Incrémentation lexicale dans les textes : une auto-organisation

Matthias Tauveron
Fonctionnements Discursifs et Traduction
UR LiLPa, Université de Strasbourg
matthias.tauveron@etu.unistra.fr

RESUME

Nous proposons une étude dynamique du lexique, en décrivant la manière dont il s'organise progressivement du début à la fin d'un texte. Pour ce faire, nous nous focalisons sur la co-occurrence généralisée, en formant un graphe qui représente tous les lemmes du texte et synthétise leurs relations mutuelles de co-occurrence. L'étude d'un corpus de 40 textes montre que ces relations évoluent d'une manière auto-organisée : la forme – et l'identité – du graphe de co-occurrence restent stables après une phase d'organisation terminée avant la 1ère moitié du texte. Ensuite, il n'évolue plus : les nouveaux mots et les nouvelles relations de co-occurrence s'inscrivent peu à peu dans le réseau, sans modifier la forme d'ensemble de la structure. La relation de co-occurrence généralisée dans un texte apparaît donc comme la construction rapide d'un système, qui est ensuite assez souple pour canaliser un flux d'information sans changer d'identité.

ABSTRACT

Lexical Incrementation within Texts: a Self-Organization

We propose here a dynamic study of lexicon: we describe how it is organized progressively from the beginning to the end of a given text. We focus on the "generalized co-occurrence", forming a graph that represents all the lemmas of the text and their mutual co-occurrence relations. The study of a corpus of 40 texts shows that these relations have a self-organized evolution: the shape and the identity of the graph of co-occurrence become stable after a period of organization finished before the first half of the text. Then they no longer change: new words and new co-occurrence relations gradually take place in the network without changing its overall shape. We show that the evolution of the "generalized co-occurrence" is the quick construction of a system, which is then flexible enough to channel the flow of information without changing its identity.

MOTS-CLES : Texte ; lexique ; co-occurrence généralisée ; auto-organisation

KEYWORDS: Text; lexicon; generalized co-occurrence; self-organization

1 Introduction

Le texte n'est pas qu'une suite linéaire de mots, propositions ou phrases. Il croît au fur et à mesure de son déroulement, à la manière d'une boule de neige et possède ainsi une dimension « incrémentielle » (Legallois, 2006). Nous proposons ici d'étudier cette incrémentation au niveau lexical. Nous montrons que le texte est un agencement complexe mais non anarchique de mots – comme (Adam, 2004, 35) l'a dit à propos des

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 3: RECITAL, pages 15–28,
Grenoble, 4 au 8 juin 2012.

propositions – et proposons une description de la dynamique de cet agencement. Notre propos porte sur la « texture » ou la « textualité », c'est-à-dire la dimension formelle des textes et leur organisation, et ne concerne pas directement le sens ni sa construction en discours.

L'agencement des mots sera envisagé ici au travers de leurs relations de co-occurrence. Sous le nom de *co-occurrence généralisée*, l'étude de l'ensemble des relations de co-occurrence entre les lemmes d'un texte ou d'un corpus permet de révéler l'organisation en réseau de son lexique (Véronis, 2004, Viprey, 2006, Paranyushkin, 2010). Ces études en restent cependant à une image statique, montrant le réseau de co-occurrences tel qu'il se déploie une fois le texte lu dans son entier. Nous proposons ici une étude de la dynamique de cette co-occurrence généralisée en montrant son évolution du début à la fin d'un texte. Notre corpus de travail est formé en premier lieu de 20 textes courts (200 à 2000 mots, total de 11.015 mots), et en second lieu de 20 textes longs (5.000 à 22.500 mots, total de 192.477 mots).

Dans les cas typiques que nous observons, le réseau de co-occurrence généralisée croît très progressivement au début du texte. Cependant, de manière surprenante, à l'issue d'une première phase terminée avant la 1ère moitié du texte, ce réseau atteint un stade qui, sans être totalement figé, n'évolue plus guère par la suite. Ce réseau s'est donc formé une identité qui reste stable malgré l'apport ultérieur d'information. L'incrémentation lexicale dans la suite du texte ne fait que renforcer cette identité. Nous pensons qu'un tel comportement obéit à la définition que Moreno (2004) a donné des systèmes auto-organisés : ce sont des systèmes qui connaissent une *construction progressive de leur identité*. Nous expliciterons les arguments qui militent pour et contre la caractérisation du réseau de co-occurrence généralisée comme système auto-organisé, en montrant les enjeux linguistiques et cognitifs sous-jacents à cette question.

2 Méthode

2.1 Corpus

Notre corpus de textes courts est fait d'éditoriaux de revues scientifiques (revues *Développement durable et territoire*, *Vertigo*, *Langages*) et universitaires (revue *Savoir(s)*, « *magazine d'information de l'Université de Strasbourg* »). Une fois les premiers résultats obtenus, la démarche a été appliquée au corpus de textes longs : articles scientifiques (*Langages* 182 sur les théories du langage et les politiques des linguistes, *Intellectica*, 40 sur la notion de représentation) et encyclopédiques (*Wikipédia*, articles biographiques et analyses d'œuvres dans le domaine musical, grands articles portant sur la vie en société, tels « Religion », « Démocratie »). Nous disposons également d'un corpus de contrôle (textes totalement différents du corpus de travail sur lesquels on applique la même méthode, pour en tester le fonctionnement, le bien-fondé et la démarche) fait de deux textes poétiques d'Arthur Rimbaud issus des *Illuminations* (total : 1221 mots), choisis pour le caractère *a priori* désordonné de leur lexique. En tant

que tels, ils illustrent un cas extrême concernant l'organisation des mots dans un texte. On verra que leur étude permet de circonscrire certains comportements particuliers rencontrés dans le corpus de travail.

2.2Formation du graphe de co-occurrences

Le graphe de co-occurrences vise à donner une représentation visuelle synthétique de la co-occurrence généralisée dans un texte. Chaque lemme est représenté sous la forme d'un point (*nœud* du graphe) et un *lien* est tracé entre deux nœuds lorsque les lemmes correspondants sont co-occurrents. Tout lien est doté d'un *poids* qui indique le nombre de fois que la co-occurrence a été constatée dans le texte. On mesure ainsi directement l'importance de chaque lien. Nous considérons que deux unités sont co-occurrentes lorsqu'elles se trouvent à une distance de moins de trois mots[1], et ce, indépendamment des coupures de phrases.

La première étape du traitement consiste en une tokenisation et une lemmatisation faites en Perl grâce au dictionnaire fourni par l'ABU[2]. Le texte résultant est parcouru pour créer deux listes : celle des lemmes et celle des relations de co-occurrence. Ensuite, un programme Perl supplémentaire effectue, par l'intermédiaire du module *Graph* de Jarkko Hietaniemi[3], le calcul de la *betweenness centrality* des nœuds (définie *infra* en 2.3.1).

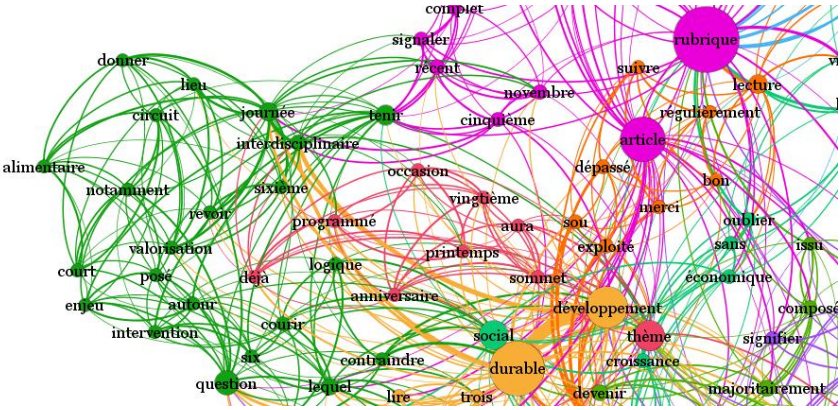

Figure 1 – Détail d'un graphe de co-occurrence pour un texte entier[4]

Notre choix de représenter les lemmes du texte sur le graphe (et non les formes rencontrées, ou encore des traits sémantiques) répond lui aussi à des raisons d'économie

[1] Le choix d'une distance aussi courte est motivée par la performance des outils de traitement, et par des questions de lisibilité du graphe. Comme nous l'a fait remarquer Tristan Vanrullen, accroître la taille de la fenêtre alourdit le graphe en multipliant les liens, ce qui génère par ailleurs un bruit important, notamment sur des textes longs portant sur plusieurs thématiques. La distance de 3 mots est un choix *a minima*.

[2] http://abu.cnam.fr/DICO/mots-communs.html

[3] http://search.cpan.org/~jhi/Graph-0.94/.

[4] Obtenu grâce à Gephi 0.7beta (www.gephi.org). Les couleurs signalent les tendances qu'ont les nœuds à être reliés : un nœud est plus lié aux nœuds de la même couleur que lui.

dans le traitement des données. Si la lemmatisation fait perdre des informations sur le sens des unités (Lemaire, 2008), nous pensons que cette perte n'est pas si significative dans notre problématique, formelle et non sémantique, d'organisation des mots dans les textes.

2.3Analyse des textes à l'aide du graphe de co-occurrences

2.3.1 L'organisation hiérarchique des graphes

L'intérêt du graphe ne réside pas dans la seule visualisation ergonomique qu'il propose. Il s'agit d'une structure mathématique pourvue de descripteurs permettant de caractériser numériquement et qualitativement les graphes, leurs nœuds et leurs liens. Nous nous intéressons ici spécifiquement à ceux qui révèlent son organisation hiérarchique[5].

Du fait des liens qu'ils nouent avec leurs voisins, et de leur position à l'échelle globale du graphe, certains lemmes ont une position centrale dans le réseau de co-occurrence. Parmi tous les indicateurs disponibles dans la littérature pour mesurer cette centralité, nous avons recours à la *centralité d'intermédiarité* (*betweenness centrality*, désormais BC). Notre choix se porte sur cette mesure pour trois raisons. En premier lieu, elle reflète l'intuition (Wasserman, Faust, 1994, 215) : les unités qui semblent les plus centrales à l'œil nu ont la BC la plus élevée. En second lieu, la BC est assez bien corrélée à la fréquence : un lemme fréquent a en général une BC élevée. La BC a cependant l'avantage de creuser les écarts entre les unités de fréquence similaire, et fait donc mieux apparaître les unités importantes. Enfin, elle renvoie à une forme pertinente d'organisation du lexique du texte : les unités ayant une BC élevée ont à la fois un rôle organisateur dans le graphe (du fait de leur position hiérarchique), et jouent un rôle d'intermédiaire entre les différentes notions du texte (Vergès, Bouriche, 2001, 69). Ce dernier aspect découle directement de la définition de ce paramètre.

La BC d'un nœud donné est en effet obtenue en additionnant la probabilité, pour tout couple de nœuds du graphe, que le nœud en question se trouve sur le plus court chemin reliant ces deux nœuds. Un nœud a donc une BC élevée si et seulement si beaucoup d'autres nœuds sont en relation directe avec lui ou sont obligés de passer par lui pour entrer en relation avec d'autres. Les nœuds dotés de la BC la plus élevée jouent donc un rôle constitutif dans le graphe de co-occurrence : c'est grâce à eux que se font la majorité des liens à échelle locale et à échelle globale. Un classement des nœuds par BC décroissante donne une image de l'organisation hiérarchique du lexique du texte. Une BC importante est le signe d'une position saillante dans le texte[6].

[5] Nous entendons par là que certains nœuds ont une position pré-éminente par rapport à d'autres. Sur la Figure 1, il s'agit notamment de *rubrique*, *article*, *développement* et *durable*. Soulignons que l'on pourrait chercher, indépendamment de cette organisation hiérarchique, d'autres formes d'organisation du graphe, notamment une organisation modulaire (Tauveron, 2012) que les paramètres que nous évoquons dans la section 5.2 permettent de caractériser.

[6] Contrairement à Boguraev, Neff (2000), notre définition de la saillance est purement interne au texte considéré, sans référence à une norme extérieure.

2.3.2 Une analyse longitudinale

L'originalité de notre étude réside dans le fait que nous nous focalisons sur l'évolution des graphes avec l'avancée du texte phrase après phrase[7]. Pour chaque texte, un premier graphe décrit la 1ère phrase, le 2ème graphe décrit les 2 premières, etc. Le texte entier n'est décrit que par le dernier graphe. Le choix d'une fenêtre s'élargissant à chaque étape (et non d'une « fenêtre glissante ») se justifie par le fait que nous travaillons sur un phénomène d'incrémentation. Une synthèse faite par un dernier programme Perl permet de comparer ces différentes étapes[8]. La comparaison est faite en relevant, dans l'évolution du texte au fil des étapes, les nœuds et les liens qui, au moins à un instant donné, ont une certaine prééminence. L'étude de l'ensemble de l'histoire de chacun de ces nœuds et de ces liens fournit une image d'ensemble des unités les plus saillantes, de ce point de vue, dans le texte. Ainsi, l'évolution du texte vue sous le seul angle de la co-occurrence généralisée permet d'étudier l'incrémentation lexicale qui a lieu dans le texte, non seulement sous son aspect quantitatif, mais d'étudier également sa construction.

3 Résultats

3.1 Trois types d'évolution

Dans notre corpus, nous avons mis en évidence trois grands types d'évolution d'ensemble de la co-occurrence généralisée.

Premier type, une croissance d'un bout à l'autre du texte : se créent sans arrêt de nouveaux nœuds et de nouveaux liens susceptibles de prendre une place prééminente assez rapidement.

Second type, une stabilisation progressive et unidirectionnelle du graphe : il commence par connaître une phase de croissance désordonnée, analogue au comportement précédent, mais cette croissance débouche rapidement sur un état stable, atteint en général avant la moitié du texte. Une fois ce palier atteint, le graphe n'évolue plus que dans le détail. C'est le comportement auto-organisé auquel nous avons fait référence en introduction ; il est d'ailleurs largement majoritaire dans notre corpus.

En troisième lieu, des cycles de stabilisation et de réorganisation successives. Le comportement précédent débouche subitement sur une nouvelle phase d'organisation au cours de laquelle est créé un nouvel état stable, différent du précédent.

Dans le corpus que nous avons utilisé, seul le second comportement est attesté de façon récurrente, comme le montre le tableau suivant :

[7] Nous ne pouvons revenir ici sur la question de la non-pertinence éventuelle de cette unité syntaxique dans l'étude de textes sur corpus, y compris dans le cas des corpus écrits.
[8] Ces synthèses sont représentées graphiquement à partir de la section 3.2.

Croissance permanente	3 textes	Textes brefs (Tous les textes de Rimbaud + un éditorial)
Stabilisation progressive et unidirectionnelle	33 textes	Textes brefs (éditoriaux) et tous les textes longs (scientifiques, encyclopédiques)
Cycles de stabilisation et de réorganisation	4 textes	Textes brefs (éditoriaux)

Tableau 1 – Répartition des textes selon les types d'évolution.

3.2 Les cas de stabilisation progressive et unidirectionnelle

3.2.1 Le corpus de textes brefs

Le début du texte[9] connaît éventuellement un sursaut de départ, phase brève (au plus un sixième du texte) au cours de laquelle les variations de centralité des nœuds et de poids des liens sont très rapides et imprévisibles, allant parfois d'un extrême à l'autre[10]. L'essentiel de l'évolution commence ensuite par une phase de stabilisation progressive. Les paramètres varient alors graduellement pour amener le texte à un certain stade organisé. Cette phase occupe en général entre un et deux cinquièmes du texte. Une fois atteint ce stade organisé, l'évolution est, qualitativement au moins, terminée. Les lemmes et les liens importants deviennent ensuite de plus en plus importants, à des vitesses différentes et imprévisibles. Ceux qui sont moins importants le restent, parfois en s'échangeant leurs places. Par contre, à quelques exceptions près, aucun nœud ou lien ne connaît d'important changement de classement passé ce stade. Il s'agit donc d'une phase d'aménagement de l'organisation. Le système que constitue la co-occurrence généralisée se donne au cours de la première phase une identité qui évolue ensuite sans se renier. Toute évolution ne fait que la raffermir, en accroissant le poids des unités importantes.

Dans notre corpus, un des cas les plus représentatifs est celui de l'éditorial du numéro de 1 de *Développement durable et territoire* (Figure 2, chaque courbe représente l'évolution d'un lemme au long du texte). Dans ce texte, la relation de co-occurrence généralisée a formé son identité pendant les 13 premières phrases du texte. La suite de l'évolution ne constitue qu'un réaménagement.

On peut mettre en évidence un phénomène analogue pour le comportement des liens. On voit sur les Figures 3 et 4 que le comportement des liens du graphe a la même histoire au cours de l'avancée du texte que celui des nœuds[11]. À ceci près que la construction de l'état

[9] Les différentes étapes sont qualitativement les mêmes dans tous les textes relevant de ce comportement. Seules changent la durée respective des phases et leur netteté.

[10] Cette forte variabilité s'explique par le fait que la mesure en question porte sur une portion du texte très réduite, et donc d'autant plus sujette à des variations aléatoires. On rencontre le même problème qu'avec un calcul de fréquence.

[11] Pour des raisons de lisibilité, nous répartissons les liens du graphe en 3 catégories. Les liens les plus importants (respectivement médians, moins importants) sont ceux qui, au moins une fois dans toute leur évolution, ont un poids qui dépasse 60% du poids maximal (respectivement 40%, 30%).

organisé du graphe a lieu sur une période plus étendue que pour les nœuds, à savoir entre la 11ème et la 15ème phrase.

Dans les détails, des différences parmi les Figures 3 et 4 révèlent un aspect de la dynamique du phénomène : moins les liens sont importants et plus ils évoluent vite au cours de la première phase. En effet, on constate pendant cette phase 4 montées abruptes sur le graphique des liens les plus importants et 10 sur le graphique suivants. C'est-à-dire que les liens les plus importants sont plus stables, et les liens les moins importants sont plus volatiles. Le poids des liens dénote donc la solidité et le caractère architectural des relations au cours d'une évolution. La relation de co-occurrence généralisée apparaît donc comme un véritable système, dont les parties les plus significatives sont pérennes, et les parties les plus légères sont adaptables au cours du temps[12].

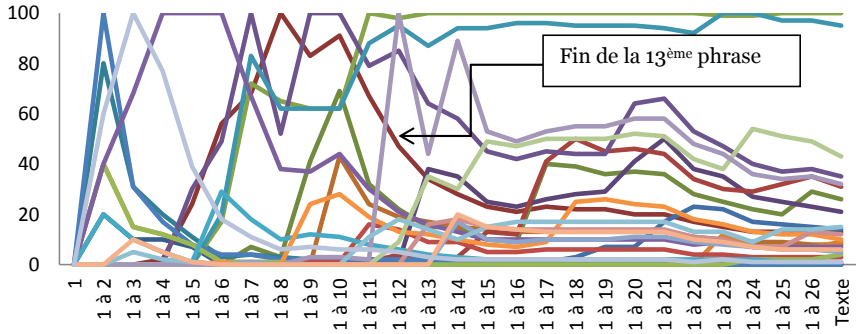

Figure 2 - Évolution de la BC des lemmes dans DDT-1

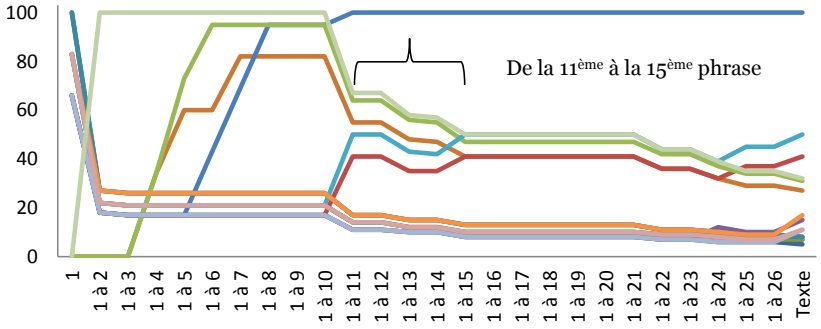

Figure 3 - Poids des liens les plus importants dans DDT-1-1

[12] On remarque sur Figure 4 que de nouveaux liens gagnent en importance dans la deuxième partie de l'évolution. Malgré le caractère unidirectionnel de ces évolutions, les parties les plus volatiles du graphe peuvent connaître des phases de croissance ponctuelles.

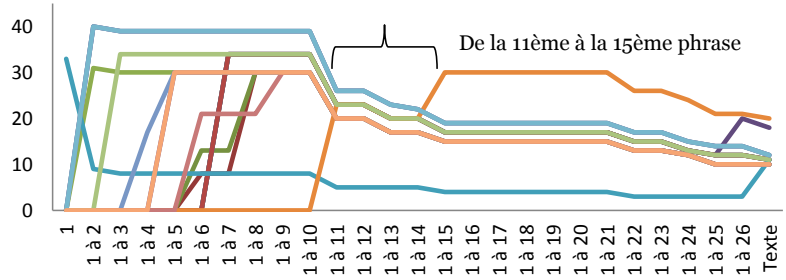

Figure 4 - Poids des liens les moins importants dans DDT-1-1

3.2.2 Le corpus de textes longs

Le comportement attesté sur la majorité des textes brefs du corpus de travail est également constatable sur l'ensemble du corpus de textes longs[13]. Il est même plus net sur ce second corpus. En effet, la phase initiale de stabilisation peut être – en proportion – largement plus courte, n'occupant qu'un dixième du texte dans certains cas. Il semble en particulier que plus le texte est long, plus la phase de stabilisation est courte en proportion. On le voit en particulier sur la Figure 5, représentant l'évolution d'un texte de 14.601 mots, divisé en 40 étapes de 15 phrases (article « Beethoven » de *Wikipédia*), et dans laquelle il apparaît bien que l'évolution de la BC des nœuds est finie dès la 4ème étape.

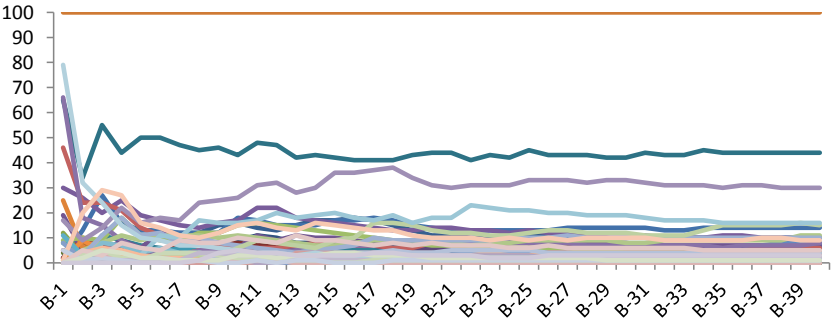

Figure 5 – BC des nœuds dans « Beethoven »

Le retard de l'évolution des liens par rapport à celle des nœuds est là encore constatable (Figure 6) : la phase d'organisation des liens du graphe s'étend jusqu'à l'étape 10, c'est-à-dire jusqu'à la fin du 1er quart du texte.

[13] Pour des raisons de lourdeur de traitement informatique, l'incrémentation est étudiée par paliers de 10 à 15 phrases, et non plus d'une seule. Le texte le plus long du corpus (22.500 mots) compte en effet 1.027 phrases.

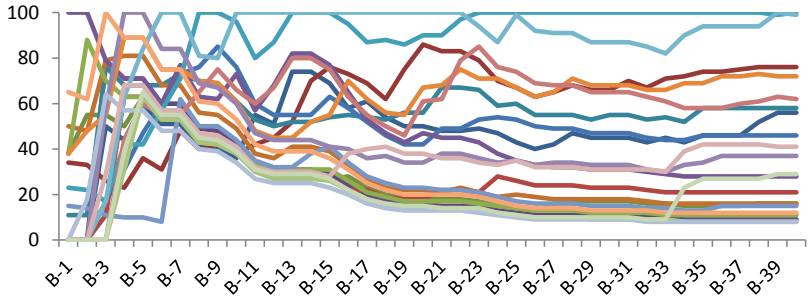

Figure 6 – Poids des liens les plus importants dans « Beethoven »

3.3 Les cas cycliques

Comme évoqué plus haut, nous avons pu mettre en évidence un fonctionnement cyclique sur un petit nombre de textes. Dans ces cas, l'évolution décrite en 3.2 se produit deux fois : une fois un état stable atteint, et maintenu pendant un laps de temps assez long, arrive une phase de réorganisation vers un autre état stable, maintenu lui aussi. On voit que le texte *DDT2-2* (Figure 7) alterne phases d'organisation (6 à 9, 21 à 30) et phases de stabilité (9 à 21, 30 à 41). La croissance de certaines unités qui a lieu lors de cette dernière phase (les deux courbes vertes correspondent à *politique* et à *culture*) ne bouleverse pas l'organisation d'ensemble. Par ailleurs, l'état du graphe lors des deux périodes de stabilité est bien distinct : c'est par exemple le lemme *notion* qui domine entre 9 et 21, alors que c'est *développement* qui domine après 30.

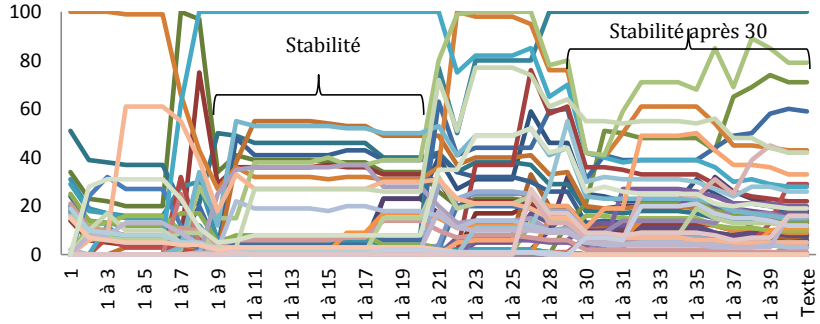

Figure 7 - Évolution de la BC des nœuds dans DDT2-2

Un retour à la lettre du texte permet de voir que ces deux phases de stabilité correspondent à des changements de thématique. La première partie du texte est consacrée à la résilience, qualifiée 5 fois de *notion*. Une phrase dénote ensuite un abandon de cette thématique (« *il semblerait qu'il n'y ait qu'un pas de la résilience à la*

résistance dans le cadre du développement durable ») et un passage à la thématique du développement durable (on trouve plus loin « *[é]videmment difficile de définir ici et en une phrase ce qu'est le développement durable* »)[14].

3.4Les cas de croissance permanente

Nous avons mené sur un corpus de contrôle fait de deux poèmes d'Arthur Rimbaud la même étude que sur le corpus de travail (Figure 8 : poème *Enfance*). Elle met en évidence un fonctionnement totalement différent des deux présentés plus haut, que l'on retrouve, de manière moins nette, sur un des textes du corpus (*Vertigo*3-1). Dans ces textes, la co-occurrence généralisée croît en permanence : au fur et à mesure qu'apparaissent de nouvelles phrases et de nouveaux mots, apparaissent également de nouveaux lemmes et de nouvelles relations de co-occurrence. On n'identifie pas de phase de stabilité à un quelconque moment. Comme sur le cas précédent, la plus haute valeur en BC est prise successivement par des lemmes différents, sans cette fois que cette alternance ne reflète des phases organisées.

Les trois textes marginaux qui connaissent ce comportement constituent en quelque sorte des cas limites. Dans les deux comportements organisés décrits plus haut, l'organisation en partie hiérarchique des graphes servait en effet à assurer la cohésion du réseau de co-occurrence généralisée, et permettait donc l'émergence d'une forme d'ensemble. Les nœuds les plus importants construisaient leur position de prééminence au cours d'une évolution mesurée, pour la garder par la suite. Au contraire, dans les cas de croissance permanente, aucun état stable n'est maintenu au cours du temps, ce qui n'empêche pas de considérer qu'on a affaire à une certaine forme d'organisation.

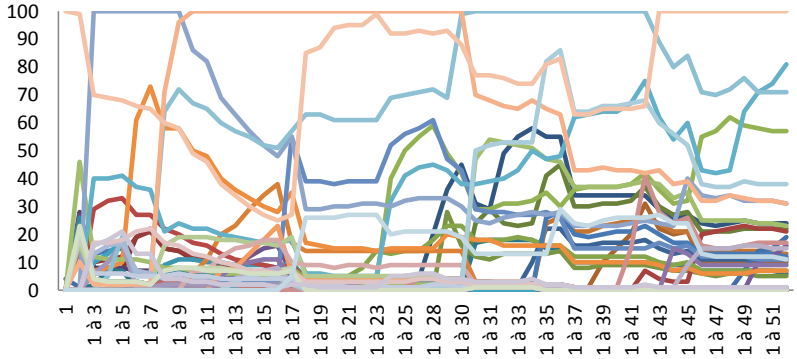

Figure 8 - Évolution de la BC des nœuds dans *Enfance* de Rimbaud

[14] Il faut noter cependant que le changement de thématique n'intervient pas à l'instant précis du changement de phase que nous identifions. Le grand nombre d'occurrences de *notion* au début du texte donne une certaine inertie à ce terme . Dans la mesure que nous faisons, sa position prééminente se maintient après le changement de thématique.

4 Discussion

4.1 Un système auto-organisé

Il est surprenant que le graphe ne connaisse pas systématiquement une croissance permanente d'un bout à l'autre du texte. Dans les faits, le cas normal est celui d'une stabilisation progressive et unidirectionnelle du graphe de co-occurrence[15]. Ce fonctionnement particulier peut s'expliquer en référence à un mécanisme général d'évolution de certains systèmes, appelé *auto-organisation* (van de Vijver, 2004).

Les observations faites montrent que l'évolution de la co-occurrence généralisée ne se fait pas au hasard, et ne prend pas la forme d'une croissance permanente. En effet, au fur et à mesure de l'avancée du texte, les mots et les relations de co-occurrence ne sont pas ajoutés de manière effrénée, ni placés anarchiquement mais sont disposés de manière systématique. Ce placement passe par une phase de stabilisation qui a lieu au début du texte, avant d'atteindre un état certes relativement stable, mais surtout susceptible de se réajuster pour permettre l'intégration des nouveaux contenus. C'est la force de cette structure : elle permet de canaliser le flux d'information continu dans la deuxième partie du texte. Elle repose sur la présence d'une certaine hiérarchie entre unités – les lemmes et les liens les plus importants, du fait de leur rôle architectural, étant moins volatiles. C'est cette organisation qui assure en premier lieu la stabilité du système et son existence (Ladrière, 2009), malgré la sollicitation (et le risque de désorganisation) que représente l'apport de nouveaux contenus. L'ensemble du processus montre donc une *construction progressive de l'identité du réseau de co-occurrence généralisée* (définitoire de l'auto-organisation chez Moreno, 2004). Ce réseau élabore donc son identité de système au cours d'une première phase. Par la suite, même si ce système ne reste pas figé et évolue, c'est sans modifier cette identité. Comme on l'a vu, comprendre l'évolution du système suppose de le considérer dans sa globalité. Cette évolution d'ensemble dépasse l'histoire de chacun des nœuds et des liens, et est imprévisible à partir de leurs évolutions ponctuelles (comme le montre le passage brusque de la phase chaotique initiale à la phase organisée). Par définition, l'organisation de la co-occurrence généralisée est donc *émergente*. Nous allons montrer de surcroît qu'elle semble *prévue* (par le scripteur du texte) *pour être identifiée* (par le lecteur) – il s'agit donc d'une *auto-organisation au sens faible* dans la typologie d'Atlan (2011, 194)[16].

Si on peut mettre le phénomène que nous avons constaté ici sur le compte d'une propriété générale de certains systèmes, nous allons montrer maintenant qu'il repose sur des contraintes et ressources de la cognition humaine.

[15] L'interprétation faite ici s'applique également, dans une moindre mesure, sur les textes présentant une évolution cyclique.

[16] Ces systèmes s'opposent chez l'auteur aux auto-organisations au sens fort, que sont par exemple les êtres vivants, dans lesquelles la structure a une finalité et une signification qui tirent leur origine de l'intérieur du système. De tels systèmes ont donc une évolution autonome. Un texte étant produit pour être lu, sa finalité et son fonctionnement sont nécessairement à mettre sur le compte du scripteur.

4.2Interprétation en termes cognitifs

4.2.1 Contraintes du système cognitif

Van Dijk, Kintsch (1983) ont supposé l'existence de mécanismes cognitifs permettant la structuration de l'information rencontrée dans les textes. Parmi eux, la suppression d'informations secondaires, et le regroupement de plusieurs informations en une seule. On peut supposer que ces mécanismes obéissent à la nécessité d'organiser le sens (et, indirectement, le lexique) du texte. Une des conséquences de ces phénomènes est la présence de thèmes (au sens de Hjørland, 2001), qui structurent le contenu sémantique du texte de manière notamment hiérarchique.

Le comportement auto-organisé que nous avons décelé révèle un des aspects de cette contrainte. En effet, le maintien de la prééminence d'unités importantes et la relative rigidité des liens les plus forts ont un rôle architectural dans la co-occurrence généralisée similaire à l'économie des thèmes dans un texte. Il apparaît donc que le caractère auto-organisé de la co-occurrence généralisée permet l'application des opérations de suppression et de structuration de l'information.

4.2.2 Ressources du système cognitif

Par ailleurs, le fonctionnement auto-organisé décrit ici n'est possible que grâce à certaines ressources du système cognitif. Deux des opérations cognitives fondamentales postulées par Langacker (1987, 116-140) – la sélection et la transformation – rendent ce type d'évolution possible.

Le fonctionnement des textes tel que nous l'avons montré repose en effet sur la capacité qu'a le système cognitif à sélectionner les unités et les relations les plus saillantes pour focaliser son attention sur elles et les placer au centre de sa construction du sens textuel. La modification permanente du système auto-organisé repose sur ses capacités de transformation : les représentations formées sont sans cesse déformées pour accepter de nouvelles informations arrivant certes par incrémentation successive, mais, et c'est notre conclusion, de manière structurée.

Le fonctionnement de ces ressources contribue à expliquer la difficulté à la lecture des deux textes poétiques de Rimbaud. L'impression de désordre qu'on peut avoir à leur sujet est démontrée ici, sur son seul versant lexical. Comme on l'a vu, ces textes font partie de ceux qui ne témoignent pas d'une stabilisation progressive : les opérations que sont la sélection et la transformation d'informations par le système cognitif sont dans ces cas inopérantes et ne permettent pas facilement l'élaboration d'un sens d'ensemble.

5 Conclusions

5.1 Perspectives d'application

Nous avons considéré ici (à la manière d'Utiyama, Isahara, 2001) que les thèmes des textes se définissaient par des relations de co-occurrence (fréquence des liens par

l'intermédiaire de leur poids, centralité). En suivant les idées de Hearst (1997) ou Ferret (2007), on pourrait appliquer ce travail en premier lieu à des fins de segmentation thématique. Nous avons en effet montré dans certains cas que notre méthode permettait de diviser les textes en différentes sections. Elle montre également dans d'autres cas (croissance permanente) que certains textes ne sont pas segmentables thématiquement.

Erkan, Radev (2004) et Xie (2005) ont par ailleurs montré que la centralité dans le lexique des textes permettait de mettre en évidence la structuration hiérarchique des informations, et donc de repérer les éléments qui doivent figurer dans leur résumé. La spécificité de notre méthode – envisager le texte dans sa dynamique – permettrait de repérer les passages qui apportent le plus d'informations. Si une telle étude confirme le fait que les informations essentielles se situent souvent au début des textes, nous avons pu montrer qu'il ne s'agit pas d'une généralité (dans les cas de croissance permanente).

En troisième lieu, la méthode présentée ici reposant sur le seul examen du lexique, c'est dans l'indexation des documents qu'elle connaîtrait ses meilleures applications. Elle pourrait permettre d'identifier les mots les plus essentiels du texte selon les passages et d'identifier leurs relations mutuelles mieux que ne le ferait une liste de mots-clés.

Enfin, l'examen des spécificités de l'organisation lexicale des textes permettrait de faire un pas dans la caractérisation du style et du genre d'un texte (voire d'un auteur)[17].

5.2 Nouvelles perspectives de recherche

La méthode présentée ici se doit d'être affinée et précisée en recourant à d'autres paramètres mathématiques donnant une meilleure description qualitative de la position des nœuds dans les graphes (*eigenvector centrality, clustering coefficient*, classes de modularité, etc.). L'effet de la taille de l'empan définissant la co-occurrence, d'un éventuel élagage du graphe ou de la lemmatisation restent également à caractériser en détail.

Références

ADAM, J.-M., (2004), *Linguistique textuelle. Des genres de discours aux textes*, Paris, Nathan.

ATLAN, H., (2011), *Le Vivant post-génomique ou qu'est-ce que l'auto-organisation ?*, Paris : Odile Jacob.

BOGURAEV, B., NEFF, M., (2000), "Lexical Cohesion, Discourse Segmentation and Document Summarization," *Proceedings of RIAO'2000*, Paris (12–14 avril 2000).

ERKAN, G., RADEV, D., "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", *Journal of Articial Intelligence Research*, 22, 457-479.

FERRET, O. 2007. "Finding document topics for improving topic segmentation". *ACL 2007*. 480-487.

[17] Cette suggestion de Tristan Vanrullen portera tous ses fruits dans un prochain travail.

HEARST, M. 1997. “TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages”. *Computational Linguistics*, 23, 1, 33-64.

HJØRLAND, B., (2001), “Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content… and Relevance”, *Journal of the American Society for Information Science and Technology*, 52, 9, 774–778.

LADRIERE, J., (2009), « Système, épistémologie », *Encyclopaedia Universalis*.

LANGACKER, R., (1987), *Foundations of Cognitive Grammar. Theoretical Prerequisites*, Stanford : Standford UP.

LEGALLOIS, D., (2006), « Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison », *Langages*, 163, 3-9.

LEMAIRE, B., (2008), « Limites de la lemmatisation pour l'extraction de significations », *JADT 2008*, 725-732.

MORENO, A., (2004), « Auto-organisation, autonomie et identité », *Revue internationale de philosophie*, 228, 135-150.

PARANYUSHKIN, D., (2010), « Text network analysis », Conférence du *Performing Arts Forum*, http://noduslabs.com/research/pathways-meaning-circulation/, (14.09.2011).

TAUVERON, M., (2012), « Variation du sens lexical en discours : la co-occurrence généralisée valide une non-correspondance entre deux langues ». La cooccurrence : du fait statistique au fait textuel. Besançon, 9 février 2012.

UTIYAMA, M. ISAHARA, H.. 2001. “A statistical model for domain-independent text segmentation”. In *ACL’01*, 491–498.

VAN DE VIJVER, G., (2004), « Auto-organisation, autonomie, identité : Introduction », *Revue internationale de philosophie*, 228, 129-133.

VAN DIJK, T. KINTSCH, W., (1983), *Strategies of Discourse Comprehension*, Orlando : Academic Press.

VERGES, P., BOURICHE, B., (2001). “L’analyse des données par les graphes de similitude”. *Sciences humaines*, http://www.scienceshumaines.com/textesInedits/Bouriche.pdf.

VERONIS, J., (2004), “HyperLex : Lexical Cartography for Information Retrieval”, *Computer Speech & Language*, 18, 3, 223-252.

VIPREY, J.-M., (2006), “Structure non-séquentielle des textes”, *Langages*, 163, 71-85.

WASSERMAN, S., FAUST, K., (1994), *Social Network Analysis*, Cambridge UP.

XIE, Z., (2005), “Centrality Measures in Text Mining : Prediction of Noun Phrases that Appear in Abstracts”. *ACL’05, Proceedings of the Student Research Workshop*. Ann Arbor.