Acta-Cognitica, pp. 227-235
ARCo'07 – Cognition, Complexité, Collectif

Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies

Mick GRZESITCHAK – Evelyne JACQUEY – Mathieu VALETTE

ATILF, UMR ATILF-CNRS Nancy Université
mickgrz@atilf.fr, evelyne.jacquey@atilf.fr, mathieu.valette@atilf.fr

Résumé – Nous abordons la question de l'interprétation semi-automatique de textes. L'approche vise à évaluer le potentiel de l'analyse sémique en matière d'identification du contenu sémantique d'un texte. Dans l'expérience relatée, nous avons annoté un corpus de textes journalistiques au moyen d'un dictionnaire de sèmes et étudié la distribution desdits sèmes dans ces textes de manière à évaluer l'apport d'un tel enrichissement. Les résultats laisse entrevoir la possibilité de qualifier en termes d'isotopies (récurrence d'un sème) les phénomènes observés. Certains aspects de notre problématique sont sensiblement proches des questions que sous-entendent les systèmes complexes dans d'autres disciplines scientifiques.

Mots-Clés – Étiquetage sémantique, Isotopie, Linguistique textuelle, Sémantique, Système complexe, Trait sémantique.

1. INTRODUCTION

En matière de structuration de la connaissance, la tradition observée actuellement, tant d'un point de vue théorique (recherches cognitives) qu'applicatif (ingénierie des connaissances) est celle, d'origine aristotélicienne, de l'indexation des mots à des référents organisés en ontologies. Cette approche considère le mot comme l'unité linguistique fondamentale et lui alloue une valeur d'étiquette. Les lexiques des langues seraient dans ces conditions des jeux d'étiquettes interchangeables, et l'activité langagière reviendrait à actualiser des concepts par le biais de ces étiquettes-mots. À rebours de cette opinion, certaines approches linguistiques, d'inspiration structuraliste, considèrent que le principal objet empirique de la linguistique n'est pas le mot, ni même la phrase, mais le texte. Il constitue alors une unité linguistique complexe, évidemment sécable et analysable en éléments discrets, mais qui obéit à une cohérence générale qu'il hérite de contraintes liées aux conditions de sa production. La constitution des corpus, en tant qu'ensemble de textes entiers réunis selon leur appartenance à une pratique linguistique et culturelle, participe de cette problématique : de même qu'il semble artificiel d'isoler un mot de son contexte pour en étudier le sens (peu y parviennent), il apparaît improbable de comprendre un texte indépendamment d'autres textes relevant d'une même pratique.

Forte de ces conditions initiales, la sémantique structurale (Greimas 1966, Pottier 1992) et surtout ses différents approfondissements théoriques récents (Rastier 1989, 2001) ont développé l'idée que le *sème*, ou *trait sémantique*, en tant qu'unité minimale de sens, participait à la cohésion du texte et, d'une certaine façon, à son *interprétabilité*. Autrement dit, il y aurait une économie du texte dépendant non pas des mots et des phrases mais d'unités plus petites, infralexicales. C'est cette économie générale, complexe et encore peu étudiée, principalement faute d'outils et de ressources *ad hoc*, que nous souhaitons questionner dans cette étude, qui s'inscrit au sein d'un projet général (DIXEM) visant à équiper la sémantique textuelle d'une instrumentation pour l'analyse et la modélisation du sens. Les perspectives applicatives en sont la recherche d'information et l'interprétation automatique de documents (Valette et al. 2006, Valette 2007).
L'unité sémantique complexe sur laquelle nous nous focaliserons ici est l'*isotopie*, c'est-à-dire
la récurrence d'un sème dans un texte, à intervalle régulier. Les isotopies, en s'organisant en faisceaux constituent le *fond sémantique* d'un texte, autrement dit les thèmes généraux qui structurent le texte (Rastier, *op.cit.*). Nous tâcherons donc de montrer que ces isotopies, traditionnellement identifiées manuellement ou par la récurrence d'unités lexicales du texte, sont observables directement en termes de sèmes.
Après une explicitation de nos arguments (2), nous présenterons brièvement nos ressources (3.1. le dictionnaire de sèmes, 3.2. le corpus de l'expérience), puis nous détaillerons notre expérience (4.1) et nos résultats (4.2).

2. PROBLÉMATIQUE

L'expérience présentée dans cet article vise à montrer que le comportement des sèmes et des textes annotés se rapproche des systèmes complexes en raison des caractéristiques suivantes.
– Ce sont des structures comportant un nombre très élevé d'unités en interaction. À titre d'exemple, pour un simple texte comme cet article d'environ 3 000 mots, on aura 36 000 sèmes (le signifié d'une unité lexicale comprend en moyenne 12 sèmes).
– Ce sens global d'un texte n'est pas la somme des sèmes le constituant, il émerge dans le cadre de *parcours interprétatif* où sont sélectionné les sèmes pertinents par rapport à une *présomption* initiale sur le contenu du texte. De même, les sèmes d'une unité lexicale ne sont pas définitifs, ils se stabilisent de façon singulière au sein d'un texte donné. Le tout est conditionné par ses parties, et les parties conditionnent le tout. Il est possible d'ajouter un niveau d'analyse supplémentaire en observant que le sens d'un texte s'établit également d'après le corpus auquel il appartient (et inversement). En somme, on observe des phénomènes d'*hétérarchie sémique*.

Nous présentons maintenant l'expérience menée afin d'approfondir notre problématique : l'observation des isotopies de quelques textes qui peuvent être considérés, du point de vue sémantique, comme des systèmes complexes.

3. RESSOURCES

3.1. Le dictionnaire de sèmes

Notre dictionnaire de traits infralexicaux (ou sèmes) a été réalisé à partir d'une extraction des mots des définition d'un dictionnaire de langue informatisé, le *Trésor de la Langue Française informatisé* (Dendien & Pierrel 2003), désormais *TLFi*. Ce dictionnaire est doté de 100 000 mots et de 270 000 définitions. Pour la constitution du dictionnaire de sèmes, nous faisons l'hypothèse : une définition est un sémème mis en texte. Ainsi, les mots pleins d'une définition (substantifs, adjectifs, verbes et certains adverbes) sont, une fois lemmatisés, considérés comme les sèmes potentiels qui constituent le signifié d'une unité lexicale en attente d'actualisation. Par exemple, pour une définition telle que : *vertex :* « point le plus élevé de la voûte crânienne », on aura les sèmes potentiels : /point/, /élevé/, /voûte/, /crânien/.

3.2. Le corpus

Nous avons constitué un corpus à partir d'articles journalistiques du *Monde Diplomatique*. Nous nous sommes attaché à constituer un corpus homogène, tant d'un point de vue discursif (discours journalistique), générique (genre de l'article) que thématique (thème de l'immigration, cf. *infra*.). Ces articles sont structurés dans un format XML, selon les recommandations TEI. Ils comprennent : une partie en-tête (header) regroupant diverses informations comme l'auteur de l'article ou sa date de parution, et le corps (body) de l'article lui-même. Le corps de l'article est subdivisé en quatre parties distinctes : le chapeau de l'article (l'introduction proche d'un résumé de l'article), l'article lui-même subdivisé par paragraphes, les notes de l'auteur qui agrémentent l'article et enfin des mots-clés qui représentent le sujet du document. Ces mots-clés nous donnent une idée de la thématique abordée dans l'article, c'est pourquoi nous nous en sommes servis pour sélectionner des articles homogènes. Nous avons choisi de regrouper les articles indexés par les unités lexicales *« France »* et *« Immigration »* ou *« Extrême-droite »*. Quinze articles ont été extraits, ce qui représente un corpus de 38 437 occurrences. Enfin, pour accroître encore l'homogénéité thématique des articles, nous les avons introduits dans le logiciel Lexico3 pour effectuer une analyse factorielle des correspondances. Nous avons sélectionné quatre articles parmi les quinze initiaux car l'analyse factorielle fournie par Lexico3 soulignait leur proximité. Nous avons donc préparé pour annotation un

corpus de quatre articles (3265 occurrences, article 96-1 « *Replis communautaires à Sarcelles* », 96-8 « *Toulon, ville amirale du front national* », 97-5 « *De l'usage des régularisations* », 97-7 « *La politique française d'immigration mise à l'épreuve* »).

4. EXPÉRIENCE & ANALYSES

4.1. Expérimentation

Nous avons étudié le corpus au travers de mesures statistiques. Afin de détecter les meilleurs pistes méthodologiques possibles pour la détection des isotopies, nous avons décidé d'extraire le maximum possible d'informations. Ainsi, nous avons étiqueté sémantiquement ces quatre textes avec notre outil, puis plusieurs mesures statistiques ont été calculées à la fois pour les textes étiquetés mais aussi pour les textes originaux (uniquement lemmatisés). En résumé, pour chaque lemme du corpus et pour chaque sème, nous avons calculé diverses mesures caractérisant ces items dans les paragraphes d'un texte, dans les textes d'un corpus et enfin dans tous les paragraphes du corpus. Nous ressortons ainsi, pour nos deux corpus, deux listes : 500 sèmes et 1 500 lemmes. Chaque liste est classée par ordre décroissant de fréquence, avec pour chaque item les valeurs des différentes mesures statistiques. Par exemple, pour le sème /société/ nous obtenons :
– ces valeurs le caractérisant dans le corpus des quatre articles ;

Tableau 1. Différentes mesures statistiques pour le trait sémantique /société/ dans le corpus d'articles du Monde Diplomatique.

classt freq.	sème	freq. Corpus	presence_ corpus	presence_ article_ moyenne	moyenne	écart-type	écart-type para
27	société	753	100 %	85.5	188	65.32	27

– et des valeurs comme celles-ci pour chaque article du corpus.

Tableau 2. Différentes mesures statistiques pour le trait sémantique /société/ dans un texte précis du corpus d'articles du Monde Diplomatique.

classt freq.	Sème	freq. texte	presence_para	moyenne	écart-type
39	Société	269	75.0 %	33.6	36.25

Toutes ces valeurs sont également générées pour les lemmes du corpus. Le tableau (tableau 1) regroupe les valeurs suivantes.

- presence_corpus : la proportion d'articles où est présent le sème/lemme.

- presence_article_moyenne : la proportion de paragraphes où est présent le sème/lemme parmi tous les articles du corpus.
- moyenne : la fréquence moyenne d'apparition du sème/lemme dans tous les articles du corpus.
- écart-type : l'écart-type dans tous les articles du corpus.
- écart-type para : La moyenne des écarts-types dans tous les paragraphes du texte.

Le tableau (tableau 2) regroupe les éléments suivants.
- presence_para : la proportion de paragraphes où est présent le sème/lemme.
- moyenne : la fréquence moyenne d'apparition du sème/lemme dans tous les paragraphes du texte.
- écart-type : l'écart-type dans tous les paragraphes du texte.

Ces tableaux de résultats présentent l'avantage de pouvoir être triés selon les colonnes qui nous intéressent. Toute la difficulté est d'attribuer une sémantique à ces tris.

4.2. Résultats et analyses

Parmi les sèmes les plus répandus, c'est-à-dire possédant de hautes fréquences, on distingue sans mal des éléments relevant du métalangage lexicographique non filtré dans le lexique utilisé pour cette expérience :
/état/, /action/, /ensemble/, /chose/, /personne/, /manière/, /lieu/, /élément/, /situation/, /exprimer/, etc.
Cette première analyse montre l'importance quantitative de ces éléments. Le bruit apparaît massif. La confusion possible entre des *pseudo-sèmes* (cf. [Valette et al. 2006]) issus du métalangage lexicographique et de véritables sèmes candidats posent de fréquents problèmes d'interprétation. Ainsi, le substantif *« pouvoir »*, souvent présent dans les hautes fréquences de nos listes, serait un trait candidat pertinent dans le corpus choisi (politique, société), mais on ne peut ignorer qu'il correspond également, en tant que verbe, au lemme de l'auxiliaire modal utilisé de façon quasi figée dans les définitions des mots suffixés en *-able*. Par exemple, la définition de l'adjectif *opposable* est « qui peut être opposé ». Dans l'état actuel de notre ressource, l'ensemble des sèmes du lexème 'opposable' est [/pouvoir/, /opposer/] alors que le sème métalinguistique correspondrait plutôt à un *potentiel*.
Les éléments relevant d'un hypothétique fond sémantique apparaissent avec une fréquence moins élevée, de façon relativement dispersée, dans un entre-deux statistique. On peut toutefois repérer dans les hautes fréquences, avec une certaine régularité dans les différents textes, des indicateurs de domaines très généraux qui pourraient compléter ceux plus précis proposés par les lexicographes. Par exemple, dans notre corpus, les sèmes /société/, /social/,

/politique/, /ville/ permettent de situer à grands traits certaines des thématiques générales du *Monde diplomatique*.
Il s'agit maintenant d'organiser les sèmes candidats à la formation d'une isotopie. En classant les traits du texte 97-7 par ordre croissant selon l'écart-type calculé à partir leurs fréquences dans la totalité des paragraphes (ecart_type 2) puis selon le taux de présence dans les paragraphes (présence_para), nous constatons que parmi les 16 premiers sèmes candidats, les items /travailler/ et /étranger/ relèvent directement de la thématique du texte (le travail clandestin).

Tableau 3. Classement des traits de l'article 97-7.

	classt freq,	trait	freq, texte	presenc e_para	moy 1	moy 2	écart-type 1	écart-type 2
2	408	essentiellement	22	87,5%	3,14	2,75	0,35	0,49
3	494	variation	18	87,5%	2,57	2,25	0,73	0,74
4	438	en particulier	21	100%	2,62	2,62	0,86	0,86
5	486	habileté	19	87,5%	2,71	2,38	0,88	0,88
6	401	prisonnier	22	100%	2,75	2,75	0,97	0,97
7	495	travailler	18	100%	2,25	2,25	0,97	0,97
8	366	prétendre	24	100%	3	3	1	1
9	422	provoquer	21	87,5%	3	2,62	1,07	1,06
10	479	ouvrage	19	87,5%	2,71	2,38	1,16	1,13
11	491	étendre	18	87,5%	2,57	2,25	1,18	1,14
12	415	éprouver	21	87,5%	3	2,62	1,2	1,17
13	463	jeu	20	87,5%	2,86	2,5	1,25	1,21
14	471	commencer	20	87,5%	2,86	2,5	1,25	1,21
15	361	étranger	24	100%	3	3	1,22	1,22

En revanche, un sème classé en 5ème position, avec une régularité de 100% pour 22 occurrences (chaque paragraphe en comprend au moins une) mérite toute notre attention : il s'agit de /prisonnier/. Il se trouve que sa racine 'prison-' est complètement absente du texte. C'est donc bel et bien d'une isotopie dont il s'agit là. Si les formes de la famille morphologique sont absentes du texte (*« prison »*, *« prisonnier »*, *« emprisonner »*, etc.) et même les synonymes possibles (comme *« détention »*), le sème, lui, participe au fond sémantique du texte.

Le même classement, sur le texte 96-8, qui traite de l'administration de la ville de Toulon par le Front National présente une situation similaire. En 10ème position, le sème candidat /harceler/ apparaît avec un taux de couverture de 90% (pour 19 paragraphes). Là encore, le mot *« harceler »* et ses dérivés morphologiques sont complètement absents du texte proprement dit.

Tableau 4. Classement des traits de l'article 96-8.

	classt freq,	trait	freq, texte	presence_para	moy 1	moy 2	écart-type 1	écart-type 2
2	482	bois	70	80%	8,75	7	2,22	2,53
3	432	nécessaire	75	70%	10,7	7,5	2,76	3,55
4	481	paraître	70	80%	8,75	7	3,73	3,69
5	472	changer	71	80%	8,88	7,1	3,76	3,72
6	402	cesse	78	90%	8,67	7,8	3,83	3,73
7	408	pop	77	90%	8,56	7,7	3,83	3,73
8	419	purement	76	90%	8,44	7,6	3,86	3,76
9	420	Négation	76	90%	8,44	7,6	3,86	3,76
10	424	inverser	76	90%	8,44	7,6	3,86	3,76
11	425	harceler	76	90%	8,44	7,6	3,86	3,76
12	426	fondamentalement	76	90%	8,44	7,6	3,86	3,76
13	443	gouvernement	74	80%	9,25	7,4	3,86	3,83
14	411	jouer	77	80%	9,62	7,7	3,84	3,84
15	470	commencer	71	80%	8,88	7,1	3,95	3,87
16	374	réellement	81	90%	9	8,1	4,08	3,97

Le texte 96-1 présente un intérêt particulier. Dans ce texte où il est essentiellement question des jeunes de la ville de Sarcelles, on a en 9ème position, le trait candidat /enfant/. Le mot « enfant » est toujours absent du texte. Mais si jusqu'à maintenant, on pouvait interpréter les traits candidats à l'isotopie comme des « idées de » (idée de harcèlement, idée de prison), il ne peut être question ici d'une « idée d'enfant ». Car c'est bien d'enfants dont il s'agit, l'auteur pratique l'euphémisme en employant des mots tels que « jeune » ou « adolescent » pour qualifier les protagonistes. Le sème isotopique isolé donne accès plus crûment à la réalité des faits : le fond sémantique nous parle d'enfants.

Tableau 5. Classement des traits de l'article 96-1.

classt freq.	trait	freq. texte	presence_para	moy 1	moy 2	écart-type 1	écart-type 2
482	changer	24	66.7 %	6	4	1	1,82
481	complet	24	66.7 %	6	4	1,22	1,91
476	pose	24	66.7 %	6	4	1,58	2,08
478	jouer	24	66.7 %	6	4	1,58	2,08
486	aboutir	24	66.7 %	6	4	1,58	2,08
392	titre	29	66.7 %	7,25	4,83	0,83	2,09
451	convenable	26	66.7 %	6,5	4,33	1,5	2,15
466	déplacement	25	66.7 %	6,25	4,17	1,64	2,16
418	enfant	28	66.7 %	7	4,67	1,41	2,23
483	artiste	24	66.7 %	6	4	1,87	2,24
461	fondamental	25	66.7 %	6,25	4,17	1,79	2,24
465	déplacer	25	66.7 %	6,25	4,17	1,79	2,24
429	surtout	27	66.7 %	6,75	4,5	1,64	2,27
438	indéfini	27	66.7 %	6,75	4,5	1,64	2,27
421	confiance	28	66.7 %	7	4,67	1,58	2,3

Plus étonnant est le classement du sème candidat /accident/, au rang 2 dans le texte 97-5, suivant les mêmes critères de classement. Le texte 97_5 aborde sensiblement le même sujet que le texte 97-7 vu ci-dessus, et là encore, il n'est nullement question d'accident dans tout le texte. Pourtant, le sème candidat est présent dans 80% des paragraphes avec régularité.

En conclusion, nous remarquons un bruit considérable dans les données analysables, mais ce bruit, pourvu que nous puissions y distinguer une structure, est peut-être créateur d'ordre. Une typologie des sèmes (actuellement en cours de réalisation, *cf.* Valette, Estacio *et al.* 2006, Ramdani 2007) croisée à une typologie des isotopies permettra à termes d'améliorer le rappel et la précision de notre méthode.

Il est intéressant de voir que les sèmes qui apparaissent saillants sont absents des textes eux-mêmes. D'une manière générale, il semble y avoir peu de recoupements entre les lemmes d'un texte et ses sèmes, ce qui confirme nos hypothèses quant à l'existence de deux niveaux d'interprétation (ou d'analyse) distincts. Outre une typologie et une hiérarchisation des sèmes, une typologie des isotopies devrait améliorer le rendement et permettre notamment davantage de prédictibilité des sèmes pertinents. En somme, le bruit, pourvu que nous puissions y distinguer une structure, est peut-être créateur d'ordre.

5. PERSPECTIVES

Dans ce papier, nous avons cherché à montrer qu'à partir d'un objet complexe (texte annoté sémiquement), on pouvait entrevoir une organisation sémantique inédite. Il reste évidemment à affiner nos outils d'observation avec les modèles adéquats.

Étant donné que les phénomènes sémantiques sur lesquels nous travaillons ont des propriétés similaires à celles qui caractérisent les systèmes complexes, il serait souhaitable que les traitements statistiques que nous avons mis en place, et que nous envisageons de mettre en place, soient réalisés à l'aide d'outils et de méthodologies éprouvés pour les systèmes complexes, comme par exemple les réseaux de neurones (Victorri, & Fuchs 1996). D'autre part, nous avons la volonté de constituer des classes sémantiques d'items (de sèmes, lemmes, portions textuelles, textes, sous-corpus, etc), ce qui est l'un des domaines d'application reconnu pour les réseaux de neurones. Enfin, de même que les sèmes permettent d'unifier l'observation du comportement sémantique des unités linguistiques quel que soit leur niveau (mot, syntagme, phrase, texte, etc), les réseaux de neurones pourraient jouer un rôle unificateur similaire au niveau des traitements statistiques que nous envisageons.

Enfin, il est admis que le lexique d'une langue s'acquiert en interaction avec d'autres locuteurs, de manière collective. Comme le montrent (Aknine *et al.* 2004), le rapprochement avec les systèmes multi-agent apparaît pertinent.

Dans (Aknine *et al.* 2004), les auteurs présentent une méthode pour analyser le contenu de documents. Le coeur de leur méthode est de former des coalitions d'agents permettant une analyse multidimensionnelle du contenu des documents. En d'autres termes, cela permet de combiner de nombreuses règles d'analyse et par consensus les agents élaborent une solution globale pertinente. Pour notre problématique, nous pouvons imaginer transplanter cette méthode pour détecter les isotopies d'un texte ou même sa thématique. Encore mieux, on pourrait imaginer que les agents mettent eux-mêmes le lexique sémantique à jour sous réserve qu'ils aient trouvé un consensus entre eux.

6. BIBLIOGRAPHIE

Aknine, S. Caillou, P., & Slodzian, A. (2004). Méthode Consensuelle de Formation de Coalitions Multi-Agents. In *congrès francophone de reconnaissance des formes et intelligence artificielle*, pp. 979-988.

Dendien, J. & Pierrel, J.-M. (2003). Le Trésor de la Langue Française informatisé : un exemple d'informatisation d'un dictionnaire de langue de référence. *TAL Vol 44/2*, 11-37.

Greimas, A. J. (1966). *Sémantique structurale*, PUF, Paris, 1966.

Pottier, B. (1992). *Théorie et Analyse Linguistique*, Hachette, Paris, 1992.

Ramdani, E. (2007) *Du dictionnaire de langue au lexique TAL - la construction d'une ressource pour l'annotation sémantique des textes*. Mémoire de Master, Paris, InaLCO.

Rastier, F. (2001). *Arts et sciences du texte*, Paris, PUF.

Rastier. F. (2003). Formes sémantiques et Textualité. *Cahiers du CRISCO*, *12*, 99-114.

Valette, M. Estacio-Moreno, A. Petitjean, E., & Jacquey, E. (2006). Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens. *Verbum ex machina*, *Actes de la 13ème conférence sur le traitement automatique des langues naturelles (TALN 06)*. Piet Mertens, Cédrick Fairon, Anne Dister, Patrick Watrin (éds). *Cahiers du CENTAL*, 2.1, UCL Presses Universitaires de Louvain. Volume 1. Pages 357-366).

Valette, M. (2007). A quoi servent les lexiques sémantiques ? Discussion et proposition, *Actes du colloque Description Linguistique pour le Traitement Automatique du Français, Congrès de l'ACFAS.* Montréal, Canada, 15-19 mai 2006,.

Victorri, B. & Fuchs, C. (1996). *La polysémie. Construction dynamique du sens*. Paris, Hermès.