HG(\mathcal{H})$ associated with the corpus under study can be seen in Figure 5. Note that the color of each node is indicative of the layer to which it belongs.

4.1 | Line graphs of a multilayer hypergraph

An extension of the line graph concept to multiplexed networks was made [56] in order to analyze multiplexed and multilayer network systems. Now, after the definition of multilayer hypergraph, it is necessary to consider the different definitions of the line graph for this type of hypergraphs.

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

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

On the other hand, as it can be easily verified, the line graph of a hypergraph is a particular case of the intersection graph concept [47]. But if the nodes belonging to each of the different layers have different nature and we can distinguish between them (for example, red, green or blue layer nodes), we can introduce different partial line graphs associated to the hypergraph, one for each layer (or type of nodes) of the multilayer hypergraph. Thus, we can establish the following definition:

Definition 3. Let $\mathcal{M} = (\{\mathcal{H}_1, \dots, \mathcal{H}_\ell\}, \Theta)$ be a multilayer hypergraph, with $\mathcal{H}_k = (X_k, \epsilon_k)$ and

$$\Theta \subset \mathcal{P} \left(\bigcup_{k=1}^{\ell} X_k \right).$$

Given $k \in \{1, \dots, \ell\}$, we will call the *partial multilinegraph of \mathcal{M} with respect to the k layer $\mathcal{H}_k = (X_k, \epsilon_k)$* the linegraph $L_k(\mathcal{M}) = (\Theta, E'_k)$, where if $h_i, h_j \in \Theta$, then

$$\{h_i, h_j\} \in E'_k \iff ((h_i \cap h_j) \cap X_k) \neq \emptyset.$$

Similarly, if $k, r \in \{1, \dots, \ell\}$, we will call the partial multilinegraph of \mathcal{M} with respect to the layers \mathcal{H}_k and \mathcal{H}_r the linegraph $L_{k,r}(\mathcal{M}) = (\Theta, E'_{k,r})$, where if $h_i, h_j \in \Theta$, then

$$\{h_i, h_j\} \in E'_{k,r} \iff ((h_i \cap h_j) \cap (X_k \cap X_r)) \neq \emptyset.$$

Obviously, the above definition can be extended to obtain the partial multilinegraph of a multilayer hypergraph with respect to three or more layers.

The following relationships among the diameters of the partial multiline graph is of particular interest in relation to the application of these concepts to the case study of the multilayer linguistic hypergraph that we will discuss in the next section. The proof is immediate, bearing in mind that all the edges belonging to the partial multilinegraphs are also in $L(\text{proj}(\mathcal{M}))$:

Proposition 3. If $\mathcal{M} = (\{\mathcal{H}_1, \dots, \mathcal{H}_\ell\}, \Theta)$ is a multilayer hypergraph, with $\mathcal{H}_k = (X_k, \epsilon_k)$ and

$$\Theta \subset \mathcal{P} \left(\bigcup_{k=1}^{\ell} X_k \right),$$

then $\forall k, r \in \{1, \dots, \ell\}$ we have that

$$d(L(\text{proj}(\mathcal{M})) \leq d(L_k(\mathcal{M})) \leq d(L_{k,r}(\mathcal{M})).$$

5 | A MULTILAYER LINGUISTIC HYPERGRAPH FOR THE AUTOMATIC EXTRACTION OF TEXT SUMMARIES

In the new model we are presenting, we combine the multilayer network structure with the hypergraph structure resulting in a multilayer hypergraph model in which the hyperedges are formed by all the words that are part of that sentence: Verbs, linking words, lexical words, and the words tagged as rest of the words so that in the derivative graph and in the homogeneity graph also appear different layers as it can be seen in Figure 5 where the color of each node indicates the layer to which it belongs.

It is important to note that the mesoscopic structure of a given text or corpus can be stratified at different levels since, for example, by considering paragraphs as a set of sentences, and chapters as a set of paragraphs, these linguistic concepts can be incorporated into the model and treated using a methodology similar to the one described in which the hyperedges are not necessarily the sentences of the text. This methodology will undoubtedly allow us to characterize a text or set of texts from the derivatives of the corresponding graphs and hypergraphs, respectively.

Thus, by calculating the derivative graph (and the homogeneity graph) associated with the linguistic hypergraph composed of all the sentences of a corpus or a text, we will obtain a greater or lesser degree of disparity (respectively, similarity) between each pair of sentences that form such text (or corpus) from the words shared by them, with the peculiarity that these quantitative measures are reflected and represented in their corresponding derivative graph and homogeneity graph. Therefore, the derivative graph of a hypergraph constitutes a specific feature of that hypergraph that is intrinsic and characteristic of it. In the case that the analysis refers to the hypergraph associated with a corpus or set of sentences, its derivative graph will provide specific and intrinsic qualitative and quantitative characteristics of that set of texts.

To extract a summary and to determine which sentences are the most representative of a text or corpus, we will use the PageRank biplex algorithm introduced in [57] and we will apply this algorithm using three different criteria, bearing in mind the specific network to which we are applying the algorithm.

PageRank and its different variants constitute a family of algorithms related to the concept of random walker and that allow to assign a certain numerical value to the nodes of a complex network and to order them according to their importance or centrality [58–62]. Specifically, the personalized PageRank of an individual term (or node) i of a network is the i component of the stationary state $\pi_0 \in \mathbb{R}^n$ ($\|\pi_0\| = 1$) of the random walker with transition matrix

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

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

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

To compare the different rankings we will obtain, we use the standard mathematical tool known as *Kendall's τ correlation coefficient*[63], which is commonly used to compare different rankings or orderings of the same set of elements. Thus, if we take two rankings r_1 and r_2 of a set of N elements, then Kendall's correlation coefficient is defined as

$$\tau(r_1, r_2) = \frac{\tilde{K}(r_1, r_2) - K(r_1, r_2)}{\binom{N}{2}}, \tag{7}$$

where $\tilde{K}(r_1, r_2)$ denotes the number of pairs (i, j) that do not change its mutual relative position with respect to both rankings r_1 and r_2 , and $K(r_1, r_2)$ denotes the number of pairs (i, j) that change its mutual relative position in that rankings.

To perform our study on the multilayer hypergraph \mathcal{H} in which the nodes are all the words that compose the corpus, distributed in the four layers that compose the hypergraph, and the hyperedges are the phrases (sets of all the words located between two periods), we will use the same methodology as in [29] and [18] to associate to each node its corresponding PageRank, with the idea of ordering not only the words of each of the linguistic layers considered (with the aim of ordering

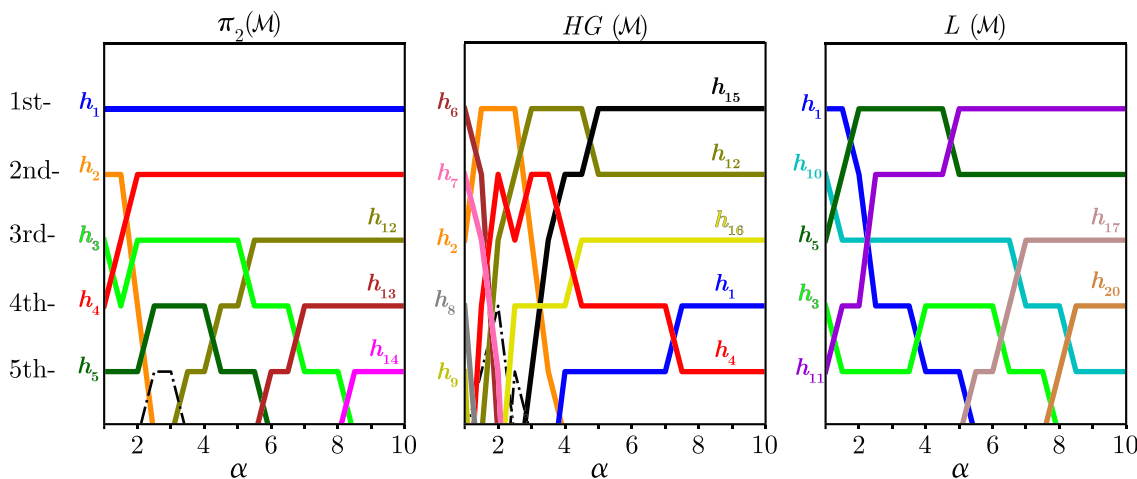


FIGURE 6 Ranking of the most relevant phrases when varying the weights of the edges pointing to the words of the lexical layer in terms of the multiplicative parameter of the weights $\alpha \in [1,10]$. A list of these specific phrases is included in Appendix A. [Colour figure can be viewed at wileyonlinelibrary.com]

the sentences to which they belong) but also the sentences themselves considered as nodes of the corresponding linear graph according to their importance [58–60, 64, 65]. For this purpose, bearing in mind that for the PageRank calculation used throughout this work we have considered the algorithm described in [57], we will apply this algorithm on three different frames obtained from the application of three different criteria:

1. **Ranking 1.** To obtain this ranking, we first have built a graph on which to apply the PageRank algorithm. To accomplish this, we convert each hyperedge of \mathcal{H} into a clique to obtain the projection graph $\Pi_2(\mathcal{H})$, which is obviously a multilayer network. After this, we transform the undirected graph into a directed graph by considering each undirected edge as an edge with the double direction. The objective is to artificially increase the weight (and the importance of the links) of the edges pointing to the words of the lexical layer, multiplying these edges by a parameter $\alpha \in [1, 20]$, in order to analyze the variations produced in the ranking by varying this parameter (and, consequently, to obtain new rankings by giving more importance to the lexical layer as we increase the value of α). Edges going from words in the lexical layer to words in other layers are left with their original value. At this point, we must highlight that a linguistic decision following the criteria of several experts was taken in order to catalog the terms (words) and assign them to one or another layer. Particularly, the expert and linguistic criteria was crucial for assigning the terms to the lexical layer. Now, taking into account that the average number of words of a sentence within the corpus under study is 18.0496 and that, therefore, the local lexical density is 18.0496, we can deduce, reasoning similarly to [66], where $\mathbb{E}(\ell)$ is the mathematical expectation, that the damping factor corresponding to this configuration is 0.9475, since

$$\begin{aligned} 18.0496 = \mathbb{E}(\ell) &= \sum_{k=0}^{\infty} k \cdot \mathbb{P}(\ell = k) = \sum_{k=1}^{\infty} k \cdot (1 - q) \cdot q^k \\ &= (1 - q) \cdot q \sum_{k=1}^{\infty} k \cdot q^{k-1} = \frac{q}{1 - q}. \end{aligned}$$

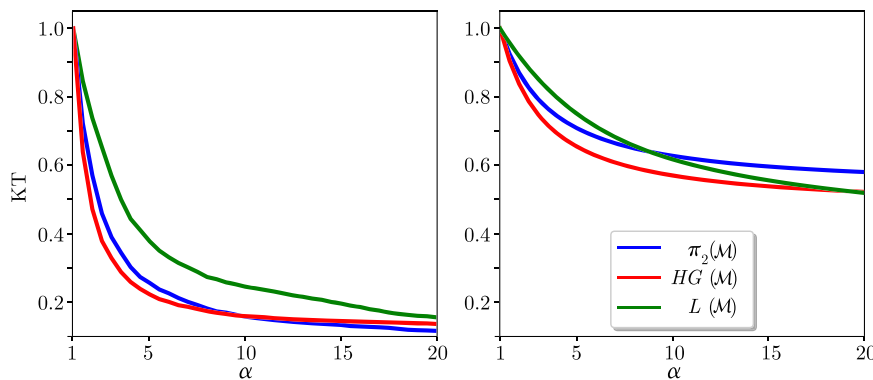


FIGURE 7 Variation of the Kendall- τ parameter when varying the α parameter. The left panel corresponds to such variation when the first 100 sentences are considered. The right panel is the variation obtained when considering the whole corpus sentences. [Colour figure can be viewed at wileyonlinelibrary.com]

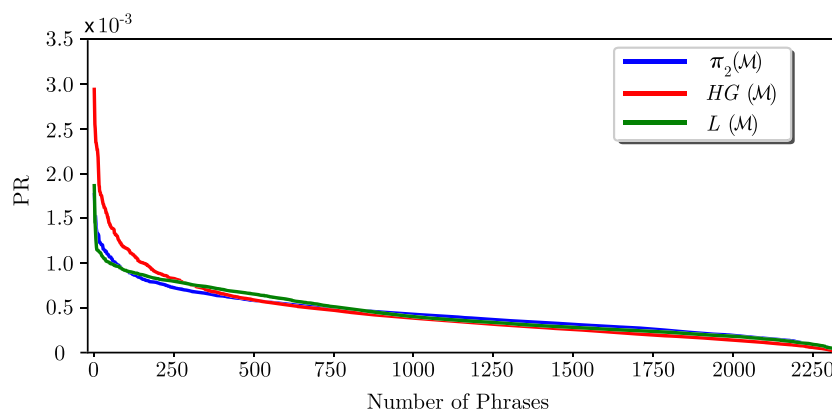


FIGURE 8 Number of sentences whose PageRank reaches a specific value for $\alpha = 20$. [Colour figure can be viewed at wileyonlinelibrary.com]

Finally, once the PageRank of each word has been obtained, this PageRank is distributed proportionally among the phrases in which this word appears to obtain the PageRank of each phrase.

2. **Ranking 2.** To determine this ranking, we first have built the derivative (multilayer) graph and its associated homogeneous graph. After this, and similar to the previous case, we transform the undirected graph into a directed graph by considering each undirected edge as an edge with the double direction and, again operating analogously to the previous case, we multiply the edges pointing to the words of the lexical layer by the parameter $\alpha \in [1, 20]$. Once this has been done, 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 17.6823 (since, after collapsing words pairs $\{i, j\}$ such that $\frac{\partial H}{\partial \{i, j\}} = 0$, the average length of sentences decreases, albeit slightly), the damping factor corresponding to this configuration is 0.9464. Now, once the PageRank of each of the words has been obtained, and proceeding in the same way as in the previous case, this PageRank is distributed proportionally among the sentences in which this word appears in order to obtain the PageRank of each sentence. In the case that two nodes from different layers collapse into one, that is, if these nodes are similar, the priority established for assigning the new node to one of the layers is as follows: Lexical layer, verbal layer, linking words layer, remaining words layer, so that, for example, if a node from the linking word layer and another from the verbal layer are similar, the new collapsed node will be assigned to the verbal layer. Figure 5 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.
3. **Ranking 3.** Finally, to obtain this ranking, we will apply the PageRank algorithm considered on the multilayer network $L(\mathcal{H}) = \Pi_2(\mathcal{H}^*)$ in which each node is a phrase of the corpus, so that two phrases will be connected if they have at least one word in common. As in the previous cases, we want to analyze the different rankings obtained according to the number of lexical words that these sentences share. Therefore, we first transform the undirected graph into a directed one by considering each undirected edge as an edge with the double direction. The weight of the edge between two sentences will be higher if they share a larger number of lexical words. For example, if sentence s_1 and sentence s_2 share three lexical words and two words from other layers, and sentence s_1 appears repeated twice, the weight of edge $w(s_1 \rightarrow s_2)$ will be $3\alpha + 2$, and the weight of edge $w(s_2 \rightarrow s_1)$ will be $2(3\alpha + 2) = 6\alpha + 4$. 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 27.4186, in this context, the damping factor corresponding to this configuration is 0.9648.

To properly apply the PageRank algorithm on each of the graphs obtained after the application of the above three criteria, the corresponding value of q is the probability that a random walker does not vary its trajectory by moving from one node to another directly connected to the current node, instead of jumping to another of the nodes of this network not necessarily connected to this node. Therefore, to complete the description of all the elements necessary to apply the PageRank algorithm, we will point out that in the case of Ranking 1 and Ranking 2 the personalization vector considered is the (relative) frequency of the set of all the words in the corpus under study, and for Ranking 3 the personalization vector considered is the (relative) frequency of each sentence included in that corpus.

Figure 6 shows important differences between the rankings obtained by applying each of the three criteria and the variations obtained in these rankings by artificially increasing the importance of the lexical layer by means of the α parameter. Noteworthy are the important variations obtained in Ranking 2 by slightly increasing the importance of the lexical layer.

On the other hand, in Figure 7, we can see the variation of the Kendall- τ parameter when varying the α parameter, and in Figure 8, it is possible to observe an outstanding difference between the way in which the slope describing PageRank varies in the case of Ranking 2 in comparison with that described in the other two cases, which allows obtaining in the first case a smaller number of representative phrases with a high PageRank. This feature makes this criterion particularly attractive in view of the development of a potential tool to summarize text automatically.

6 | CONCLUSIONS AND FUTURE WORKS

We introduce, in the framework of the multilayer hypergraphs defined in this work, the derivative graph and the homogeneity graph of this type of hypergraphs, and we analyze these concepts and their new associated tools as two useful structures that allow us to associate a great variety of characteristic properties of a given hypergraph.

Based on the explicit relation obtained between the value of the derivative of a hypergraph with respect to two nodes i and j and the Jaccard index of its corresponding stars $\varepsilon(i)$ and $\varepsilon(j)$, we obtain certain characteristic features and properties of the hypergraph under study and of the model represented by this structure. Moreover, a new structure, the partial

multiline graphs of a multiline hypergraph and some of its relationship with the projection graph of the multiline hypergraph, is also introduced allowing us to study the relationship between the different layers from the perspective of a partial linegraph. This makes it possible to analyze the relationships between sentences in a text according to whether they share, for instance, a lexical word, a verb, or a linking word.

The application of the algorithm introduced in [65] on the corpus under study has made it possible to obtain three rankings (according to the criteria used) in which the most representative nodes of the complete hypergraph associated with the corpus under study are ordered in order to identify the most representative phrases, advancing in the idea of obtaining tools capable of extracting automatic summaries of texts. This methodology can also be used for each of the hypergraphs associated with any type of text, and can even be adapted to obtain the key words and the most characteristic terms of the text under analysis. This facilitates the answer to the first three questions posed in the introduction on the location of the words, their combinations, and the most representative phrases of a text.

The hypergraphs corresponding to any kind of texts, as well as their corresponding derivative graphs (and their associated homogeneity graphs), make it possible to establish similarities and differences between these texts. In particular, the tools and concepts introduced allow us not only to associate a numerical index useful to quantify the similarity or dissimilarity between the different texts of a given corpus but also to highlight numerical elements and parameters of the multilayer derivative graph that, from this work, we can associate to a text (or corpus), and that can be considered as a kind of “mathematical signature” characteristic of that text. Moreover, the methodology developed can be applied at different levels. Thus, for example, for a given text, it is possible to consider not only the structure of the defined multilayer hypergraph in which the nodes are the words and the hypergraphs are the sentences, but also, and successively, a hypergraph in which the nodes are the words and the hypergraphs are the paragraphs, and another in which the nodes are the sentences and the hypergraphs are the paragraphs. Considering this sequence of mathematical structures (multilayer derivative graphs) and the different characteristic parameters of these structures (such as the diameter or degree distribution, the centrality of the hypergraph, the efficiency, and others), we will be able to point out singular elements that allow us to characterize and compare different texts, and even group them according to some of the characteristics that constitute their hallmark in terms of style.

This specific study, as well as the possible usefulness of the tools introduced to obtain technical characteristics related to the styles of the different authors and the level of linguistic competence of any text written in English, will be analyzed and developed in future works.

We believe it is important to highlight the usefulness of these concepts for their possible application to text classification, text summarization, machine translation, stylometry, and authorship detection.

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

ACKNOWLEDGEMENTS

This work has been partially supported by projects PGC2018-101625-B-I00 (Spanish Ministry, AEI/FEDER, UE) and M1967 Grant (Rey Juan Carlos University, Spain). The authors acknowledge the usage of the resources, technical expertise, and assistance provided by the supercomputer facility CRESCO of ENEA in Portici (Italy).

CONFLICT OF INTEREST STATEMENT

This work does not have any conflicts of interest.

ORCID

Ángeles Criado-Alonso  <https://orcid.org/0000-0001-8608-1806>

David Aleja  <https://orcid.org/0000-0002-4477-1638>

Miguel Romance  <https://orcid.org/0000-0001-9259-9716>

Regino Criado  <https://orcid.org/0000-0001-7954-6169>

REFERENCES

1. R. Albert and A. L. Barabasi, *Statistical mechanics of complex networks*, Rev. Mod. Phys **74** (2002), 47–97.
2. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, *Complex networks: Structure and dynamics*, Phys. Rep. **424** (2006), 175–308.
3. J. Cong and H. Liu, *Approaching human language with complex networks*, Phys. Life Rev. **11** (2014), no. 4, 598–618.
4. R. Criado, J. Flores, A. García del Amo, and M. Romance, *Analytical relationships between metric and centrality measures of a network and its dual*, J. Comput. Appl. Math. **235** (2011), no. 7, 775–1780.
5. E. Omodei, M. De Domenico, and A. Arenas, *Evaluating the impact of interdisciplinary research: A multilayer network approach*, Netw. Sci. **5** (2017), no. 2, 235–246.
6. L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti, *Eigenvector centrality of nodes in multiplex networks*, Chaos **23** (2013), no. 3, 33131.
7. F. Battiston, V. Nicosia, and V. Latora, *The new challenges of multiplex networks: Measures and models*, Eur. Phys. J. Spec. Top. **226** (2017), 401.
8. F. Battiston, V. Nicosia, and V. Latora, *Structural measures for multiplex networks*, Phys. Rev. E **89** (2014), no. 3, 2804.
9. S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, *The structure and dynamics of multilayer networks*, Phys. Rep. **544** (2014), no. 1, 1–122.
10. M. De Domenico, A. Solé-Ribalta, E. Cozzo, M. Kivela, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas, *Mathematical formulation of multi-layer networks*, Phys. Rev. X **3** (2013), 41022.
11. M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, *Ranking in interconnected multilayer networks reveals versatile nodes*, Nat. Commun. **6** (2015), 6868.
12. F. Pedroche, R. Criado, J. Flores, E. García, and M. Romance, *On PageRank versatility for multiplex networks: Properties and some useful bounds*, Math. Methods Appl. Sci. **43** (2020), no. 14, 8158–8176.
13. I. Amburg, N. Veldt, and A. R. Benson, *Clustering in graphs and hypergraphs with categorical edge labels*, Proceedings of the Web Conference, 2020, pp. 706–717.
14. A. Benson, *Three hypergraph eigenvector centralities*, SIAM J. Math. D.S. **1** (2019), no. 2, 293–312.
15. K. Kovalenko, M. Romance, E. Vasilyeva, D. Aleja, R. Criado, D. Musatov, A. M. Raigorodskii, J. Flores, I. Samoylenko, K. Alfaro-Bittner, M. Perc, and S. Boccaletti, *Vector centrality in hypergraphs*, Chaos Solitons Fractals **162** (2022), 112397.
16. R. Lambiotte, M. Rosvall, and I. Scholtes, *From networks to optimal higher-order models of complex systems*, Nat. Phys. **15** (2019), 313–320.
17. J. Borge-Holthoefer and A. Arenas, *Semantic networks: Structure and Dynamics*, Entropy **12** (2010), 1264–1302.
18. A. Criado-Alonso, E. Battaner-Moro, D. Aleja, M. Romance, and R. Criado, *Enriched line graph: A new structure for searching language collocations*, Chaos Solitons Fract. **142** (2021), no. 1, 110509.
19. H. F. De Arruda, L. D. F. Costa, and D. R. Amancio, *Using complex networks for text classification: Discriminating informative and imaginative documents*, EPL (Europhysics Letters) **113** (2016), no. 2, 28007.
20. H. F. De Arruda, S. Nascimento, V. Q. Marinho, D. R. Amancio, and L. D. F. Costa, *Representation of texts as complex networks: A mesoscopic approach*, J. Complex Netw. **6** (2018), no. 1, 125–144.
21. S. N. Dogorovtsev and J. F. F. Mendes, *Language as an evolving word web*, Proc. R. Soc. Lond. B **268** (2001), 2603–2606.
22. R. Ferrer i Cancho and R. V. Solé, *The small world of human language*, Proc. of the Royal Soc. of London B **286** (2001), 2261–2266.
23. R. F. i Cancho, O. Riordan, and B. Bollobás, *The consequences of Zipf's law for syntax and symbolic reference*, Proc. Biol. Sci/The Royal Society **272** (2005), no. 1562, 561–565.
24. H. Liu and F. Hu, *What role does syntax play in a language network?* EPL (Europhysics Letters) **83** (2008), 18002.
25. H. Liu, C. Xu, and J. Liang, *Dependency distance: A new perspective on syntactic patterns in natural languages*, Phys. Life Rev. **21** (2017), 171–193.
26. S. Martincic, D. Margan, and A. Mestrovic, *Multilayer network of language: A unified framework for structural analysis of linguistic subsystems*, Phys. Rev. E **74** (2016), 26102.
27. A. Mehler, A. Lücking, S. Banisch, P. Blanchard, and B. Frank-Job Eds., *Towards a theoretical framework for analyzing complex linguistics networks* Edited by A. Mehler, A. Lücking, S. Banisch, P. Blanchard, and B. Frank-Job, Springer-Verlag, 2016.
28. R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, *Language networks: Their structure, function, and evolution*, Complexity **15** (2010), no. 6, 20–26.
29. A. Criado-Alonso, E. Battaner-Moro, D. Aleja, M. Romance, and R. Criado, *Using complex networks to identify patterns in specialty mathematical language: A new approach*, Social Netw. Anal. Mining **10** (2020), no. 1, 1–10.
30. F. N. Silva, D. R. Amancio, M. Barsodova, L. da F. Costa, and O. N. Oliveira, *Using network science and text analytics to produce surveys in a scientific topic*, J. Infometrics **10** (2016), 487–502.
31. T. K. Landauer, P. W. Foltz, and D. Laham, *An introduction to latent semantic analysis*, Discourse Process. **25** (1998), no. 2-3, 259–284.
32. D. R. Amancio, M. G. Nunes, O. N. Oliveira Jr, and L. D. F. Costa, *Extractive summarization using complex networks and syntactic dependency*, Phys. A Stat. Mech. App. **391** (2012), no. 4, 1855–1864.
33. L. Antigueira, O. N. Oliveira Jr., L. F. Costa, and M. G. V. Nunes, *A complex network approach to text summarization*, Inform. Sci. **179** (2009), no. 5, 584–599.
34. D. R. Amancio, *A complex network approach to stylometry*, PLoS ONE **10** (2015), no. 8, e0136076.

35. D. R. Amancio, *Authorship recognition via fluctuation analysis of network topology and word intermittency*, *J. Stat. Mech.: Theory Experiments* **2015** (2015), no. 3, P03005.
36. X. Chen, P. Hao, R. Chandramouli, and K. Subbalakshmi, *Authorship similarity detection from email messages*, *Machine Learning and Data Mining in Pattern Recognition*, Springer, 2011, pp. 375–386.
37. A. Mehri, A. H. Darooneh, and A. Shariati, *The complex networks approach for authorship attribution of books*, *Phys. A* **391** (2012), no. 7, 2429–2437.
38. J. A. Danowski, B. Yan, and K. Riopelle, *A semantic network approach to measuring sentiment*, *Qual Quant* **55** (2021), 221–255.
39. L. Bowker and J. Pearson, *Working with specialized language: A practical guide to using corpora*, Routledge, 2002.
40. M. A. K. Halliday and C. M. I. M. Matthiessen, *Introduction to functional grammar*, 3rd ed., Routledge, Taylor & Francis Group, London and New York, 2004.
41. A. Criado-Alonso, D. Aleja, M. Romance, and R. Criado, *Derivative of a hypergraph as a tool for linguistic pattern analysis*, *Chaos Solitons Fractals* **163** (2022), 112604.
42. R. Criado, J. Flores, A. García del Amo, J. Gómez-Gardeñes, and M. Romance, *A mathematical model for networks with structures in the mesoscale*, *Int. J. Comput. Math.* **89** (2012), no. 3, 291–309.
43. M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, *Multilayer networks*, *J. Complex Netw.* **2** (2014), no. 3, 203–271.
44. C. Berge, *Hypergraphs*, Combinatorics of finite sets, North-Holland, 1989.
45. A. Bretto, *Hypergraph theory (an introduction)*, Mathematical Engineering, Springer International Publishing Switzerland, 2013.
46. R. Tyshkevich and V. E. Zverovich, *Line hypergraphs: A survey*, *Acta Applicandae Mathematicae* **52** (1998), 209–222.
47. R. J. Naik, *Intersection graphs of graphs and hypergraphs: A survey*. arXiv:1809.08472.
48. P. Jaccard, *Distribution de la flore alpine dans le bassin des dranses et dans quelques regions voisines*, *Bull. de la Société Vaudoise des Sciences Naturelles* **37** (1901), 241–272.
49. M. Brusco, J. D. Cradit, and D. A. Steinley, *Comparison of 71 binary similarity coefficients: The effect of base rates*, *PLoS One* **16** (2021), no. 4, e0247751.
50. L. D. F. Costa, *Further generalizations of the Jaccard index*. <https://www.researchgate.net/publication/355381945> (Online Accessed 21 August 2021).
51. L. D. F. Costa, *On the effects of text preprocessing on paragraph similarity networks*. <https://www.researchgate.net/publication/361553289> (Online Accessed 20 June 2022).
52. L. D. F. Costa, *On similarity*, *Phys. A* **599** (2022), 127456.
53. L. Costa and F. da, *Coincidence complex networks*, Vol. **3**, 2022, pp. 15012.
54. S. Talaga and A. Nowak, *Structural measures of similarity and complementarity in complex networks*, *Scientific Rep.* **12** (2022), 16580.
55. M. K. Vijaymeena and K. Kavitha, *A survey on similarity measures in text mining*, *Mach. Learn. Appl.* **3** (2016), no. 1, 1–28.
56. R. Criado, J. Flores, A. García del Amo, M. Romance, E. Barrena, and J. A. Mesa, *Line graphs for a multiplex network*, *Chaos* **26** (2016), 65309.
57. D. Aleja, R. Criado, A. García del Amo, Á. Pérez, and M. Romance, *Non-backtracking PageRank: From the classic model to Hashimoto matrices*, *Chaos, Solitons Fractals* **126** (2019), 283–2918.
58. P. Boldi, M. Santini, and S. Vigna, *PageRank: Functional dependencies*, *Inf. Syst.* **27** (2009), no. 4, 19–23.
59. S. Brin and L. Page, *The anatomy of a large-scale hypertextual Web search engine*, *Comput. Netw.* **30** (1998), 107.
60. S. Brin, L. Page, R. Motwani, and T. Winograd: *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab, 1999.
61. E. García, F. Pedroche, and M. Romance, *On the localization of the personalized PageRank of complex*, *Netw. Lin. Algebra Appl.* **439** (2013), 640–652.
62. A. Langville and C. Meyer, *Deeper inside PageRank*, *Internet Math.* **1** (2005), no. 3, 335–380.
63. M. G. Kendall, *A new measure of rank correlation*, *Biometrika* **30** (1938), no. 1-2, 81–89.
64. A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: The science of search engine ranks*, Princeton Univ. Press, 2006.
65. F. Pedroche, M. Romance, and R. Criado, *A bplex approach to PageRank centrality: From classic to multiplex networks*, *Chaos* **26** (2016), 65301.
66. R. Criado, S. Moral, Á. Pérez, and M. Romance, *On the edges' PageRank and line graphs*, *Chaos* **28** (2018), no. 7, 75503.

How to cite this article: Á. Criado-Alonso, D. Aleja, M. Romance, and R. Criado, *A new insight into linguistic pattern analysis based on multilayer hypergraphs for the automatic extraction of text summaries*, *Math. Meth. Appl. Sci.* (2023), 1–17. DOI 10.1002/mma.9201

APPENDIX A: LIST OF MOST RELEVANT PHRASES OF THE DATASET

The list of the most relevant sentences in the corpus under study is shown below. Their relative position can be seen in Figure 6. Bear in mind that commas and other punctuation symbols have been removed, and that the corpus has been treated as the data set on which we have applied the algorithms, so it cannot be considered as a “homogeneous” text.

- h*₁ This recommendation is based on one hand on the similarity between authors preference and the submitted query and on the other hand on the betweenness centrality of authors found on the search path.
- h*₂ It is worth adding that those driver agents do not possess any super power of any sort but they simply temporarily become informed leader as they happened to have discerned the danger first any other agent in the swarm could be driving the group as long as it is subjected to specific external cues which are not made available globally to the whole swarm.
- h*₃ The applied approach is a supervised machine learning approach where we attempt to learn a model for link formation based on a set of topological attributes describing both positive and negative example.
- h*₄ Positive systems are often found in the modeling of biology hydrology engineering and industrial process whose variables represent quantities that don't make sense unless they are nonnegative for example time in stochastic game algorithms money and goods in Leontief model data packets flowing in a network quantity of bacteria in a epidemiological model etc.
- h*₅ As an example, we apply this study to a multilayered network formed by two layers the social network of collaboration of the Spanish scientific community of statistical physics and the telecommunication network of each institution.
- h*₆ In this sense, when studying the factors involved in learning carried out by system associated with e-learning, the existence of numerous variables involved in the development of these learning and teaching process is observed such as educational level of students knowledge of ICT by students and lecturers places and moments for learning and teaching electronic devices used and personal and institutional situations.
- h*₇ After a careful review of dozens of review experiences over the last 15 years and applying the knowledge gained from the authors' direct involvement in the most relevant experiences, 5 additional criteria were identified to be included in the review methodology and software development, encompassing all non-software components of the review system.
- h*₈ Since the implemented solution is responsible for the safeguard of the properties inherent to democratic elections as well as voter rights, the authors consider that it is a safer option to wait until a sufficient budget has been allocated rather than starting a project without enough resource.
- h*₉ In this article, a remote electronic voting system is defined as a voting system used in a remote noncontrolled environment through electronic means in which the vote is sent partially or totally via an internet connection from a personal computer or mobile device which has not been specifically designed as a specialized electronic voting machine.
- h*₁₀ One of the main differences between this model and classical stage structured model is that in the current model we can alter the number of adults contributing to eggs production.
- h*₁₁ In this paper, we propose a random network to model the evolution of the academic performance focused on the educational level of bachillerato in Spain.
- h*₁₂ While the match between our theory and numerical simulation on synthetic network of both types is excellent Gleeson network may be parameterized to fit the degree distribution and clustering spectrum of real world network thus allowing us to achieve a significant improvement over standard non clustered theory in this case.
- h*₁₃ The steps of the methodology are the following network transformation into a weighted graph with trunk pipes segregation based on a factor and pipes weights community detection in the reduction graph through label propagation algorithm entrance definition to each sector based on an energy assessment.
- h*₁₄ We show numerically that the information of edge correlation between layer offers substantial insight in the complex multiplex structure thus contributing to simplified and faster analysis and design of multiplex graph with desired consensus or synchronization properties.
- h*₁₅ Although weak subcriticality sets the necessary condition for a freezing out dynamics as we crossover a secondary bifurcation, we find out an unexpected critical behavior in regard of the standard Kibble–Zurek prediction.

- h*₁₆ Although our system can work at a realtime frame rate for the considered video resolution, its tracking accuracy is affected by factors like the scene illumination condition, the contrast of the targets with respect to the background, the velocity of each target, and the frame rate of the video.
- h*₁₇ The properties of the network are compared with their corresponding spatial scale in order to derive allometric scaling laws.
- h*₁₈ While such an approach has been successfully applied in the context on simple network, different options can be applied to extend it to the multiplex network context.