RÉCITAL 2005, Dourdan, 6-10 juin 2005

Cent mille milliards de poèmes et combien de sens ? Une étude d'analyse potentielle

Florentina Vasilescu Armaselu
Université de Montréal, Département de littérature comparée
armaselu@sympatico.ca

Mots-clefs – Keywords

Unité du discours, réseaux de cohésion, analyse thématique, littérature potentielle

Discourse unity, networks of cohesion, thematic analysis, potential literature

Résumé – Abstract

A partir du concept de *cohésion* comme mesure de l'*unité du texte* et du modèle oulipien de la *littérature par contraintes*, notre étude propose une méthode d'*analyse potentielle* sur ordinateur dans le cas des *Cent mille milliards des poèmes*. En s'appuyant sur un ensemble de contraintes initiales, notre programme serait capable d'analyser tous les textes potentiels produits par la machine en utilisant ces contraintes.

Using the concept of *cohesion* as a measure for the *unity of text* and the Oulipian model of the *literature by constraints*, our study proposes a computational method of *potential analysis* for *One hundred billions sonnets*. Starting from a set of initial constraints, our program would be able to analyze all the potential texts produced by the machine under these constraints.

1 Introduction

Les mécanismes de compréhension du *sens* d'un texte dans son ensemble restent encore peu connus. Des modèles du processus d'interprétation (comme compréhension, pas comme interprétation critique) ont été déjà proposés : la théorie des cadres (Minsky, 1975), la hiérarchisation des expressions anaphoriques (Lakoff, 1976), les réseaux de cohésion (Halliday, Hasan, 1976), l'hypothèse de la connectivité (Gentner, 1981). (Rastier, 1987) relie la notion de *sens* d'un texte à la modalité de perception de *l'unité du texte,* dans le processus d'interprétation. A partir de l'étude des réseaux de cohésion (Stoddard, 1991) et de la notion de cohésion lexicale (Morris, Hirst, 1991) nous proposons une nouvelle approche d'analyse sur ordinateur des réseaux de cohésion comme mesures de l'unité d'un texte, dans le cas des *Cents mille milliards de poèmes* (Queneau, 1961). Notre analyse s'appuie sur des relations d'ordre sémantique, syntaxique et cognitif qui contribuent à la perception d'un texte comme un tout cohésif. Ne disposant pas d'un dictionnaire électronique comportant ce type d'information, nous avons annoté manuellement (voir 4.1) les mots considérés significatifs (noms, verbes, adjectifs) des dix sonnets originaux de Queneau. A partir de cet ensemble d'annotations et en utilisant un mécanisme combinatoire permettant l'engendrement de nouveaux sonnets, notre programme de composition et d'analyse produit des diagrammes, des calculs estimatifs et des descriptions thématiques (voir 4.1, 4.2, 5), en simulant la « compréhension » d'un sonnet en termes d'unité thématique et de relations de cohésion établies entre les mots.

Le choix des *Cents mille milliards de poèmes* comme banc d'essai pour notre étude n'a pas été aléatoire. Premièrement, parce que le modèle oulipien de la création par contraintes et son mécanisme combinatoire (voir 2), représenteraient un point de départ pour une *analyse potentielle* du texte. Nous entendons par cela un programme qui, à partir d'un ensemble de contraintes initiales (dans notre cas, les dix sonnets originaux annotés), serait capable d'*analyser*, par des procédés combinatoires, tous les *textes potentiels* produits par la machine en utilisant ces contraintes. Deuxièmement, les dix poèmes originaux, doués, selon Queneau, d'un « thème » et d'une « continuité », joueraient le rôle de *base de comparaison* pour notre analyse. Troisièmement, en tenant compte que notre objet d'étude est une collection de sonnets, notre intérêt porte sur les enjeux d'un traitement automatique du *sens* dans le cas des textes littéraires. En d'autres mots, il s'agissait de reformuler en termes interrogatifs la bien connue citation de Turing placée par Queneau en tête de ses *Cent mille milliards de poèmes* : *Est-ce qu'une machine peut apprécier un sonnet écrit par une autre machine ?*

2 Une machine à fabriquer des sonnets

Les deux tendances majeures de la recherche oulipienne sont la reprise des œuvres du passé et l'invention de nouvelles règles de création littéraire, à partir de *contraintes formelles* (Le Lionnais, 1986). Selon (Motte W.F. Jr., 1986), les *Cent mille milliards de poèmes* représentent le modèle de l'entreprise oulipienne, par la reprise d'une forme poétique traditionnelle, le sonnet, et par l'invention d'une nouvelle forme poétique *combinatoire* permettant à chaque vers d'être intégré dans l'ensemble quasi-infini de sonnets potentiels. Comme l'affirme (Queneau, 1961), *Cent mille milliards de poèmes* est une « machine » à fabriquer 10^{14} poèmes différents. Le fonctionnement de cette machine s'appuie sur un ensemble de contraintes formelles, internes et combinatoires. Les *contraintes internes* concernent la forme de chaque sonnet (deux quatrains et deux tercets), les rimes qui ne doivent pas « être trop banales [...] trop rares ou uniques » et l'existence d'un « thème » et d'une « continuité » pour chacun des dix sonnets d'origine. Les *contraintes combinatoires* exigent une structure grammaticale invariante et l'absence des désaccords en genre et en nombre pour toute substitution de vers possible. En tenant compte de cette immense mais encore limitée *potentialité créative*, qu'est-ce qu'on pourrait dire alors sur la *potentialité de sens* de ce type de machine ?

3 Cohésion et unité du texte

Le concept de *sens* fait l'objet d'étude de plusieurs disciplines dans le cadre des sciences humaines. Comme nous avons déjà mentionné, notre démarche s'intéresse seulement à une partie plus restreinte de ce concept, reliée à la modalité par laquelle nous percevons *l'unité d'un texte* dans le processus d'interprétation (Rastier, 1987). Selon (Morris, Hirst, 1991), le texte ou le discours n'est pas une simple succession de mots et de phrases faisant référence à des choses différentes, mais un ensemble d'entités reliées l'une à l'autre qui portent sur un même sujet. C'est une propriété qui confère au texte la qualité d'*unité* et qui est appelée *cohésion.* La cohésion n'est pas pourtant une caractéristique inhérente au texte, elle dépend aussi de lecteur. Dans l'acception de (Stoddard, 1991), la cohésion est un mécanisme unificateur que nous construisons pendant le processus d'interprétation et qui nous aide à dériver beaucoup plus de sens du texte dans son ensemble que de la simple somme des sens des mots et des phrases qui le composent. La cohésion impliquerait ainsi la construction de liens mentaux entre les parties composantes d'un texte, dans le processus d'interprétation.

Il y a plusieurs types de relations déterminant la cohésion. Notre étude s'intéresse aux relations *sémantiques* existant entre les mots (partie/tout, co-occurrence dans des contextes similaires, appartenance à un même domaine) et déterminant la *cohésion lexicale* (Morris, Hirst, 1991). De plus, notre analyse s'appuie sur le modèle des *réseaux de cohésion* (Stoddard, 1991), utilisé dans l'analyse des articles définis, des pronoms et des dislocations d'agents verbaux. Le réseau de Stoddard comporte un *nœud* (le référent, par exemple *Abraham Lincoln*) et des *éléments de cohésion* (par exemple, les pronoms *he, him, his*) reliés au nœud par des *relations de cohésion* sémantico-syntaxiques. Stoddard fait ainsi une distinction entre la *cohésion*, une caractéristique sémantico-syntaxique, et la *cohérence* une mesure de « l'unité de sens d'un texte » qui entraîne « l'environnement cognitif » et « l'expérience » du lecteur. Nous avons adapté ce modèle, en considérant le réseau de cohésion comme une structure de *nœuds* (noms, verbes, adjectifs préalablement annotés) reliés entre eux par des *relations de cohésion* qui dépendent de la nature des attributs attachés aux nœuds et qui impliquent des connaissances d'ordre lexico-sémantique, syntaxique et cognitif (voir 4.1).

4 Ouvroir d'analyse potentielle

Le programme permet à la fois la construction et l'analyse d'un poème. Il y a deux modalités de construire un poème : choisir un des dix sonnets originaux en indiquant son numéro dans un champ de saisie ou composer un nouveau poème, en combinant les vers des dix sonnets à l'aide de 14 listes déroulantes. Le module d'analyse utilise les attributs attachés manuellement aux mots, comme des *contraintes initiales.* Après la composition d'un sonnet tel indiqué ci-dessus, le programme compare les attributs et construit un lien entre deux mots (nœuds), s'il y détecte au moins une valeur commune, indifféremment du type des attributs.

4.1 Les contraintes initiales

L'annotation des mots (en format XML) comporte un attribut obligatoire (le *lemme)* et un ensemble d'attributs optionnels[1] (voir Figure1), selon les quatre types de relations considérés :

1. Relations de type *sémantique,* décrites par l'attribut *domaine,* une classe sémantique reliée à « l'expérience d'un groupe » et qui encode une « pratique sociale » (Rastier, 1997). Ce type d'attribut permet de construire, par exemple, un lien entre *Tamise* et *bateaux* (domaine = navigation), *climat* et *bise* (domaine = météo), *Socrate et Platon* (domaine = philosophie), etc.
2. Relations de type *syntaxique* entre un nom et son complément (ou son attribut) et un verbe et son complément. Ces relations sont mises en évidence par l'attribut *relatif_à* associé au complément ou à l'attribut : *« climat londonien »* (londonien, relatif_à = climat) ; *« Sa sculpture est illustre »* (illustre, relatif_à = sculpture) ; « on transporte et le marbre …» (marbre, relatif_à = transporter).
3. Relations extratextuelles supposant des *connaissances du monde*, définies par les attributs *appartenance* et *allusion.* Le programme mettrait ainsi en relation *Grèce* (lemme = Grèce) avec *Platon* (appartenance = Grèce) ; *londonien* avec *Tamise* (appartenance = Angleterre) ; *Elgin* (allusion = Parthénon, Turquie) avec *Parthénon* (lemme = Parthénon) et *Turc* (appartenance = Turquie) ; *frissonner* avec *bise* (allusion = froid), etc. A la différence de (Stoddard, 1991) nous avons considéré ces connaissances du monde comme facteurs déterminant la cohésion du texte.
4. Relations *étymologiques* (*gaucho, pampa, maté* reliés par leur attribut *étymologie* = espagnol).

[1] Leurs valeurs ont été suggérées par *Le Petit Larousse, Le Grand dictionnaire terminologique* et *EURODICAUTOM.*

Comme dans le modèle oulipien, l'annotation s'appuie sur des contraintes initiales, internes et combinatoires (voir 2), supposant une interprétation au niveau de chacun des sonnets originaux et une sorte de méta-interprétation qui devrait tenir compte des combinaisons possibles des mots d'un sonnet avec les mots des autres. Il s'agissait ainsi de prévoir des valeurs d'attribut appropriées de façon que *Rameaux* (sonnet 3) soit par exemple relié à *cloche* (sonnet 1), *corne* (sonnet 1) soit relié à *taureau* (sonnet 1) et à *veau* (sonnet 5) mais pas à *chat* (sonnet 9) ou à *baleine* (sonnet 3), dans un sonnet potentiel. Un exemple de diagramme de cohésion produit par le programme est présenté ci-dessous :

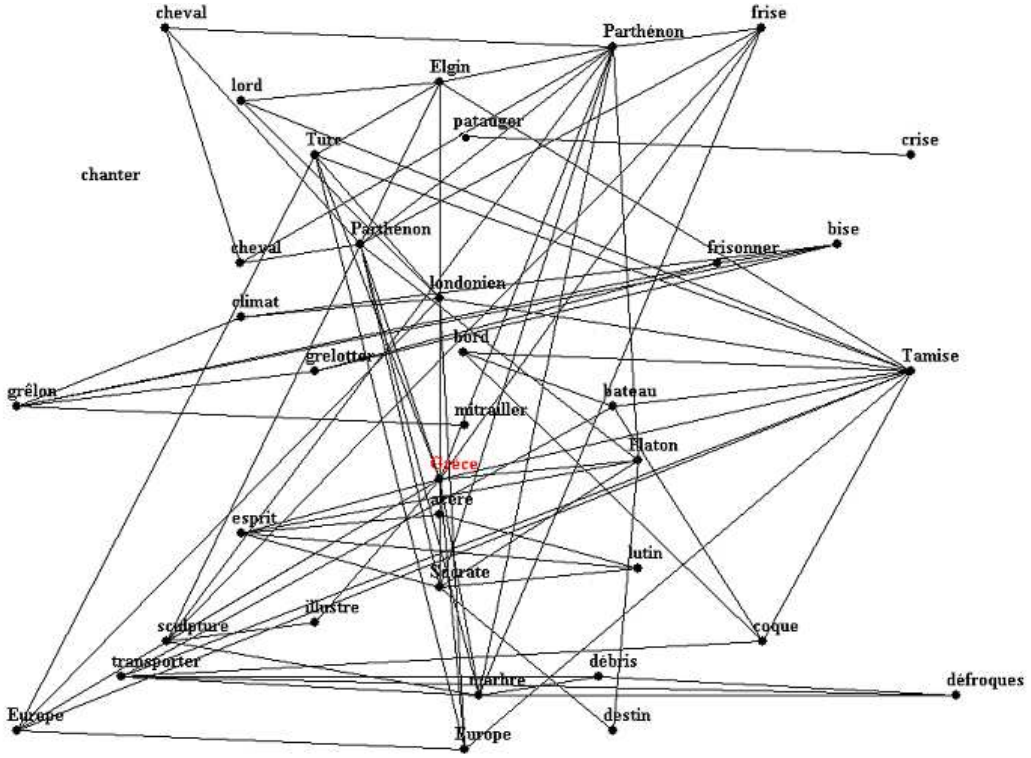

Figure 1. Réseaux de cohésion. Sonnet 2

4.2 Le banc d'essai

Les *Cent mille milliards des poèmes* nous ont servi également comme banc d'essai pour des estimations quantitatives (sur le contenu, la cohésion, le thème), considérées comme des indicateurs globaux de l'unité d'un poème. Le programme détermine la composition (en %) de chaque poème analysé, par rapport aux sonnets originaux (par exemple : 100% sonnet2 ; 50% sonnet1 +50% sonnet2, etc). C'est une mesure par laquelle on pourrait apprécier, en grandes lignes, le caractère homogène ou hétérogène d'un poème quant à sa composition. A partir de l'hypothèse qu'un texte nous paraîtrait plus unitaire si la plupart de ses mots sont reliés entre eux, nous avons défini le *coefficient de cohésion globale* (CCG) comme :

$$CCG = \frac{NRD}{TN} \cdot 100$$, où NRD = le nombre de Nœuds du Réseau de cohésion Dominant,

TN = le nombre Total des Nœuds pour le texte analysé.

Comme un texte pourrait renfermer plusieurs réseaux de cohésion indépendants (ce qui indiquerait une fragmentation de son unité), le réseau dominant regrouperait le plus grand

nombre de nœuds connectés entre eux par des liens de cohésion, pour le texte donné. Un coefficient de cohésion de 100% caractériserait ainsi un texte où tous les mots analysés soient reliés entre eux, en formant un seul réseau. Le sonnet 2 (voir Figure 1) présente un coefficient de cohésion globale assez élevé, puisque son réseau de cohésion dominant inclut presque tous les nœuds, sauf trois (*chanter, patauger, crise*). Une autre mesure, utilisée comme indicateur du degré de cohésion entre les mots, est la *densité moyenne* (DM), i.e. le nombre moyen de connexions par nœud pour un texte donné (Stoddard, 1991). Une valeur élevée de la densité signalerait une cohésion forte, une valeur basse indiquerait une cohésion faible (un texte ayant moins de relations entre ses éléments). Comme base de comparaison, le programme affiche les valeurs du CCG et de la DM des dix sonnets originaux, par ordre décroissant.

Le programme propose également un *thème,* i.e. l'entité à laquelle les mots du texte font le plus souvent référence et qui appartient au texte (valeur de l'attribut *lemme*) ou est extérieure au texte (valeur d'un attribut *domaine*, *allusion*, *appartenance*, etc). Le programme compte le nombre de fois qu'une valeur d'attribut apparaît dans l'annotation d'un sonnet, en choisissant comme thème la ou les valeurs les plus fréquentes. Pour une caractérisation plus détaillée, chaque thème est accompagné des *centres de focalisation* du sonnet, i.e. les mots comportant le plus grand nombre de connexions (par exemple, *Grèce* pour le sonnet 2, Figure 1). Cette description thématique serait une représentation condensée du sens global d'un texte (voir 5).

5 Observations sur les résultats

La Figure 2 présente les résultats d'analyse pour les dix sonnets de départ et pour dix sonnets composés. Le tableau indique une cohésion moins forte (densité plus basse) pour les sonnets composés, bien que des valeurs plus élevées soient également possibles (No 12). Le coefficient de cohésion globale présente des valeurs variables, comparables parfois avec les sonnets originaux. Les valeurs basses de cet indicateur montreraient l'existence de plusieurs réseaux de cohésion, dont aucun ne domine de façon marquante, et alors une fragmentation de contenu (14, 19). Cette caractéristique ne semble pas reliée nécessairement au caractère trop hétérogène d'un sonnet, une combinaison de tous les dix sonnets originaux pouvant déterminer des valeurs assez élevées (20). Comme le montre le tableau, les descriptions thématiques obtenues par le mécanisme combinatoire d'analyse peuvent reprendre les thèmes initiaux, avec un changement éventuel de centre de focalisation (11), engendrer de nouveaux thèmes (14, 15, 16, 19, 20) ou des thèmes composites (12, 13, 16, 17, 18). Dans ce dernier cas on pourrait avoir parfois une certaine compatibilité entre les éléments y impliqués (12, 16), parfois un caractère plus hétérogène, bien que pas tout à fait incompatible (13, 17, 18). D'un autre côté, les résultats d'analyse semblent fort dépendants de la description XML, i.e. de la subjectivité et du niveau de connaissances utilisés dans l'interprétation des sonnets originaux.

No	Sonnet/Contenu	CCG %	DM	Thème	Centres de focalisation (no. liens/centre)
			Sonnets originaux		
1	S1	*62.8*	*2.7*	*Amérique du Sud*	*Amérique du Sud; pampa (8)*
2	S2	*91.8*	*4.7*	*Europe*	*Grèce (14)*
3	S3	*72.5*	*9.5*	*mer*	*poisson; dorade; molve_lotte (20)*
4	S4	*72.5*	*2.8*	*noblesse*	*blason; baron (8)*
5	S5	*92.3*	*4.0*	*Europe*	*latin (13)*
6	S6	*96.6*	*2.8*	*urbanisme*	*escroc; provincial (6)*
7	S7	*62.0*	*3.9*	*généalogie, famille*	*généalogiste; adultérin; parent (11)*
8	S8	*92.5*	*6.5*	*langues*	*métromane (18)*
9	S9	*89.1*	*9.6*	*alimentation*	*turbot; requin (25)*
10	S10	*62.5*	*3.3*	*mort*	*mort (12)*

No	Sonnet/Contenu	CCG %	DM	Thème	Centres de focalisation (no. liens/centre)
			Sonnets composés		
11	**50%**S6,S10	*71.4*	*2.3*	*mort*	*croque-morts; tissu; pâlotte (6)*
12	**50%**S3,S9	*95.1*	*9.9*	*alimentation, zoologie*	*poisson; dorade; molve lotte (25)*
13	**50%**S1,S4	*62.1*	*2.4*	*géographie, Angleterre, noblesse*	*baron ; Malabar ; lord (6)*
14	**28.5%**S4,S8 ; **21.4%** S2,S6	*26.1*	*2.1*	*Angleterre*	*Tamise (8)*
15	**35.7%**S7,S8 ; **28.5%**S9	*70.5*	*2.5*	*psychologie*	*idiot ; métromane (6)*
16	**35.7%**S1; **21.4%**S3,S7,S2	*63.6*	*2.1*	*navigation, eau*	*marin (7)*
17	**28.5%**S5,S1 ; **21.4%**S2,S9	*65.8*	*2.6*	*Europe, zootechnie*	*taureau ; veau ; Grec (7)*
18	**14.2%**S1,S2,S3,S4,S8,S9,S10	*78.7*	*2.4*	*alimentation, emploi, transporter*	*chauffeur ; sel ; marbre ; marbrier ; pompier (5)*
19	**14.2%**S2,S4,S6,S7,S8 ; **7.1%**S3,S5,S9,S10	*35.4*	*1.9*	*art*	*aède ; poète (6)*
20	**21.4%**S4; **14.2%**S2,S5; **7.1%**S1,S3,S6,S7,S8,S9,S10	*70.2*	*2.3*	*boissons*	*Beaune, Chianti, anglais (6)*

Figure 2 : Résultats d'analyse. CCG – coefficient de cohésion globale, DM – densité moyenne

6 Conclusion

Notre étude du *sens* des *Cent mille milliards de poèmes* s'appuie sur la notion de *cohésion* comme expression de l'*unité d'un texte*. La question posée par le titre est rhétorique. L'étude complète du sens des *Cent mille milliards de poèmes* supposerait une analyse des 10^{14} sonnets potentiels, ce qui dépasse évidemment le but de notre projet. Notre démarche s'intéresserait plutôt aux enjeux d'une *analyse potentielle*, i.e. à la capacité d'un programme d'analyser un ensemble quasi-infini de textes, à partir d'un nombre fini de contraintes initiales.

Références

GENTNER D. (1981), Verb semantic structures in memory for sentences: Evidence for componential representation. *Cognitive psychology*, Vol. 13, pp. 56-83.

HALLIDAY M.A.K., HASAN R. (1976), *Cohesion in English,* London, Longman.

LAKOFF G. (1976), Pronouns and reference, *Syntax and semantics, Notes from the linguistic underground,* Vol. 7, pp. 275-335.

LE LIONNAIS F. (1986), Lipo. First Manifesto, In *Oulipo, A Primer of Potential Literature,* Translated and edited by Warren F. Motte Jr., Lincoln, University of Nebraska Press.

MINSKY M. (1975), A framework for representing knowledge. In *The Psychology of Computer Vision*, P.H. Winston Editor, New York, McGraw-Hill.

MORRIS J., HIRST G. (1991), Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of the Text, *Computational Linguistics*, Vol. 17, pp. 21-45.

MOTTE W.F. Jr. (1986), Introduction, In *A Primer of Potential Literature,* Translated and edited by Warren F. Motte Jr., Lincoln, University of Nebraska Press.

QUENEAU R. (1961), *Cent mille milliards de poèmes,* Paris, Gallimard

RASTIER F. (1987), *Sémantique interprétative,* Paris, Presses Universitaires de France.

RASTIER F. (1997), *Meaning and Textuality,* Toronto, University of Toronto Press.

STODDARD S. (1991), *Text and Texture: Patterns of Cohesion,* Norwood, Ablex Publishing.