Graph Mining under Linguistic Constraints for Exploring Large Texts

Solen Quiniou[1], Peggy Cellier[2], Thierry Charnois[3, 4], and Dominique Legallois[5]

[1] LINA, LUNAM Université de Nantes, Nantes, France

[2] IRISA, INSA de Rennes, Rennes, France

[3] GREYC, Université de Caen Basse-Normandie, Caen, France

[4] MoDyCO, Université Paris-Ouest Nanterre La Défense, Paris, France

[5] CRISCO, Université de Caen Basse-Normandie, Caen, France

solen.quiniou@univ-nantes.fr, peggy.cellier@irisa.fr,
thierry.charnois@unicaen.fr, dominique.legallois@unicaen.fr

Abstract. In this paper, we propose an approach to explore large texts by highlighting coherent sub-parts. The exploration method relies on a graph representation of the text according to Hoey's linguistic model which allows the selection and the binding of adjacent and non-adjacent sentences. The main contribution of our work consists in proposing a method based on both Hoey's linguistic model and a special graph mining technique, called CoHoP mining, to extract coherent sub-parts of the graph representation of the text. We have conducted some experiments on several English texts showing the interest of the proposed approach.

Keywords. Text coherence, graph representation, graph mining, Hoey's linguistic model.

Minería de grafos bajo restricciones lingüísticas para exploración de textos grandes

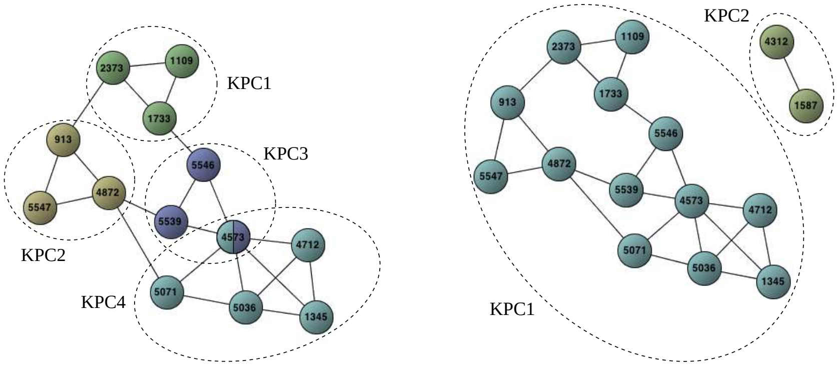
Resumen. En este artículo se propone el enfoque para la exploración de textos grandes destacando las sub-partes coherentes. El método de exploración se basa en la representación del texto mediante un gráfo de acuerdo con el modelo lingüístico de Hoey, el cual permite la selección y vinculación de frases adyacentes y no adyacentes. La principal aportación de este trabajo es la propuesta del método basado en el modelo lingüístico de Hoey por un lado y por otro lado en la técnica especial de minería de grafos llamada minería CoHoP, con el fin de extraer las sub-partes coherentes de la representación gráfica del texto. Se realizaron unos experimentos sobre varios textos en inglés mostrando el interés del enfoque propuesto.

Palabras clave. Coherencia de texto, representación con un grafo, minería de grafos, el modelo lingüístico de Hoey.

1 Introduction

Due to the availability of huge corpora, linguists, humanities scholars or other researchers can easily have access to large collections of texts in order to give a critical interpretation, or a discursive and textual analysis of them. However, such tasks are not easy to apply on large texts. For instance, linguists could want to discover new knowledge without knowing exactly what they are looking for. To do so, they analyze a text, and try to formulate and validate some assumptions. The main issue is the treatment of large texts. Indeed, in this case it is difficult to formulate and validate hypotheses by hand over the whole text. It is therefore crucial to design automatic methods to help the experts by highlighting some relevant and coherent parts of the texts. In addition, it could be useful to use some parameters to set the size of the visualized coherent parts so as to tune correlatively the granularity level of lexical cohesion in the textual parts.

On the one hand, visualization, automatic summarization, and clustering techniques can help the linguists to explore, or analyze large texts. Visualization tools can allow a user to explore a text collection by highlighting frequent textual patterns within the collection [4]. Summarization



(a) $k = 3$ **(b)** $k = 2$

Fig. 1. CoHoPs extracted from the same two attributes, for two values of k

approaches aim at producing a reduced text made up of salient sentences either selected or generalized from the original text [8]. Although visualization and summarization techniques allow to pinpoint the relevant sentences of a text, they do not provide a view of the relations between the sentences which can be interesting to analyze a text. Clustering is a well-known technique used in the field of text mining [5] to automatically group similar objects (*e.g.*, sentences) that share some similarities (*e.g.*, topics). The drawback of such approaches is that each sentence belongs to one and only one cluster although some sentences may refer to several topics. Nevertheless, clustering offers a good baseline for evaluating our approach (see Section 5.2). On the other hand, computational linguistic models like the ones based on the Rhetorical Structure Theory (RST) [11] aim at identifying elementary discourse units (*e.g.*, sentences, clauses) and relations between them. However, these relations only hold between adjacent units.

A linguistic model to analyze non-narrative texts based on lexical repetitions, the Hoey model, is presented in [6]. The approach highlights the organization of the text (development of a text, conceptual content), by revealing the binding of adjacent and non-adjacent sentences. This approach is interesting for several tasks, like retrieving a logical reasoning about a specific subject in a text, studying the lexical cohesion of a text [9], or summarizing a text [14]. Whereas this approach is hard to apply by hand on large texts, few works are based on a computational implementation of the Hoey model [9, 14]. The main drawback of these implementations is that the sentence networks thus built are very large. Therefore, it is difficult to display the whole networks in a user-friendly way.

In this paper, we propose an approach to automatically extract, from a text, subsets of sentences that are coherent from a lexical point of view. Furthermore, the subsets are represented by graphs which offer a view of the relationships between the sentences. In addition, the size of those sentence subsets is manageable for linguists to analyze them. The main contribution of our work consists in proposing a method based on both an implementation of Hoey linguistic model to represent the text as a graph and a special graph mining technique to extract coherent sub-parts of this graph. Graph mining has gained an increasing interest in the field of data mining for discovering new knowledge [16]. In this paper, we focus on the mining of a certain type of patterns called *collections of homogeneous k-clique percolated components* (CoHoPs) [12]. We use them to extract homogeneous parts of sentence networks. Moreover, some constraints can be set to mine the graphs which makes it possible to vary the size of the sub-graphs and their degree of coherence. To our knowledge, this graph mining technique has never been used in the field of natural language processing. In our approach, the mining is said to be done “under linguistic constraints” because the original structure of the graph is built according to Hoey’s model.

The rest of the paper is organized as follows. Section 2 introduces the Hoey linguistic model and Section 3 presents the used graph mining technique. Then, our approach based on mining sentence networks under linguistic constraints is described in Section 4. Finally, Section 5 reports some experimental results.

2 Hoey's Linguistic Model

Based on lexical repetitions, the main idea of the Hoey model [6] is the identification of sentences sharing at least three lexical units. A *lexical repetition* can be the strict repetition of the lexical unit (*e.g.*, brain/brain) but also lexical units that share the same lemma or the same stem (*e.g.*, produce/production), a synonymy relation (*e.g.*, buy/purchase), etc. When two sentences share at least three lexical units, the pair of sentences is *bounded*. A set of at least three sentences such that each sentence is bounded directly or indirectly with all the other sentences of the set is called a *sentence network*. Figures 6a and 6b show excerpts of sentence networks. In these examples, the lexical repetition is only based on shared lemmas. It is interesting to note that the distance between the sentences can be really high (the position of the sentence in the text is given in square brackets at its beginning). The set of sentence networks of a text is called the *hypotext*. Note that unbounded sentences do not appear in the hypotext.

The Hoey linguistic model is useful to represent a text so as to analyze its lexical cohesion. However, the main drawback of the Hoey-based approaches is that the sentence networks thus built are too wide to be entirely displayed which make tedious the analysis of large texts. That is why, we need a method to extract homogeneous parts of the sentence networks so as to ease the analysis of the networks. For that purpose, we introduce the CoHoP mining approach.

3 Graph Mining: CoHoP Patterns

A CoHoP mining algorithm, as the one proposed by [12], allows the extraction of CoHoP patterns from boolean attributed graphs. A CoHoP can be seen as a set of communities where the elements share similar properties: a community corresponds to what is called a *k-clique percolated component* (k-PC).

3.1 k-clique Percolated Components (k-PCs)

In a graph, a *k-clique* is a set of k vertices in which every pair of distinct vertices is connected by an edge. A *k-clique percolated component* (k-PC) is a relaxed version of the concept of cliques. A k-PC was defined by [3] as the union of all the k-cliques connected by overlaps of $k-1$ vertices. Therefore, in a k-PC, each k-clique can be reached from any other k-clique through a series of adjacent k-cliques and each vertex of a k-PC can be reached from any other vertex through well connected subsets of vertices (the k-cliques).

In Figure 1a, there are 4 k-PCs: {913, 4872, 5547}, {1109, 1733, 2373}, {4573, 5539, 5546}, and {1345, 4573, 4712, 5036, 5077}. The first three k-PCs only contain one 3-clique whereas the last k-PC contains five overlapping 3-cliques: {1345, 4573, 4712}, {1345, 4573, 5036}, {1345, 4712, 5036}, {4573, 4712, 5036}, and {4573, 5036, 5071} (with $k = 3$, the overlaps of 3-cliques contain two vertices). Note that a clique is contained in at most one k-PC but a vertex can be part of several k-PCs as it can belong to several k-cliques.

3.2 Collections of Homogeneous k-PCs (CoHoPs)

A *collection of homogeneous k-PCs* (CoHoPs) was defined by [12] as a set of vertices such that, with k, α, and γ being positive integers defined by users:

— all vertices are *homogeneous*, *i.e.* they share at least α true-valued attributes,

— the collection contains at least γ k-PCs,

— and all k-PCs showing the same true-valued attributes are in the collection.

Figure 1a illustrates such a CoHoP extracted from a set of two attributes $\{a_1, a_2\}$ and containing four k-PCs ($\alpha = 2, k = 3, \gamma = 4$). Note that, as opposed to the computation of the k-PCs, the extraction of the CoHoPs is done from the sets of attributes of the vertices. In Figure 1a, the sets of attributes of the vertices are not displayed (in order not to overload the figure) but each vertex, V_i, is labeled with a set of attributes, A_i, that contains at least a_1 and a_2.

Therefore, parameter α allows the setting of the minimum number of attributes needed to be shared by the vertices of the extracted CoHoPs, whereas γ allows the setting of the minimum number of

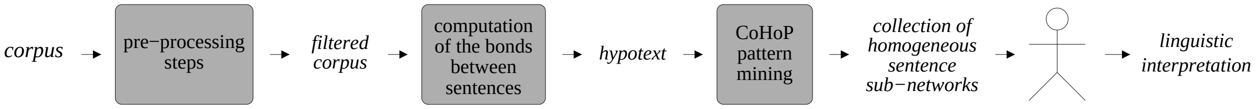

corpus
pre−processing steps
filtered corpus
computation of the bonds between sentences
hypotext
CoHoP pattern mining
collection of homogeneous sentence sub−networks
linguistic interpretation

Fig. 2. Overview of our approach

k-PCs in the CoHoPs. Parameter k has an important impact on the structure of the extracted CoHoPs. Indeed, increasing it also increases the coherence that need to have the vertices belonging to the same k-PC. Figure 1b represents the CoHoP extracted from the same set of attributes as in Figure 1a but when choosing $k = 2$. This CoHoP now contains 15 vertices distributed in only two k-PCs, the biggest one corresponding to the four k-PCs of Figure 1a. Thus, choosing the value of k allows setting the wanted level of cohesion between the vertices of each k-PC. Indeed, the vertices need to be more strongly bounded when increasing the value of k.

4 Methodology

In this section, we propose a new approach to extract coherent parts of sentence networks: it is based on both the Hoey linguistic model and the extraction of CoHoP patterns. Figure 2 illustrates the various steps of the approach that are presented in greater details in the following sub-sections.

4.1 Pre-Processing and Construction of the Hypotext

First, the text is POS-tagged using TreeTagger [15] and split into sentences at punctuation marks of the following set: {“.”, “?”, “!", “:“}. The sentences are then filtered so as to keep only their relevant lexical units. In our case, it consists in keeping their lexemes (nouns, adjectives, adverbs, and verbs except auxiliaries). Actually, we consider the lemmas of these lexemes. Therefore, each sentence of the filtered text is represented by its lexeme lemmas. For example, the sentence “*Online emotional experiences may be compared to receiving a salary without earning it by hard work.*” is represented by the set {*online, emotional, experience, compare, receive, salary, earn, hard, work*}.

From the filtered text, we build its graph representation (hypotext) by applying the Hoey linguistic model. To create the hypotext as defined in Section 2, we bound all pairs of sentences that share at least three lexeme lemmas. Note that unbounded sentences do not appear in the hypotext.

4.2 Mining Sentence Networks Under Linguistic Constraints

The goal of this final step is to extract homogeneous parts of the hypotext created as described previously. The hypotext can be seen as an attributed graph where each vertex represents a sentence and each edge represents a bond between two sentences that share at least three lexical units. Furthermore, the set of lexical units of a sentence is associated as a set of attributes to its corresponding vertex. With this representation of the hypotext as an attributed graph, we can use CoHoP mining algorithms, as presented in Section 3. In our approach, the mining is said to be done “under linguistic constraints” because the original graph is built according to the Hoey linguistic model. Moreover, the set of attributes labeling a vertex corresponds to the lexical units of the underlying sentence.

Each extracted CoHoP pattern corresponds to what we call a *collection of homogeneous sentence sub-networks* (CoHoSS). In the same way a CoHoP is made up of homogeneous k-PCs (*i.e.*, sets of vertices that share the same set of attributes), a CoHoSS is made up of homogeneous sentence sub-networks (*i.e.*, sets of sentences that share the same set of lexical units). Each sentence sub-network corresponds to the definition of a k-PC. Thus, in a sentence sub-network, each sentence is either directly bounded by an edge to the other sentences of the sub-network (if they share at least three lexical units), or indirectly reachable from any other sentence through well connected subsets of sentences (each subset corresponds to a k-clique, as defined in Section 3.1). Therefore, CoHoSSs represent collections of sub-networks of the overall sentence network that have a certain lexical cohesion w.r.t.

Table 1. Quantitative results on the hypotext construction

Corpus	*Speech*	*Love*
#Sentences	5 308	5 571
#Words	127 563	112 325
#Total lexemes	59 657	53 035
#Bonds	50 277	131 497
#Sentence networks	2	2
%Sentences in the hypotext	75.6 %	79.0 %

the considered set of lexical units from which they are extracted. The structure of the CoHoSSs can then be analyzed by linguists, for example to interpret each of the sub-network and the way they are connected.

5 Experimental Results

In this section, we report two sets of experiments on the implementation of Hoey's model and more particularly on the extraction of CoHoSS patterns. The first experiment is done on two large English texts (see Section 5.1) and the second experiment is done on a short scientific paper (see Section 5.2).

5.1 Mining Sentence Networks from Large Texts

5.1.1 Settings: Data and Tools

First, to evaluate our proposed approach, we chose two large corpora, each one corresponding to an expositive English text: "*The Origin of Speech*" [10] (denoted *Speech*) and "*Love Online: Emotions on the Internet*" [2] (denoted *Love*). These texts contain 416 and 302 pages, respectively. Note that, after the pre-processing steps presented in Section 4.1, each sentence of the texts is represented by its corresponding set of filtered lexical units.

In order to extract the CoHoPs as presented in Section 4.2, we use *CoHoP Miner* [12]. It allows the extraction of CoHoPs by setting the various parameters of the mining process (k, α, γ).

5.1.2 Quantitative Results on Applying Hoey's Linguistic Model

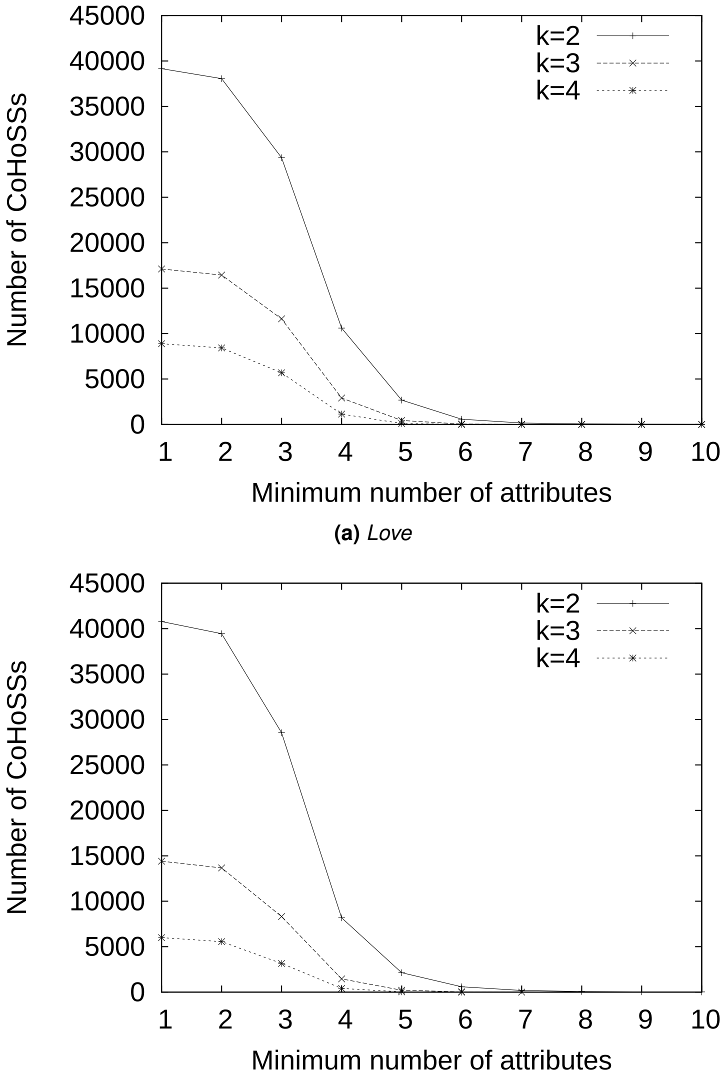
The quantitative results on the hypotext created to represent each considered corpus are summarized in Table 1. We can note that a sentence contains on average 10 lexemes for *Speech* and 11 lexemes for *Love* whereas it contains on average 24 words for *Speech* and 20 words for *Love*. Therefore, representing sentences by their lexemes allows a reduction of the number of attributes describing a sentence without losing meaningful information. Furthermore, the hypotexts are very large w.r.t. the number of sentences: more than 75%. It suggests a strong lexical cohesion in the texts (each sentence of the hypotext is bounded on average with 13 sentences for *Speech* and with 30 sentences for *Love*). We can note that, for each corpus, few sentence networks, only two, are found: a very small sentence network with very few sentences and a very large one. The analysis of the large network is not manageable by hand and therefore requires methods to extract coherent sub-parts from this network as proposed by our approach.

5.1.3 Quantitative Results on the Extracted CoHoSSs

The number of extracted CoHoSSs using *CoHoP Miner* depends on the value of the parameters k, α and γ. The value of γ allows to choose the minimum number of sub-sentence networks that compose the CoHoSSs (see Section 3.2). In the experiments, we set γ to 1 *i.e.* we do not limit the number of sub-sentence networks in the CoHoSSs.

Figures 3a and 3b plot the number of extracted CoHoSSs for various values of k w.r.t. the minimum number of attributes, for both corpora. Each point of the curves corresponds to the number of CoHoSSs extracted from at least α attributes. For example, in Figure 3a, with $k = 3$, 11 624 CoHoSSs were extracted from a set of at least 3 attributes. We can see that the majority of the CoHoSSs are based on 1 to 6 attributes. Furthermore, most of the CoHoSSs are based on at most 4 attributes. It means that the topics in the CoHoSSs are expressed by less than 4 lexical units. The behaviour of the curves is the same on both corpora and for the various values of k.

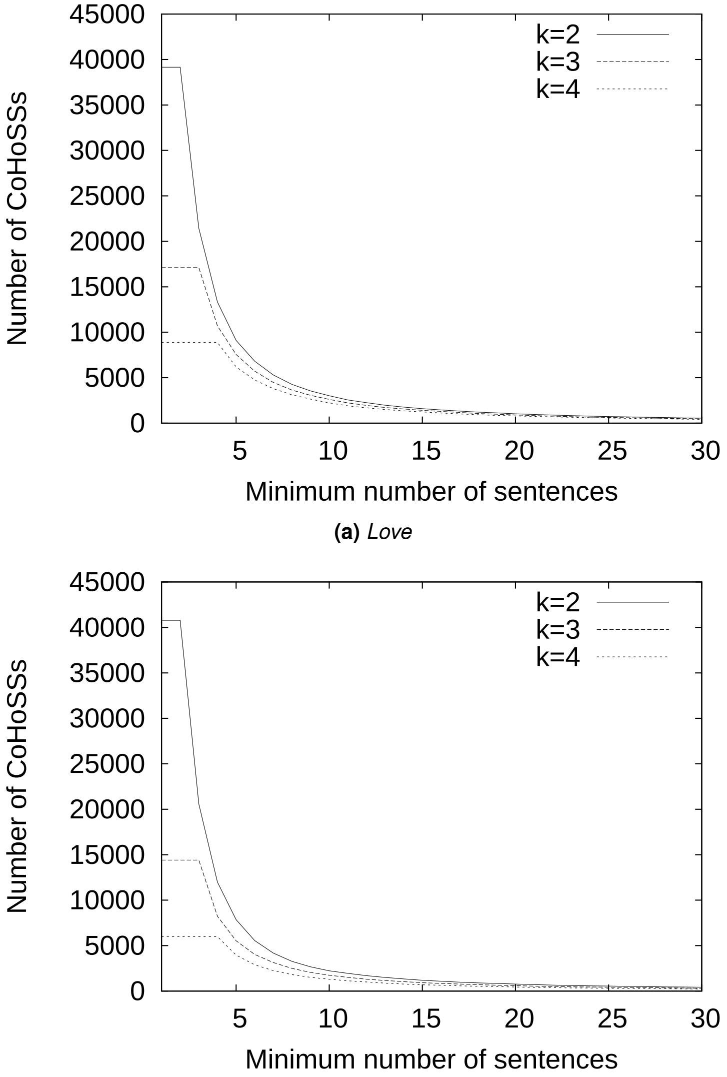
Figures 4a and 4b plot the number of extracted CoHoSSs for various values of k (from 2 to 4)



(a) *Love*

(b) *Speech*

Fig. 3. Number of extracted CoHoSSs w.r.t. attributes



(a) *Love*

(b) *Speech*

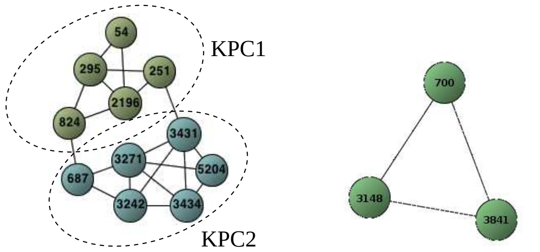
Fig. 4. Number of extracted CoHoSSs w.r.t. sentences

w.r.t. the minimum number of sentences, n, that belong to them, for both corpora. Each point of the curves corresponds to the number of CoHoSSs that contain at least n sentences. For example, in Figure 4a, 7 559 CoHoSSs contain at least 5 sentences, for $k = 3$. We can see that the majority of the CoHoSSs contain at most 20 sentences. Furthermore, most of the CoHoSSs contain less than 10 sentences. It means that we extract a lot of small sets of sentences that are thus easier to analyze from a linguistic point of view than the whole hypotext. The behaviour of the curves is also the same on both corpora and for the various values of k. Moreover, the CoHoSSs that contain a lot of sentences are actually based on a single attribute which is a lexical unit with a general meaning relatively to the considered corpus. For example, with $k = 3$, for the text "*The Origin of Speech*", the CoHoSS from the word *"speech"* contains 608 sentences whereas the CoHoSS from the word *"origin"* contains 590 sentences.

Finally, we can see that the number of extracted CoHoSSs decreases when the value of k increases. This is because k sets the granularity level of lexical cohesion for the sub-networks in the CoHoSSs (see Section 3.2). When k increases, the level of lexical cohesion increases too, which limits the number of extracted CoHoSSs. In conclusion, the value of k may be chosen according to the granularity level of lexical cohesion needed in the CoHoSSs. Furthermore, the value of γ may be chosen to limit the total number of extracted CoHoSSs by selecting the largest ones. Finally, the setting of α allows to focus the linguistic analysis on bounded sentences that share at least a minimum number of lexical units.

5.1.4 Examples of Extracted CoHoSS and Linguistic Interpretation

Figure 5a illustrates the first considered CoHoSS, extracted from the text "*The Origin of Speech*", and Figure 6a gives the corresponding sentences of the text. The CoHoSS was extracted from the attribute "*adaptation*", using the following values for the mining parameters: $k = 3, \alpha = 1, \gamma = 1$. It is made up of two sub-networks. The first sub-network (KPC_1) deals with the general topic of the CoHoSS, *i.e.* the phenomenon of adaptation. This sub-network is relatively coherent whereas the distance between its sentences is very high (corresponding to a span of more than 2 000 sentences in the text). The second sub-network (KPC_2) develops a more specific topic of adaptation: the specialization of the left-hemispheric. This sub-network starts with sentence 687 which is connected to the prior sub-network by sentence 824. We can see that the span of the CoHoSS is very large since the CoHoSS starts at sentence 54 and ends at sentence 5204. This interesting property of sentence non-contiguousness in the sentence networks can therefore be seen in the CoHoSSs extracted from the networks but also in the sub-networks of the CoHoSSs.



(a) {*adaptation*}

(b) {*person, further, develop, relationship*}

Fig. 5. CoHoSSs extracted from given attributes ($k = 3$)

A second example of extracted CoHoSS, from "*Love Online: Emotions on the Internet*", is illustrated by Figure 5b (Figure 6b gives its corresponding sentences). The CoHoSS was extracted from the attributes "*person, further, develop, relationship*" using the following values for the mining parameters: $k = 3, \alpha = 4, \gamma = 1$. It highlights the three main stages of a relationship between two persons: the keen attention to the signals conveyed by the other person; the development of the relationship after the first face-to-face meeting; the principle of reality when the two partners know each other better.

5.2 Mining Sentence Networks from a Scientific Paper

5.2.1 Settings: Data and Experimental Protocol

To show the interest of our approach based on the extraction of CoHoSSs, we evaluate the coherence of the CoHoSSs w.r.t. a baseline clustering method. Because we have to evaluate by hand the coherence of all the extracted CoHoSSs, it would be too tedious to do so on one of the corpora used in Section 5.1 since too many CoHoSSs

[54] I take the standpoint of an **evolutionary**$_1$ biologist who, according to Mayr (1982), "studies the forces that bring about changes in faunas and floras ... [and] studies the steps by which have **evolved**$_2$ the miraculous **adaptations**$_3$ so characteristic of every aspect of the organic world" (pp.69 – 70).

[251] An important connotation of the tinkering metaphor, for Jacob, is that **adaptations**$_3$ exploit whatever is available in order to respond successfully to selection pressures, whether or **not**$_4$ they originally **evolved**$_2$ for the use they're now put to.

[295] "**language**$_5$ cannot be as novel as it seems, for **evolutionary**$_1$ **adaptation**$_3$ does **not**$_4$ **evolve**$_2$ out of the blue" (p.7).

[824] Indeed, the same claim about the genes could be made for organisms without **language**$_5$ and culture, because the **evolutionary**$_1$ process **involves**$_2$ **adaptation**$_3$ to a particular niche.

[2196] "**language**$_5$ cannot be as novel as it seems, for **evolutionary**$_1$ **adaptations**$_3$ do **not**$_4$ **evolve**$_2$ out of the blue" (Bickerton, 1990, p.7).

[687] In my **view**$_{15}$, **speech**$_1$ is an **adaptation**$_2$ that made the rich message-sending **capacity**$_3$ of spoken **language**$_4$ possible.

[3242] The most prevalent **view**$_{15}$ of the **origin**$_5$ of the **hand**$_{16}$ – mouth relationship in the latter part of the last century was that the **adaptation**$_2$ in tool use which occurred in **Homo**$_6$ **habilis**$_7$ about 2 million years ago led to a **left-hemispheric**$_8$ specialization for manual " praxis " (basically motor skill) and that the first **language**$_4$ was a gestural **language**$_4$ built on this basis.

[3271] This led to the **conclusion**$_{14}$ that the **origin**$_5$ of the human **left-hemispheric**$_8$ praxic specialization, commonly thought to be a basis for the **left-hemisphere**$_9$ **speech**$_1$ **capacity**$_3$, cannot be attributed to the tool-use **adaptation**$_2$ in **Homo**$_6$ **habilis**$_7$ (MacNeilage, in press).

[3431] One implication of the **origin**$_5$ of a **left-hemisphere**$_9$ routine-action-control **specialization**$_{10}$ in early vertebrates is that this already-existing **left-hemisphere**$_9$ action **specialization**$_{10}$ may have been put to use in the form of the right-side dominance associated with the clinging and leaping motor **adaptation**$_2$ characteristic of everyday early **prosimian**$_{13}$ life.

[3434] If so, then the **left-hemisphere**$_9$ action-control **capacity**$_3$ favoring right-sided **postural**$_{11}$ support may have triggered the asymmetric reaching **adaptation**$_2$ favoring the **hand**$_{16}$ on the side less dominant for postural support – the left **hand**$_{16}$ – before the manual-predation **specialization**$_{10}$ in vertical clingers and leapers, and its accompanying ballistic reaching **capacity**$_3$, **evolved**$_{12}$.

[5204] As evidence for the highly specialized nature of this emergent **adaptation**$_2$, he cites the **conclusion**$_{14}$ of the **postural**$_{11}$ **origins**$_5$ theory that left-**hand**$_{16}$ preferences for prehension **evolved**$_{12}$ in **prosimians**$_{13}$ (see Chapter 10).

(a) {*adaptation*}

[700] However, the online lover, lacking many types of sensory information, must be sensitive to every signal conveyed by the other **person**$_1$ – otherwise, their **relationship**$_4$ cannot **develop**$_3$ **further**$_2$.

[3148] When there is no significant discrepancy between the imagined partner and the one **revealed**$_5$ in the first face-to-face meeting, there is a good chance that the **relationship**$_4$ will **develop**$_3$ **further**$_2$, as each **person**$_1$ already has a positive attitude toward the other.

[3841] As the **relationship**$_4$ **develops**$_3$ **further**$_2$, more negative aspects about the **person**$_1$ will be **revealed**$_5$, thus making this **person**$_1$ more real.

(b) {*person, further, develop, relationship*}

Fig. 6. Corresponding sentences for the CoHoSSs of Figure 5

were extracted. That is why we chose to do this evaluation on one of our scientific papers [13]. This paper contains 12 pages and 188 sentences that were pre-processed as presented in Section 4.1. In addition, each sentence is actually represented by the corresponding set of its filtered lexical units that are used to build the hypotext: the total number of filtered lexical units is 498.

As a baseline clustering method, we used a k-means clustering with a cosine distance. Each sentence is represented by a vector of 498 elements, each element being set to 1 or 0 depending on whether the sentence contains or not the corresponding lexical unit. To extract the clusters, we used Elki [1] with the kMeansLloyd algorithm and the cosine distance. To set the value of k (the number of clusters) we chose empirically the value so as to maximize the number of clusters that contain between 3 to 10 sentences. Indeed, in the rest of the evaluation, we will only consider clusters and CoHoSSs that contain 3 to 10 sentences. These values were chosen because assessing the coherence of very small clusters or CoHoSSs (containing 2 sentences) is not interesting and it is difficult to obtain quite large coherent clusters or CoHoSSs (the upper bound of 10 sentences represents clusters or CoHoSSs containing 5% of the sentences of the text). Therefore, the value of k (the number of clusters) is set to 60: 38 of the 60 clusters contain 3 to 10 sentences. To extract the CoHoSSs, we used *CoHoP Miner* with the following settings: $k = 3, \alpha = 1, \gamma = 1$. Out of the 509 extracted CoHoSSs, 457 contain 3 to 10 sentences: only the latter CoHoSSs will be used for the evaluation.

Table 2. Mean and standard deviation of the scores

Judge	Shared CoHoSSs	All CoHoSSs	Clusters
J_1	2.5 ± 0.7	2.6 ± 0.6	2.2 ± 0.8
J_2	2.3 ± 0.8	2.3 ± 0.8	1.8 ± 0.8
J_3	2.4 ± 0.7	2.4 ± 0.7	1.7 ± 0.8
All judges	2.4 ± 0.6	2.4 ± 0.7	1.8 ± 0.7

For the evaluation, CoHoSSs and clusters are presented to three judges: the 38 clusters, 50 CoHoSSs shared by the three judges (randomly selected among the 457 CoHoSSs), and 135-137 CoHoSSs owned only by each judge (randomly selected among the remaining 407 CoHoSSs). In order to perform a blind evaluation, we randomly mix the clusters and the CoHoSSs presented to each judge. Therefore, each judge has to evaluate the coherence of 223-225 lists of sentences without knowing whether the lists correspond to CoHoSSs or clusters. Note that the sentences in the lists are

Table 3. Distribution of the scores, s, attributed to CoHoSSs and clusters

Judge	$1 \leq s < 2$		$2 \leq s < 3$		$s = 3$	
	CoHoSSs	Clusters	CoHoSSs	Clusters	CoHoSSs	Clusters
J_1	6.5%	28.9%	27.6%	26.3%	65.9%	44.7%
J_2	21.1%	44.7%	29.7%	28.9%	49.2%	26.3%
J_3	13.9%	55.3%	30.5%	23.7%	55.6%	21.1%
All judges	13.6%	47.4%	32.2%	42.1%	54.3%	10.5%

ordered according to their position in the text. As the evaluation, the judges were asked to determine the coherence of the lists of sentences on a scale from 1 to 3 (a score of respectively 1, 2, and 3 means that respectively 0-33%, 33-75%, and 75-100% of the sentences belonging to a cluster or a CoHoSS are considered coherent). The following definition given in [7] is used to assess the *coherence*: *“A paragraph is coherent when the information in successive sentences follows some pattern of inference or of knowledge with which the hearer is familiar. To signal such inferences, speakers usually use relations that link successive sentences in fixed ways”*.

5.2.2 Human Evaluation of the CoHoSSs w.r.t. Clusters

Table 2 gives the mean and the standard deviation of the scores given to the CoHoSSs and clusters by each judge as well as by all of them. In the latter case, the score of each CoHoSS or cluster is either the score given by one judge (if it was only evaluated by one judge) or the mean of the scores given by the three judges. We can see that a better mean score is given to the CoHoSSs. Thus, the lists of sentences obtained through the CoHoP mining process are judged more coherent than the ones obtained with a baseline clustering algorithm.

Table 3 gives the distribution of the scores attributed to the CoHoSSs and clusters by each judge as well as by all of the judges. When considering the scores of all the judges, we can see that more than half of the CoHoSSs were given the highest score of 3 whereas a little less than half the clusters were given the lowest score of 1. Furthermore, as the total number of CoHoSSs is higher than the total number of clusters, the CoHoP mining process extracts more coherent CoHoSSs that could be analyzed from a linguistic point of view. Indeed, in absolute values, 248 CoHoSSs are coherent whereas only 4 clusters are coherent.

This manual evaluation of the coherence of CoHoSSs showed the interest of our proposed approach w.r.t. a state of the art clustering method to extract coherent sets of sentences from a text. Another advantage of our approach is that we do not need to set the number of CoHoSSs to extract whereas the number of clusters to create has to be set. Furthermore, in a clustering method, each sentence of the text is assigned to one and only one cluster whereas some sentences may not be informative and some of them may be associated to several lists of sentences. Hence the advantage of extracting CoHoSSs where a sentence may belong to several CoHoSSs or to no CoHoSS at all.

6 Conclusion

In this paper, we have proposed an automatic approach to explore large texts based on both a linguistic model (Hoey’s model) to represent the text as a graph and a graph mining method (CoHoP pattern mining) to extract relevant parts of it. The method allows to discover subsets of sentences (aka collections of homogeneous sentence sub-networks) that are coherent from a lexical point of view. The advantages are twofold. First, the graph representation offers a view of the relationships between the sentences. Second, graph mining techniques allows the scalability of Hoey’s linguistic model. In particular, tuning the parameters allows selecting relevant parts of the sentence network representing the text and refining the needed granularity level of the extracted collection of homogeneous sentence sub-networks. In linguistic terms, it highlights the lexical cohesion of the extracted sentences. We have conducted some experiments on two large English corpora to validate this approach. We have also compared our approach to a state of the art clustering method on a short scientific text.

Acknowledgments

This work is partly supported by the French Région Basse-Normandie and by the ANR (French National Research Agency) funded project Hybride ANR-11-BS02-002. The authors would also like to thank Pierre-Nicolas Mougel and Christophe Rigotti (LIRIS, Lyon) for the availability of *CoHoP Miner*.

References

1. **Achtert, E., Goldhofer, S., Kriegel, H.-P., Schubert, E., & Zimek, A.** (**2012**). Evaluation of clusterings – metrics and visual support. In *Proc. of ICDE'12*.
2. **Ben-Ze'ev, A.** (**2004**). *Love Online: Emotions on the Internet*. Cambridge Univ. Pr.
3. **Derenyi, I., Palla, G., & Vicsek, T.** (**2005**). Clique percolation in random networks. *Physical Review Letters*, 94.
4. **Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B., & Plaisant, C.** (**2007**). Discovering interesting usage patterns in text collections: integrating text mining with visualization. In *Proc. of CIKM'07*.
5. **Feldman, R. & Sanger, J.** (**2006**). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge Univ. Pr.
6. **Hoey, M.** (**1991**). *Patterns of Lexis in Text.* Describing English Language. Oxford Univ. Pr.
7. **Hovy, E.** (**1988**). Planning coherent multisentential text. In *Proc. of ACL'88*.
8. **Jones, K. S.** (**2007**). Automatic summarising: The state of the art. *Information Processing & Management*, 43(6).
9. **Legallois, D., Cellier, P., & Charnois, T.** (**2011**). Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle. In *Actes de TALN'11*.
10. **MacNeilage, P.** (**2008**). *The Origin of Speech*. UOP Oxford.
11. **Mann, W. & Thompson, S.** (**1988**). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3).
12. **Mougel, P.-N., Rigotti, C., & Gandrillon, O.** (**2012**). Finding collections of k-clique percolated components in attributed graphs. In *Proc. of PAKDD'12*.
13. **Quiniou, S., Cellier, P., Charnois, T., & Legallois, D.** (**2012**). What about sequential data mining techniques to identify linguistic patterns for stylistics? In *Proc. of CICLing'12*.
14. **Renouf, A. & Kehoe, A.** (**2004**). Textual Distraction as a Basis for Evaluating Automatic Summarisers. In *Proc. of LREC'04*.
15. **Schmid, H.** (**1994**). Probabilistic part-of-speech tagging using decision trees. In *Proc. of KDD'94*.
16. **Washio, T. & Motoda, H.** (**2003**). State of the art of graph-based data mining. *SIGKDD Explorations*, 5(1).

Solen Quiniou is an Assistant Professor in the department of Computer Science of the IUT at the University of Nantes, France. She holds a PhD degree in Computer Science from the INSA of Rennes, France. Her research interest in the LINA laboratory include natural language processing, data mining, knowledge discovery and multimodal applications.

Peggy Cellier is an Associate Professor in the Computer Science Department at INSA (Engineer School), in Rennes, France. She holds a PhD degree in Computer Science from the University of Rennes 1, France. Her research interests include software engineering, natural language processing, knowledge discovery and data mining.

Thierry Charnois is an Assistant Professor at IUT Caen and at the GREYC Laboratory (CNRS UMR 6072), University of Caen, France. He holds a PhD in Computer Science (1999) from LIPN Laboratory, University of Paris 13, and a Habilitation degree in 2011 from the University of Caen. His research interests include natural language processing, knowledge discovery, information extraction from texts, and their applications, notably in bioinformatics, digital humanities or opinion analysis.

Dominique Legallois holds a PD (habilitation) and a PhD in linguistics from the University of Caen. He is currently a full-time associate Professor in the Department of language and linguistics where he teaches both graduate and undergraduate courses in French grammar, discourse analysis and corpus linguistics. His research interests include text linguistics (lexical repetition and textual coherence), pragma-semantics of grammar and collocation frameworks in corpora. He is also Assistant Director of the CEMU (distance learning) of the University of Caen.

Article received on 07/12/2012; accepted on 11/01/2013.