Fouille de graphes sous contraintes linguistiques pour l'exploration de grands textes

Solen Quiniou[1, 2] Peggy Cellier[3] Thierry Charnois[1] Dominique Legallois[2]
(1) GREYC Université de Caen Basse-Normandie, Campus 2, 14000 Caen
(2) CRISCO Université de Caen Basse-Normandie, Campus 1, 14000 Caen
(3) IRISA-INSA de Rennes, Campus de Beaulieu, 35042 Rennes Cedex
solen.quiniou@unicaen.fr, peggy.cellier@irisa.fr,
thierry.charnois@unicaen.fr, dominique.legallois@unicaen.fr

RÉSUMÉ

Dans cet article, nous proposons une approche pour explorer des textes de taille importante en mettant en évidence des sous-parties cohérentes. Cette méthode d'exploration s'appuie sur une représentation en graphe du texte, en utilisant le modèle linguistique de Hoey pour sélectionner et apparier les phrases dans le graphe. Notre contribution porte sur l'utilisation de techniques de fouille de graphes sous contraintes pour extraire des sous-parties pertinentes du texte (c'est-à-dire des collections de sous-réseaux phrastiques homogènes). Nous avons réalisé des expérimentations sur deux textes anglais de taille conséquente pour montrer l'intérêt de l'approche que nous proposons.

ABSTRACT

Graph Mining Under Linguistic Constraints to Explore Large Texts

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 the Hoey linguistic model which allows the selection and the binding of sentences in the graph. Our contribution relates to using graph mining techniques under constraints to extract relevant sub-parts of the text (*i.e.*, collections of homogeneous sentence sub-networks). We have conducted some experiments on two large English texts to show the interest of the proposed approach.

MOTS-CLÉS : Fouille de graphes, réseaux phrastiques, analyse textuelle, navigation textuelle.

KEYWORDS: Graph Mining, sentence networks, textual analysis, textual navigation.

1 Introduction

L'interprétation critique des textes et l'analyse textuelle et discursive ont été renouvelées ces dernières années grâce à la numérisation de nombreux textes. Cependant, ce renouvellement s'accompagne de difficultés, notamment techniques : la numérisation ne suffit pas en elle-même, l'investigation des textes doit s'appuyer sur des méthodes et outils offrant à la fois une visualisation et une navigation pertinentes dans les textes. Les chercheurs peuvent ainsi par exemple focaliser leur analyse sur des thématiques particulières. La nécessité de tels méthodes et outils est d'autant plus forte que les textes sont généralement de taille conséquente. Deux types d'approches peuvent aider les linguistes dans des tâches d'exploration ou d'analyse de textes : les

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 253–266,
Grenoble, 4 au 8 juin 2012.

méthodes de résumé automatique et les techniques de visualisation de collections de textes.

D'un côté, les méthodes de résumé automatique visent à produire une vue contiguë du texte sous la forme d'un texte réduit formé de phrases saillantes. Il existe deux principales catégories d'approches pour le résumé automatique. Le premier type d'approche s'appuie sur l'extraction de phrases du texte original (Lin et Hovy, 2002). Un sous-ensemble de phrases saillantes du texte original est ainsi sélectionné. Dans le second type d'approche, appelé compression de phrases(Knight et Marcu, 2000), l'objectif est de réduire les phrases tout en préservant leur sens. Cependant, les méthodes de résumé automatique ne permettent pas de produire une vue relationnelle de la structure ou de l'organisation des différentes parties du texte.

D'un autre côté, les techniques de visualisation de collections de textes ont connu un intérêt grandissant ces dernières années (Newman *et al.*, 2010; Don *et al.*, 2007; Fekete et Dufournaud, 2000; Plaisant *et al.*, 2006). Par exemple, Newman *et al.* (2010) utilisent un modèle probabiliste pour produire un ensemble de thématiques décrivant une collection afin que l'utilisateur puisse effectuer une recherche de documents liés à une thématique particulière. Don *et al.* (2007) ont proposé un outil pour visualiser une collection de textes et permettre à l'utilisateur de l'explorer en affichant des motifs textuels fréquents (par exemple, des mots fréquents ou des ensembles de trigrammes). Les occurrences des motifs sont alors mises en relief dans le texte. Cependant, les approches de visualisation présentent un inconvénient commun aux approches de résumé automatique : le texte est visualisé de manière globale, sans mettre en évidence les relations entre les phrases.

Parmi les travaux intéressants en linguistique concernant l'exploration de textes, Hoey a présenté un modèle linguistique pour analyser des textes non-narratifs en s'appuyant sur les répétitions lexicales (Hoey, 1991). L'approche met en évidence l'organisation du texte (par exemple, le développement du texte ou son contenu conceptuel) en révélant les appariements entre phrases contiguës ou non contiguës, ce qui permet de construire une représentation du texte sous forme de *réseaux phrastiques*. Cette approche est intéressante pour plusieurs tâches tel que le raisonnement logique sur un sujet particulier du texte, l'étude de la cohésion lexicale du texte (Legallois *et al.*, 2011), le résumé de texte (Renouf et Kehoe, 2004) ou encore la segmentation de texte (Sardinha, 1999). Plusieurs études ont montré l'intérêt de la méthode sur des textes en anglais (Hoey, 1997; Károly et Francis, 2000) mais aussi sur des textes en français (Legallois, 2006). Alors qu'il est difficile de l'appliquer à la main sur des textes conséquents, peu de travaux utilisent une implémentation du modèle de Hoey. Dans Renouf et Kehoe (2004), les auteurs ont développé un outil de résumé basé sur le modèle de Hoey : SEAGULL (*Summary Extraction Algorithm Generated Using Lexical Links*). Ils ont réalisé des expérimentations pour montrer que leur outil obtient de meilleures performances que d'autres outils de résumé automatique mais ils n'ont utilisé pour cela qu'un petit texte en anglais (730 mots). Dans Legallois *et al.* (2011) les auteurs ont proposé un processus automatique appliquant le modèle de Hoey sur des textes de grande taille, afin d'étudier la cohésion lexicale de ces textes. Les expérimentations menées sur différents types de textes en français (narratif, expositif) ont permis de montrer l'intérêt de ce modèle pour cette tâche. Cependant, la grande taille des réseaux phrastiques obtenus par l'application de ce modèle demeure un inconvénient. En effet, cette représentation ne permet pas de visualiser de longs textes en entier et l'extraction de sous-parties potentiellement intéressantes n'est pas prévue par le modèle.

Dans cet article, nous proposons une approche pour extraire automatiquement, à partir d'un texte, des sous-ensembles de phrases cohérents d'un point de vue lexical. De plus, les sous-

ensembles sont représentés par des graphes, ce qui offre une vue des relations entre les phrases de ces sous-ensembles. Enfn, la taille des sous-ensembles de phrases étant raisonnable, cela permet aux linguistes de les analyser. Notre approche s'appuie sur une représentation du texte sous forme de graphe par application du modèle linguistique de Hoey. Pour pouvoir analyser de grands textes, nous proposons une implémentation du modèle de Hoey permettant de traiter des textes de grande taille. La principale contribution est l'utilisation d'une technique particulière de fouille de graphes, appelée *fouille de CoHoP*, pour extraire des sous-parties cohérentes du texte représenté sous forme de graphes. À notre connaissance, cette technique de fouille de graphes n'a jamais été utilisée dans le domaine du traitement automatique des langues. Dans notre approche, le processus de fouille est dit « sous contraintes linguistiques » car le graphe initial représentant le texte est construit par application du modèle de Hoey. D'autres contraintes linguistiques sur les sommets du graphe guident également le processus de fouille.

La fouille de graphes a connu un intérêt grandissant dans le domaine de la fouille de données pour la découverte de nouvelles connaissances (Washio et Motoda, 2003), et plus particulièrement la fouille de graphes *enrichis* (des attributs sont alors associés aux sommets). De telles méthodes de fouille ont été utilisées avec succès pour des tâches comme le clustering (Ge *et al.*, 2008; Zhou *et al.*, 2010) ou l'extraction de sous-graphes approximatifs (Tong *et al.*, 2007). Dans cet article, nous nous focalisons sur la fouille d'un certain type de motifs à partir de graphes enrichis : des *collections de k-PC homogènes* (CoHoP) (Mougel *et al.*, 2012). Nous les utilisons pour extraire des sous-parties homogènes du texte.

Dans la suite de l'article, nous présentons le modèle linguistique de Hoey, dans la section 2, puis les techniques de fouille de graphes pour extraire des motifs de type CoHoP, dans la section 3. Nous décrivons ensuite notre approche permettant d'extraire des sous-parties cohérentes des réseaux phrastiques en s'appuyant sur des méthodes de fouille de graphes, dans la section 4. Nous discutons enfin des expérimentations menées sur deux textes anglais, dans la section 5.

2 Modèle linguistique de Hoey

Le modèle linguistique introduit dans Hoey (1991) repose sur la notion de répétition lexicale au sein d'un texte. Il consiste alors à identifier les phrases du texte qui partagent au moins trois unités lexicales. Une *répétition lexicale* peut correspondre à la stricte répétition de l'unité lexicale (*cerveau/cerveau*) mais aussi à la répétition d'unités lexicales partageant le même lemme ou une autre forme dérivée (*produire/production*). La répétition lexicale peut également correspondre à une reprise anaphorique, une relation de synonymie (*acheter/acquérir*), une relation d'hypo/hyperonymie (*chien/animal*), une relation « implicative » (*conduire/voiture*) ou encore une suite ordonnée (*lundi/mardi/mercredi/...*).

Lorsque deux phrases partagent au moins trois unités lexicales, la paire de phrases est appariée. Un *appariement* entre deux phrases correspond ainsi à un chemin entre ces phrases. On appelle alors *réseau phrastique* un ensemble d'au moins trois phrases tel que, quelles que soient deux phrases de cet ensemble, ces phrases sont soit directement appariées, soit indirectement reliées par une succession de chemins dans l'ensemble de phrases [1]. L'ensemble des réseaux phrastiques d'un texte est appelé *hypotexte*. L'hypotexte peut être vu en quelque sorte comme un résumé du

1. Notons qu'en théorie des graphes, un réseau phrastique correspond à une composante connexe du graphe constitué des appariements.

texte original. Il est à noter que les phrases non appariées n'apparaissent pas dans l'hypotexte.

La figure 1 montre un extrait d'un réseau phrastique de "*Love Online: Emotions on the Internet*" (Ben-Ze'ev, 2004). Dans cet exemple, la répétition lexicale repose uniquement sur les lemmes des lexèmes communs des phrases (les *lexèmes* correspondent aux noms, adjectifs, adverbes et verbes). Ainsi, seules les répétitions strictes sont considérées. Le réseau phrastique est interprétable avec quelques « aménagements » mineurs. Dans l'exemple de la figure 1, certains mots ont été ajoutés au texte original et mis entre crochets afin de faciliter la lecture des enchaînements phrastiques du réseau (par exemple, [*However*]). La numérotation au début de chaque phrase (entre crochets) correspond à la position de la phrase dans le texte original ; il est alors intéressant de constater que l'empan de cet extrait de réseau est relativement conséquent.

Le modèle de Hoey permet de représenter un texte afin d'analyser sa cohésion lexicale. Cependant, la représentation sous forme d'hypotexte construite par ce modèle est de taille trop importante pour être visualisée en entier, ce qui rend fastidieuse l'exploration et l'analyse du texte. Ainsi, il est intéressant de disposer d'une méthode permettant d'extraire des sous-parties homogènes de réseaux phrastiques afin de faciliter l'analyse de ces réseaux. Dans ce but, nous introduisons, dans la section suivante, une approche de fouille de graphes permettant d'extraire des motifs appelés CoHoP.

3 Fouille de graphes : motifs de type CoHoP

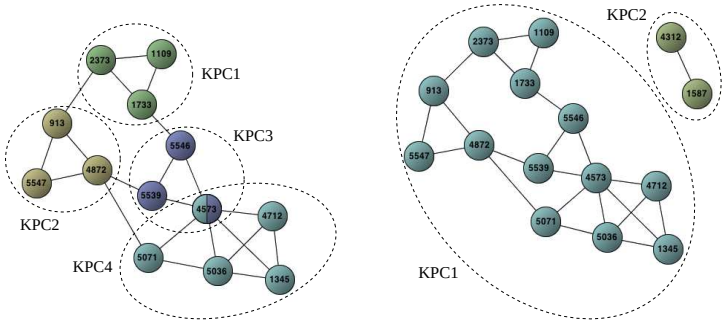
Les CoHoP (*collection de k-cliques percolées*) sont des motifs extraits à partir de graphes attribués booléens (Mougel *et al.*, 2012). Elle peuvent être vues comme des ensembles de communautés dont les éléments partagent des propriétés communes : chaque communauté correspond à une *k-clique percolée* (*k*-PC). Les CoHoP rendent ainsi compte d'une structure cachée, sous-jacente au graphe initial. Dans cette section, nous présentons les deux principales notions sur lesquelles s'appuie cette technique particulière de fouille de graphe : les *k*-PC et les CoHoP.

3.1 *k*-cliques percolées (*k*-PC)

Dans un graphe, une *k-clique* est un ensemble de k sommets dans lequel chaque paire de sommets distincts est connectée par une arête. La notion de *k-clique percolée* (*k*-PC) peut être vue comme une version relâchée du concept de clique. Une *k*-PC a été définie par Derenyi *et al.* (2005) comme étant l'union de toutes les *k*-cliques connectées par des chevauchements de $k-1$ sommets. Ainsi, dans une *k*-PC, chaque *k*-clique peut être atteinte de n'importe quelle autre *k*-clique par un chemin de *k*-cliques adjacentes et chaque sommet d'une *k*-PC peut être atteint

[1109] **Online**$_1$ **emotional**$_2$ **experiences**$_3$ may be compared to receiving a salary without earning it by hard work.
[1733] [However] **Online**$_1$ **relationships**$_4$ have a profound impact upon our **emotional**$_2$ **experiences**$_3$.
[2373] Since **emotional**$_2$ self-disclosure is more important to the **experience**$_3$ of intimacy than factual self-disclosure, 13 **online**$_1$ **relationships**$_4$ often have a higher degree of intimacy than offline **relationships**$_4$.

FIGURE 1: Extrait d'un réseau phrastique de "*Love Online: Emotions on the Internet*"



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

FIGURE 2: Exemple de CoHoP extraite à partir des attributs $\{a_1, a_2\}$, pour deux valeurs de k

par n'importe quel autre sommet de cette k-PC par un chemin de sous-ensembles de sommets bien connectés (les k-cliques).

Dans la figure 2a, il y a quatre k-PC ($k = 3$) : {1109, 1733, 2373}, {913, 4872, 5547}, {4573, 5539, 5546} et {1345, 4573, 4712, 5036, 5077}. Les trois premières k-PC contiennent une seule 3-clique alors que la dernière k-PC contient cinq 3-cliques (*e.g.*, {1345, 4712, 5036} et {1345, 4712, 4573}). Revenons sur la création de cette dernière k-PC. Nous pouvons tout d'abord constater que les sommets 1345, 4573, 4712 et 5036 sont directement connectés les uns aux autres : ils appartiennent ainsi à la même k-PC. Le sommet 5071 appartient également à cette k-PC puisqu'il est accessible à partir de chacun des quatre sommets précédents, par une série de k-cliques se chevauchant (le paramètre k a un impact sur le nombre de sommets à considérer dans les k-cliques ; dans cet exemple, sa valeur est fixée à 3) : par exemple, pour aller du sommet 5071 au sommet 4712, un chemin de 3-cliques se chevauchant peut être {4712, 4573, 5036} suivi de {4573, 5036, 5071} (avec $k = 3$, les chevauchements de 3-cliques contiennent deux sommets). En revanche, le sommet 4872 n'appartient pas à cette k-PC. En effet, pour cela il faudrait qu'il y ait une 3-clique entre les sommets 4573, 5071 et 4872, ce qui n'est pas le cas.

Il est à noter que le calcul des k-PC est indépendant des ensembles d'attributs associés aux sommets (les graphes utilisés étant attribués). De plus, une k-clique ne peut appartenir qu'à au plus une k-PC alors qu'un sommet peut se trouver dans plusieurs k-PC, puisqu'il peut appartenir à plusieurs k-cliques. Un sommet appartenant à plusieurs k-PC est appelé nœud relais ou *bridging node* (Musial et Juszczyszyn, 2009). Ainsi, dans la figure 2a, le sommet 4573 est un nœud relais car il appartient à deux k-PC : KPC_3 et KPC_4.

3.2 Collections de k-PC homogènes (CoHoP)

Une *collection de k-PC homogènes* (CoHoP) a été définie par (Mougel *et al.*, 2012) comme étant un ensemble de sommets tels que, étant donnés k, α et γ des entiers positifs définis par des utilisateurs :

– tous les sommets sont *homogènes*, c'est-à-dire qu'ils partagent au moins α attributs ;

– la collection contient au moins γ k-PC ;
– toutes les k-PC ayant les mêmes attributs sont présentes dans la collection (contrainte de *maximalité*).

La figure 2a représente ainsi une CoHoP extraite à partir de l'ensemble d'attributs $\{a_1, a_2\}$; comme vu dans la section 3.1, elle contient quatre k-PC ($\alpha = 2, k = 3, \gamma = 4$). Il est à noter que, contrairement au calcul des k-PC, l'extraction des CoHoP dépend fortement de l'ensemble d'attributs associés aux sommets du graphe. Sur la figure 2a, les ensembles d'attributs des sommets ne sont pas illustrés (pour ne pas surcharger la figure) mais chaque sommet est en fait étiqueté par un ensemble d'attributs qui contient au moins les attributs a_1 et a_2. En effet, cette CoHoP a été extraite à partir de ces deux attributs.

Les trois paramètres - k, α et γ - ont un impact inportant sur la structure des CoHoP extraites. Comme précisé précédemment, le paramètre α fixe le nombre minimal d'attributs communs associés aux sommets des CoHoP extraites et le paramètre γ fixe le nombre minimal de k-PC présentes dans les CoHoP. Le paramètre k a également un impact important sur la structure des CoHoP extraites. En effet, augmenter sa valeur a pour conséquence d'augmenter le degré de cohésion entre les sommets appartenant à une même k-PC. La figure 2b représente la CoHoP extraite à partir du même ensemble d'attributs que celle illustrée par la figure 2a mais en fixant cette fois $k = 2$. Cette CoHoP comporte maintenant 15 sommets répartis en seulement deux k-PC, la plus grosse k-PC (KPC_1) correspondant en fait aux quatre k-PC de la figure 2a. Ainsi, le choix de la valeur de k permet de choisir le degré de cohésion souhaité entre les sommets de chaque k-PC. En effet, un plus grand nombre de sommets doit être directement relié les uns aux autres lorsque la valeur de k augmente (la valeur de k représente ce nombre minimal de sommets).

4 Fouille de réseaux phrastiques

Dans cette section, nous proposons une nouvelle approche qui s'appuie à la fois sur le modèle de Hoey et sur la fouille de motifs de type CoHoP. Notre approche permet d'extraire des sous-parties homogènes de réseaux phrastiques afin de faciliter leur analyse.

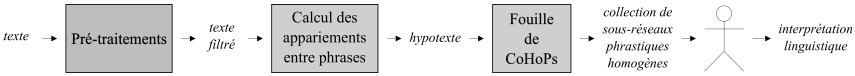


FIGURE 3: Vue d'ensemble de l'extraction de collections de sous-réseaux phrastiques homogènes

La figure 3 illustre les différentes étapes de notre approche, allant du pré-traitement du texte à l'extraction des collections de sous-réseaux phrastiques homogènes, en passant par la construction de l'hypotexte représentant le texte étudié. Ces différentes étapes sont décrites plus en détail dans les sous-sections suivantes.

4.1 Construction de l'hypotexte

Le texte est tout d'abord étiqueté à l'aide de TreeTagger (Schmid, 1994). Le texte étiqueté est ensuite découpé en phrases à chaque signe de ponctuation de l'ensemble suivant : { « . », « ? »,

« ! », « : »}. Les phrases sont finalement filtrées afin de ne conserver que leurs unités lexicales pertinentes. Dans notre cas, cela consiste à ne garder que les *lexèmes* (noms, adjectifs, adverbes et verbes hormis les auxiliaires). En fait, nous considérons les lemmes de ces lexèmes. À l'issue de cette étape, chaque phrase du texte filtré est alors représentée par les lemmes de ses lexèmes. Par exemple, la phrase « *Online emotional experiences may be compared to receiving a salary without earning it by hard work.* » est représentée par « *online emotional experience compare receive salary earn hard work* ». Tous les mots non pertinents sont ainsi filtrés et les mots restants (*e.g.*, les lexèmes) sont remplacés par leur lemme.

À partir du texte filtré, nous construisons ensuite la représentation du texte sous forme de graphe en appliquant le modèle linguistique de Hoey, comme présenté à la section 2. L'hypotexte est ainsi créé en appariant toutes les paires de phrases qui partagent au moins trois unités lexicales communes, correspondant dans notre cas aux lemmes des lexèmes. Il est également à noter que les phrases non appariées ne sont pas conservées dans l'hypotexte.

4.2 Fouille de collections de sous-réseaux phrastiques homogènes

L'objectif de cette dernière étape est d'extraire des sous-parties homogènes de l'hypotexte. L'hypotexte créé comme décrit précédemment peut être vu comme un graphe attribué. En effet, un hypotexte est un graphe dont chaque sommet représente une phrase appariée et dont chaque arête représente un appariement entre deux phrases qui partagent au moins trois unités lexicales communes. De plus, l'ensemble des unités lexicales d'une phrase peut étiqueter le sommet correspondant, en tant qu'ensemble d'attributs.

Avec cette représentation de l'hypotexte comme un graphe attribué, nous pouvons utiliser des algorithmes de fouille de CoHoP, comme présenté à la section 3. Dans notre approche, le processus de fouille est effectué « sous contraintes linguistiques » car, d'une part, la structure initiale du graphe est construite en utilisant le modèle linguistique de Hoey et, d'autre part, les ensembles d'attributs qui étiquettent les sommets sont les unités lexicales des phrases correspondantes.

Dans notre approche, chaque CoHoP extraite correspond alors à ce que nous appelons une *collection de sous-réseaux phrastiques homogènes* (CoHoSS). En effet, de la même façon qu'une CoHoP est constituée de k-PC homogènes (*i.e.*, des ensembles de sommets partageant un sous-ensemble de α attributs communs), une CoHoSS est constituée de sous-réseaux phrastiques homogènes (*i.e.*, des ensembles de phrases partageant un sous-ensemble de α unités lexicales communes). Chaque sous-réseau phrastique correspond alors à la définition d'une k-PC. Ainsi, dans un sous-réseau phrastique, chaque phrase est soit directement appariée aux autres phrases du sous-réseau par une arête (si elle partage au moins trois unités lexicales avec chacune de ces phrases), soit indirectement accessible depuis n'importe quelle autre phrase par une série de sous-ensembles de phrases bien connectés (chaque sous-ensemble correspond à une k-clique, comme défini à la section 3.1). Ainsi, la CoHoP représentée par la figure 2a correspond à une collection de sous-réseaux phrastiques homogènes (*i.e.*, une CoHoSS) dont toutes les phrases partagent les mots *emotional* et *experience* (correspondant aux attributs a_1 et a_2, respectivement). En effet, chaque sommet de la CoHoP correspond à une phrase, le numéro du sommet indiquant la position de la phrase dans le texte (la figure 1 donne les phrases associées aux sommets 1109, 1733 et 2373 de la CoHoP). De plus, les ensembles d'attributs associés aux sommets correspondent aux ensembles d'unités lexicales représentant les phrases. Par exemple, l'ensemble d'attributs A_{4712} de la phrase 4712 correspond à l'ensemble d'unités lexicales suivant :

{*parallel, world, help, preserve, actual, not, give, exciting, emotional, experience*}.

Les CoHoSS représentent ainsi des collections de sous-réseaux phrastiques de l'hypotexte initial qui ont une certaine cohésion lexicale par rapport à l'ensemble des unités lexicales à partir desquelles elles ont été extraites. Les CoHoSS ainsi que leur structure peuvent ensuite être analysées par des linguistes pour interpréter, par exemple, chacun des sous-réseaux ainsi que la façon dont ils sont connectés les uns aux autres, notamment par les *phrases relais* (correspondant aux nœuds relais présentés dans la section 3.1).

5 Résultats expérimentaux

Dans cette section, nous présentons les résultats expérimentaux sur deux textes anglais. Les textes ainsi que les outils utilisés pour extraire et visualiser les CoHoP sont tout d'abord présentés dans la section 5.1. Nous discutons ensuite les résultats quantitatifs sur l'utilisation du modèle de Hoey et sur les CoHoSS extraites, dans les sections 5.2.1 et 5.2.2 respectivement. Enfin, dans la section 5.2.3, nous présentons un exemple de CoHoSS extraite et son interprétation linguistique afin de montrer l'intérêt de notre approche pour l'exploration de textes.

5.1 Paramètres : données et outils

5.1.1 Textes étudiés

Pour évaluer l'approche que nous proposons, nous avons choisi deux textes de grande taille, chacun correspondant à un texte expositif en anglais : "*The Origin of Speech*" (MacNeilage, 2008) et "*Love Online: Emotions on the Internet*" (Ben-Ze'ev, 2004). Les caractéristiques de ces textes sont données dans la table 1.

Texte	Titre	Auteur	Année	Nb. pages
Speech	*The Origin of Speech*	Peter F. MacNeilage	2008	416
Love	*Love Online : Emotions on the Internet*	Aaron Ben-Ze'ev	2004	302

TABLE 1: Caractéristiques des textes étudiés

5.1.2 Extraction de motifs de type CoHoP

Afin d'extraire les CoHoP, comme présenté dans la section 4.2, nous utilisons *CoHoP Miner* (Mougel *et al.*, 2012). Cet outil permet d'extraire des CoHoP en fixant les valeurs des différents paramètres utilisés lors du processus de fouille (k, α et γ). Il inclut également une interface graphique pour visualiser les CoHoP extraites, comme l'illustre la figure 4. Les k-PC de la CoHoP affichée peuvent ainsi être visualisées, chaque k-PC étant représentée par une couleur différente. Les sommets appartenant à plusieurs k-PC peuvent également être visualisés afin de se focaliser sur eux lors de l'analyse des CoHoP, par exemple. Il est également possible d'afficher

l'ensemble d'attributs de chaque sommet. En fait, cet outil offre une visualisation globale des motifs extraits puisqu'il permet de sélectionner les motifs à analyser par la suite selon l'ensemble des attributs à partir desquels ils ont été extraits ou encore selon leur structure (par exemple, le nombre de *k*-PC qu'ils contiennent ou la présence de sommets appartenant à plusieurs *k*-PC). Cependant, afin d'analyser les CoHoP, les linguistes ont besoin d'afficher les phrases correspondant aux sommets ainsi que les mots à partir desquels les phrases ont été appariées (et conduisant à des arêtes entre les sommets). C'est pourquoi nous utilisons également un outil de visualisation de graphes.

5.1.3 Visualisation des réseaux phrastiques

Cet outil de visualisation de graphes a été développé en Java. Il offre une visualisation des sous-réseaux phrastiques correspondant aux CoHoP extraites, comme illustré par la figure 5. La CoHoSS affichée correspond à la CoHoP de la figure 4 : les sommets correspondent aux phrases du texte et les arêtes sont étiquetées par les unités lexicales partagées par chaque paire de phrases. Cet outil permet aux linguistes d'analyser plus facilement les réseaux phrastiques en offrant une visualisation focalisée sur une collection particulière de sous-réseaux phrastiques homogènes. Cependant, il n'offre pas d'affichage des différentes *k*-PC de la CoHoP extraite, ce qui rend plus difficile l'analyse des phrases appartenant à plusieurs sous-réseaux (*i.e.*, les phrases relais), par exemple. Il est donc intéressant d'utiliser cet outil en complément de *CoHoP Miner*.

5.2 Expérimentations sur la fouille de réseaux phrastiques

5.2.1 Résultats quantitatifs sur les réseaux phrastiques

Nous présentons tout d'abord des résultats quantitatifs sur l'hypotexte créé pour représenter chaque texte. Ces résultats sont donnés dans la table 2. Nous pouvons tout d'abord remarquer que chaque phrase contient en moyenne 10 lexèmes pour *Speech* et 11 lexèmes pour *Love* alors que les phrases comportent respectivement 24 et 20 mots, en moyenne. Représenter les phrases par leurs lexèmes permet ainsi de réduire le nombre d'attributs qui leur est associé. Nous pouvons également constater que peu de réseaux phrastiques ont été créés : deux pour chaque texte.

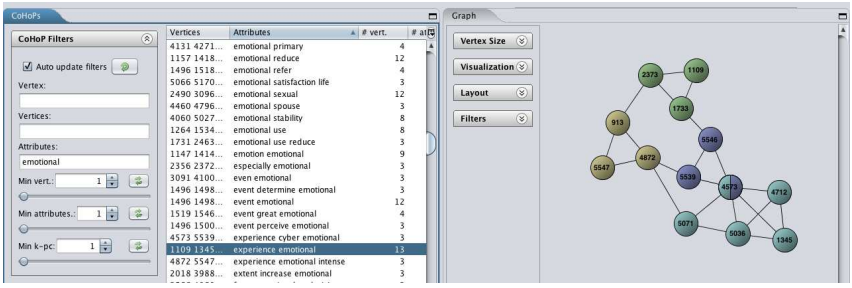


FIGURE 4: Visualisation d'une CoHoP du texte *Love*, avec *CoHoP Miner*

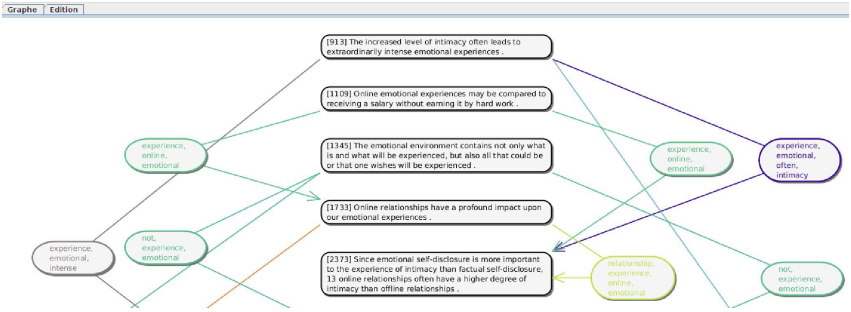


FIGURE 5: Extrait de la visualisation d'une CoHoSS du texte *Love*

De plus, l'hypotexte de chacun des textes (constitué ainsi des deux réseaux phrastiques) est de taille importante puisqu'il contient plus de 75 % des phrases. Cela suggère une forte cohésion lexicale à l'intérieur des textes (chaque phrase de l'hypotexte est en moyenne appariée avec 13 phrases pour *Speech* et 30 phrases pour *Love*).

5.2.2 Résultats quantitatifs sur les CoHoSS extraites

La table 3 donne le nombre de CoHoSS extraites pour chaque texte, pour différentes valeurs de k, α et γ. Par exemple, avec $(k = 3, \alpha = 2, \gamma = 2)$, 523 CoHoSS sont extraites pour le texte *Love* (voir la ligne 2 de la table 3a).

Une première série d'expériences a été menée en fixant $\gamma = 2$: cela signifie que l'on fixe le nombre de sous-réseaux phrastiques contenus dans une CoHoSS à au minimum deux. La table 3a donne le nombre de CoHoSS extraites avec ce paramétrage de γ et en faisant varier α et k. Nous observons que quelle que soit la valeur de k, le fait d'augmenter la valeur de α (c'est-à-dire le nombre d'attributs communs aux phrases) implique logiquement une diminution significative du nombre de CoHoSS (environ 50 % en faisant varier α de 1 à 2). Plus particulièrement, nous remarquons qu'aucune CoHoSS n'est extraite pour une valeur $\alpha \geq 3$. Cela signifie que les CoHoSS sont extraites à partir d'un ou deux attributs (ici, les lemmes des lexèmes des phrases).

Une autre série d'expériences a été menée en fixant cette fois $\alpha = 2$: cela signifie que l'on fixe le nombre d'attributs à au minimum deux. La table 3b donne alors le nombre de CoHoSS extraites avec ce paramétragede α et en faisant varier γ et k. Nous constatons qu'augmenter la valeur de γ

Texte	Nb. mots	Nb. total lexèmes	Nb. lexèmes différents	Nb. phrases	Nb. appariements	Nb. réseaux phrastiques	% phrases dans hypotexte
Speech	127 563	59 657	4 728	5 308	50 277	2	75,6%
Love	112 325	53 035	6 919	5 571	131 497	2	79,0%

TABLE 2: Résultats quantitatifs sur les hypotextes

Texte	k	α		
		1	2	3
Love	2	1 010	555	0
	3	924	523	0
	4	729	403	0
Speech	2	1 420	793	0
	3	973	523	0
	4	678	384	0

(a) $\gamma = 2$

Texte	k	γ				
		1	2	3	4	5
Love	2	38 061	555	15	0	0
	3	16 437	523	41	9	1
	4	8 425	403	51	13	4
Speech	2	39 442	793	25	2	0
	3	13 673	523	84	17	2
	4	5 552	384	75	24	2

(b) $\alpha = 2$

TABLE 3: Nombre de CoHoSS extraites en fixant la valeur (a) de γ et (b) de α

(c'est-à-dire le nombre de sous-réseaux phrastiques contenus dans une CoHoSS) diminue moins rapidement le nombre de CoHoSS extraites pour des valeurs élevées de k. En effet, comme présenté dans la section 3, augmenter la valeur de k permet d'obtenir des sous-réseaux avec un degré de cohésion plus élevé. Cela s'accompagne d'un nombre plus élevé de sous-réseaux dans les CoHoSS et donc d'une valeur plus élevée pour γ. Ainsi, en fixant les valeurs de α et γ, le nombre de sous-réseaux dans les CoHoSS extraites sera généralement plus élevé lorsque la valeur de k est également élevée.

D'une manière générale, nous pouvons observer que le nombre de CoHoSS extraites diminue lorsque les valeurs de k, α et γ augmentent. En conclusion, la valeur de k doit être choisie de manière judicieuse en fonction du degré de cohésion lexicale souhaitée pour les sous-réseaux phrastiques contenus dans les CoHoSS extraites. La valeur de γ, quant à elle, permet de limiter le nombre total de CoHoSS extraites, en ne choisissant que les plus grosses en termes de nombre de sous-réseaux phrastiques qu'elles contiennent. Enfin, la valeur de α joue un rôle important lorsque l'on souhaite focaliser l'analyse linguistique sur les relations entre les phrases partageant un certain nombre d'unités lexicales.

5.2.3 Exemple de CoHoSS extraite et interprétation linguistique

Nous présentons maintenant un exemple de CoHoSS. La figure 6 illustre la CoHoP extraite à l'aide de *CoHoP Miner*, sur le texte *Speech*, et la figure 7 donne les phrases correspondantes.

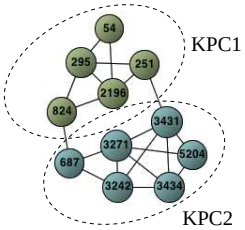


FIGURE 6: CoHoP correspondant à la CoHoSS extraite à partir de {*adaptation*} ($k = 3$)

La CoHoSS a été extraite à partir de l'attribut *adaptation*, en utilisant les valeurs suivantes pour

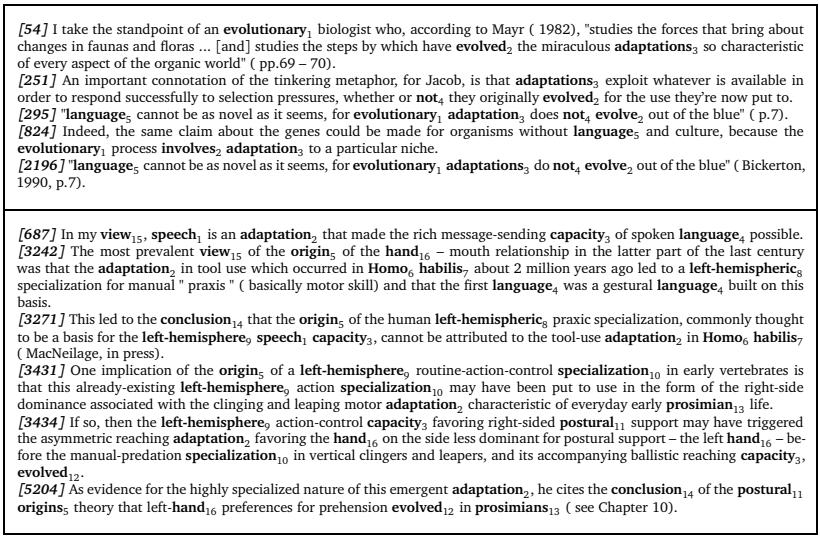


FIGURE 7: Phrases correspondant à la CoHoP de la figure 6

les paramètres de fouille : $k = 3, \alpha = 1, \gamma = 2$. Cette CoHoSS est constituée de deux sous-réseaux phrastiques. Le premier réseau (contenant les phrases 54, 251, 295, 824 et 2196) traite du thème général de la CoHoSS : le phénomène d'adaptation. Nous pouvons remarquer que ce réseau est relativement cohérent alors qu'il parcourt un ensemble considérable du texte (correspondant à un empan de plus de 2 000 phrases). Le second réseau, quant à lui, développe une thématique plus spécifique de l'adaptation : la spécialisation de l'hémisphère gauche. Ce sous-réseau commence avec la phrase 687 qui est connectée au sous-réseau précédent par la phrase 824. Nous pouvons également constater que les deux sous-réseaux se « chevauchent » au niveau du texte puisque la phrase 687 appartient au second sous-réseau alors que les phrases 824 et 2196 appartiennent au premier. De plus, nous pouvons remarquer que l'empan de la CoHoSS (mais aussi de chacun des sous-réseaux) est relativement important puisqu'elle commence à la phrase 54 et qu'elle se termine à la phrase 5204. Ainsi, cette propriété intéressante de non-contiguïté des phrases des réseaux phrastiques se retrouve également au niveau des sous-réseaux phrastiques constituant les CoHoSS extraites.

6 Conclusion

Dans cet article, nous avons proposé une approche pour explorer des textes de taille conséquente en se focalisant sur des sous-parties cohérentes. Cette méthode d'exploration s'appuie sur une représentation du texte à l'aide d'un graphe, en utilisant le modèle linguistique de Hoey

pour sélectionner et apparier les phrases conservées dans le graphe. Notre contribution porte sur l'utilisation de techniques issues de la fouille de graphes pour extraire des sous-parties du texte cohérentes d'un point de vue lexical (c'est-à-dire des collections de sous-réseaux phrastiques homogènes) dont la taille permet à un linguistique de les analyser. Nous avons réalisé des expérimentations sur deux textes anglais de la taille d'un livre pour valider cette approche. Cela nous a permis de montrer que le graphe généré à l'aide du modèle de Hoey était difficilement exploitable par un humain à cause du trop grand nombre de sommets et d'arêtes. En utilisant notre approche pour sélectionner des sous-parties pertinentes du graphe, il est alors possible d'appliquer le modèle de Hoey sur de grands textes. De plus, les différents paramètres utilisés lors du processus de fouille du graphe offrent la possibilité de définir le niveau de granularité des collections de sous-réseaux phrastiques homogènes extraites. D'un point de vue linguistique, cela signifie que le degré de cohésion lexicale entre les phrases des sous-réseaux phrastiques extraits est mis en évidence.

Remerciements

Les auteurs tiennent à remercier chaleureusement Pierre-Nicolas Mougel et Christophe Rigotti (LIRIS, Lyon) pour la mise à disposition de *CoHoP Miner*.

Ce travail bénéficie du soutien de la région Basse-Normandie et de l'ANR (projet Hybride ANR-11-BS02-002).

Références

Ben-Ze'ev, A. (2004). *Love Online : Emotions on the Internet*. Cambridge University Press.

Derenyi, I., Palla, G. et Vicsek, T. (2005). Clique percolation in random networks. *Physical Review Letters*, 94:160–202.

Don, A., Zheleva, E., Gregory, M., Tarkan, S., Auvil, L., Clement, T., Shneiderman, B. et Plaisant, C. (2007). Discovering interesting usage patterns in text collections : integrating text mining with visualization. *In Proc. of the Conference on Information and Knowledge Management*, pages 213–222.

Fekete, J. et Dufournaud, N. (2000). Compus : visualization and analysis of structured documents for understanding social life in the 16th century. *In Proc. of the ACM Conference on Digital Libraries*, pages 47–55.

Ge, R., Ester, M., Gao, B., Hu, Z., Bhattacharya, B. et Ben-Moshe, B. (2008). Joint cluster analysis of attribute data and relationship data. *ACM Transactions on Knowledge Discovering Data*, 2(2):1–35.

Hoey, M. (1991). *Patterns of Lexis in Text*. Describing English Language. Oxford University Press.

Hoey, M. (1997). Language and the subject. *In* Simms, K., éditeur : *Critical Studies*, chapitre The discourse's disappearing (and reappaering) subject : an exploration of the extent of Intertextual interference in the production of texts, pages 245–264. Rodopi.

KÁROLY, S. et FRANCIS, G. (2000). *Pattern Grammar, a corpus-driven approach to the lexical grammar of English*. John Benjamins.

KNIGHT, K. et MARCU, D. (2000). Statistics-Based Summarization — Step One : Sentence Compression. *In Proc. of the National Conference of the American Association for Artificial Intelligence*, pages 703–710.

LEGALLOIS, D. (2006). Des phrases entre elles à l'unité réticulaire du texte. *Langages*, 164:56–70.

LEGALLOIS, D., CELLIER, P. et CHARNOIS, T. (2011). Calcul de réseaux phrastiques pour l'analyse et la navigation textuelle. *In Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*.

LIN, C.-Y. et HOVY, E. (2002). From single to multi-document summarization. *In Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 457–464.

MACNEILAGE, P. (2008). *The Origin of Speech*. UOP Oxford.

MOUGEL, P.-N., RIGOTTI, C. et GANDRILLON, O. (2012). Finding collections of k-clique percolated components in attributed graphs. *In Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. À paraître.

MUSIAL, K. et JUSZCZYSZYN, K. (2009). Properties of bridge nodes in social networks. *In Proc. of the International Conference on Computational Collective Intelligence*, pages 357–364.

NEWMAN, D., BALDWIN, T., CAVEDON, L., HUANG, E., KARIMI, S., MARTÍNEZ, D., SCHOLER, F. et ZOBEL, J. (2010). Visualizing search results and document collections using topic maps. *Web Semantics Science Services and Agents on the World Wide Web*, 8(2-3):169–175.

PLAISANT, C., ROSE, J., YU, B., AUVIL, L., KIRSCHENBAUM, M., SMITH, M., CLEMENT, T. et LORD, G. (2006). Exploring erotics in emily dickinson"s correspondence with text mining and visual interfaces. *In Proc. of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 141–150.

RENOUF, A. et KEHOE, A. (2004). Textual distraction as a basis for evaluating automatic summarisers. *In Proc. of the International Conference on Language Resources and Evaluation*.

SARDINHA, T. B. (1999). Looking at discourse in a corpus : The role of lexical cohesion. *In Proc. of the World Congress of Applied Linguistics*.

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proc. of the International Conference on Knowledge Discovery and Data Mining*.

TONG, H., GALLAGHER, B., FALOUTSOS, C. et ELIASSI-RAD, T. (2007). Fast best-effort pattern matching in large attributed graphs. *In Proc. of the International Conference on Knowledge Discovery and Data Mining*.

WASHIO, T. et MOTODA, H. (2003). State of the art of graph-based data mining. *SIGKDD Explorations*, 5(1):59–68.

ZHOU, Y., CHENG, H. et YU, J. (2010). Clustering large attributed graphs : An efficient incremental approach. *In Proc. of the International Conference on Data Mining*, pages 689–698.