Université Montpellier 2

Mémoire d'Habilitation à Diriger les Recherches

Spécialité : informatique

Lexique et analyse sémantique de textes
-
structures, acquisitions, calculs, et jeux de mots

par

Mathieu Lafourcade

PRÉSENTATION LE 7 DÉCEMBRE 2011 DEVANT LE JURY COMPOSÉ DE :

Christian Boitet	Université Joseph Fourier - Grenoble 1 - LIG	Examinateur
Vladimir Fomichov	Higher School of Economics National Research University, Moscou	Rapporteur
Violaine Prince	Université Montpellier 2 - LIRMM	Examinatrice
Christian Retoré	Université Bordeaux 1 - LaBRI	Rapporteur
Eric Wehrli	Université de Genève - LATL	Examinateur
Michael Zock	Université de Marseille - LIF	Examinateur
Pierre Zweigenbaum	LIMSI-CNRS	Rapporteur

Version du 22 novembre 2011 à 17:50

Résumé

L'analyse sémantique de textes nécessite en préalable la construction d'objets relevant de la sémantique lexicale. Les vecteurs d'idées et les réseaux lexicaux semblent de bons candidats et constituent ensemble des structures complémentaires. Toutefois, faut-il encore être capable dans la pratique de les construire. Les vecteurs d'idées peuvent être calculés à partir de corpus de définitions de dictionnaires, de thésaurus ou encore de textes. Ils peuvent se décliner en des vecteurs conceptuels, des vecteurs anonymes ou des vecteurs lexicaux - chaque type présentant un équilibre différent entre précision, couverture et praticité. Quant aux réseaux lexicaux, ils peuvent être acquis efficacement via des jeux, et c'est précisément l'objet du projet JeuxDeMots. L'analyse sémantique peut être abordée par l'analyse thématique, et ainsi servir de moyen de calcul à des vecteurs d'idées (bouclage). Nous pouvons modéliser l'analyse comme un problème d'activation et de propagation. La multiplicité des critères pouvant intervenir dans une analyse sémantique, et la difficulté inhérente à définir une fonction de contrôle satisfaisante, nous amène à explorer l'usage de métaheuristiques bio-inspirées. Plus précisément, nous introduisons un modèle d'analyse par colonies de fourmis artificielles. À partir d'un texte, l'analyse vise a construire un graphe contenant les objets du texte (les mots), des objets identifiés comme pertinents (des syntagmes, des concepts) ainsi que des relations pondérées et typées entre ces objets.

Mots-clés

Traitement Automatique des Langues, analyse sémantique de textes, sémantique lexicale, vecteurs d'idées, réseaux lexico-sémantiques, acquisition lexicale, jeux sérieux.

Abstract

The semantic analysis of texts requires beforehand the building of objects related to lexical semantics. Idea vectors and lexical networks seems to be adequate for such a purpose and are complementary. However, one should still be able to construct them in practice. Vectors can be computed with definition corpora extracted from dictionaries, with thesaurii or with plain texts. They can be derived as conceptual vectors, anonymous vectors or lexical vectors - each of those being a particular balance between precision, coverage and practicality. Concerning lexical networks, they can be efficiently constructed through serious games, which is precisely the goal of the JeuxDeMots project. The semantic analysis can be tackled from the thematic analysis, and can serve as computing means for idea vectors. We can modelise the analysis problem as actviations and propagations. The numerous criteria occuring in the semantic analysis and the difficulties related to the proper definition of a control function, lead us to explore metaheuristics inspired from nature. More precisely, we introduce an analysis moodel based on artificial ant colonies. From a given text, the analysis aims at building a graph holding objects of the text (words, phrases, sentences, etc.), highlighting objects considered as relevant (phrases and concepts) as well as typed and weighted relations between those objects.

Keywords

Natural Language Processing, text semantic analysis, lexical semantics, idea vectors, lexical network, lexical acquisition, serious games.

Avant-propos

Dans le cadre du Traitement Automatique du Langage Naturel (TALN), mes thèmes de recherches concernent l'analyse sémantique et l'acquisition de ressources lexicales comme supports à cette analyse. L'exposé de mes travaux porte sur les problèmes liés à la représentation en sémantique vectorielle lexicale, l'acquisition de ces données, ainsi que leur validation et exploitation. Plusieurs définitions et mises en œuvre de l'analyse sémantique sont proposées à l'aide d'algorithmes de propagation. L'acquisition des données en sémantique lexicale est un problème difficile qui peut trouver une solution opératoire via des jeux en ligne proposés à des internautes non spécialistes en linguistique. Ce document se veut être une synthèse des travaux que j'ai menés sur ces questions depuis 1995. Chaque partie aborde une thématique particulière en essayant à la fois d'en présenter les grande lignes, d'offrir souvent une reformulation synthétique avec des résultats non publiés par ailleurs, et enfin d'inclure une ou plusieurs publications représentatives.

Remerciements

Je tiens à exprimer ma reconnaissance à tous ceux qui de près ou de loin ont été impliqués dans les travaux que je présente ici. Je ne pourrais les nommer tous, mais je pense en particulier :

à Ch. Boitet et à V. Prince, *mes mentors*, qui durant toutes ces années ont su non seulement m'encourager, mais également impulser, raturer, biffer, triturer, malaxer, digérer, reformuler ma production scientifique ;

à J. Chauché pour avoir été l'initiateur des vecteurs sémantiques, précurseurs des vecteurs d'idées, mais également pour avoir réalisé Sygmart et SygFran sans lesquels bien peu aurait été fait ;

à F. Guinand avec lequel nous avons joué aux crypto-entomologues dont les approches bioinspirées ont elles-mêmes été inspiratrices d'un modèle d'analyse sémantique de textes à colonies de fourmis artificielles ;

à D. Schwab avec qui de nombreuses idées présentes dans ce mémoire ont mûri et qui a su leur faire prendre vie et surtout les diffuser dans d'autres lieux ;

à V. Zampa sans qui PtiClic n'aurait pu voir le jour ;

à l'ensemble de mes amis et collègues de Malaisie, Thaïlande et Vietnam pour les différents projets que nous avons menés ensembles ;

à l'ensemble de la communauté de JeuxDeMots, non seulement pour avoir joué encore et encore, mais pour un (pas si) petit groupe parmi eux, d'avoir eu envie de faire évoluer l'idée et le logiciel pour devenir ce qu'il est maintenant. Ma reconnaissance en particulier à Caillouteux, k@tsof, Lyn, Mym, N@t, niniefaitlesmots, ..Syl.. et tout ceux que j'oublie.

aux membres du LIRMM et d'ailleurs pour avoir su se prêter amicalement à de nombreuses expériences autour de l'acquisition lexicale ; aux étudiants de master informatique 1 et 2, qui ont bien voulu jouer les cobayes avec enthousiasme ;

aux relecteurs de ce mémoire, en particulier Agnès, Alain, Cédric, Christian, Didier, ma mère Pierrette, Violaine et Virginie qui ont été d'une efficacité redoutable et d'une patience insoupçonnée ;

à Zélie, à David et à Brigitte qui ont supporté tout cela.

Table des matières

Table des figures

Introduction

De nombreuses applications du TALN (Traitement Automatique des Langues Naturelles), comme l'indexation de textes, la traduction automatique, ou encore le résumé automatique, sont potentiellement demandeuses d'une analyse sémantique aussi fine que possible des textes. Il peut s'agir, par exemple, d'extraire la thématique générale d'un document (ou à une granularité plus fine, des paragraphes ou des phrases) sous la forme de sélection de concepts prédéterminés. L'extraction de termes-clés présents dans le texte est également potentiellement utile en indexation de documents pour des moteurs de recherche, avec possiblement le *calcul* de termes absents du texte mais thématiquement pertinents. L'identification des syntagmes, de leurs fonctions syntaxiques et des relations qu'ils peuvent entretenir entre eux est nécessaire en traduction automatique. En résumé automatique, les demandes peuvent être différentes selon qu'il s'agit de contraction (où dans ce cas, des groupes prépositionnels supprimables seront recherchés), d'abstraction (où les relations saillantes entre syntagmes devront être identifiées en vue d'un paraphrasage), ou d'extraction (où l'identification des phrases-clés — celles à la fois porteuses du sens du texte et saillantes — est nécessaire). Quoi qu'il en soit, au moins trois questions délicates se posent si nous nous plaçons dans le cadre d'une analyse sémantique généraliste : 1) quels résultats — structures — souhaitons-nous obtenir ? 2) quels algorithmes seraient susceptibles de calculer ces résultats, et 3) quelles connaissances peuvent servir de support aux algorithmes ?

L'*analyse sémantique de texte* peut être définie comme une tâche visant à effectuer un certain nombre de traitements relatifs à la *sémantique lexicale* (de façon restrictive) ou à la *compréhension du sens* (selon une perspective globale). Nous adoptons volontairement une approche généraliste ne visant aucune application en particulier. Nous pouvons citer parmi les tâches possibles : la levée des ambiguïtés lexicales, le rattachement correct des groupes prépositionnels (parmi ceux autorisés par la syntaxe), la résolution des références (pronoms, adjectifs possessifs, identités, etc.), l'identification des rôles prédicatifs (agent, patient, instrument, etc.), de l'explicitation d'idées ou de concepts implicites. Bien entendu, cette liste est loin d'être exhaustive, et savoir précisément quels sont les traitements utiles n'est pas une question facile.

Des lexiques, des vecteurs et des réseaux

L'acquisition et la structuration des ressources lexicales sont des problèmes en eux-mêmes. Si nous y adjoignons la problématique de la représentation du sens, nous nous trouvons alors à l'intersection entre les bases de données lexicales et la sémantique lexicale. Les ressources s'organisent traditionnellement en lexiques qui constituent des listes d'éléments plus ou moins structurés. Le point d'entrée est usuellement qualifié de *vedette* et sera la forme lemmatique dans la plupart des dictionnaires d'usage, monolingues ou multilingues. Dans le contexte que nous considérons, les lexiques sont d'abord à usage calculatoire (ou machinal, à opposer à un usage humain). Toutefois, les possibilités d'exploration, de lecture et d'exploitation par des humains sont souhaitées, ne serait-ce que pour

vérifier la qualité des données ou les confronter à l'usage. La multiplication sur Internet des dictionnaires à vocation contributive s'affranchissant avec plus ou moins de bonheur des approches lexicographiques traditionnelles en est une illustration. La présentation de ces informations à des humains nécessite en général un traitement informatique, qui est une forme soit de (pré)compilation, soit de mise en cache en fonction des usages. La compilation pourra produire des structures permettant de comparer entre eux des objets lexicaux. Deux types de modèles pour représenter de l'information lexicale (et ontologique) sont envisagés ici : les *vecteurs d'idées* et les *réseaux lexicaux*.

Une première expérience de constitution de lexique via les projets de dictionnaires Français-Anglais-Malais (avec l'Université des Sciences de Malaisie à Penang, FeM [Gut *et al.*, 1996]), puis Français-Anglais-Thai (Université Chulalongkorm à Bangkok en Thaïlande, FeT), et Français-Anglais-Vietnamien (Université de Danang au Vietnam, FeV) a clairement mis en évidence certaines difficultés non seulement dans un cadre multilingue, mais également selon une optique visant à produire des données destinées aux individus mais aussi à un usage machinal [Lafourcade, 1998]. Ces premières expériences ont toutefois permis non seulement d'obtenir des lexiques exploitables ultérieurement, mais également de dégager les problèmes émaillant leur constitution (problèmes de l'inadéquation des outils, de la parcellisation non redondante des données à produire, et difficultés liées à la mise à disposition efficace sous forme électronique). Ces travaux ont évolué vers un cadre plus général par la constitution de bases de données lexicales multilingues avec le projet Papillon (avec G. Sérasset et M. Mangeot, [Mangeot-Lerebours *et al.*, 2003]). L'approche préconise la constitution d'une base pivot d'acceptions interlingues, qui correspondent à des unités de sens, à la fois (potentiellement) reliées entre elles et reliées aux formes monolingues. Cependant, d'un point de vue pratique, ces acceptions restent des symboles, dont l'interprétation n'est pas encodée.

L'encodage au niveau lexical, soit du champ thématique, soit d'une projection du sens, peut être réalisé à l'aide de vecteurs. Les vecteurs conceptuels permettent de représenter efficacement les idées associées à un segment textuel (sens, mot, groupe, texte). Ils forment une structure d'espace vectoriel. Selon leur mode de calcul, ils peuvent représenter un champ thématique, un champ ontologique (relation *est-un*), ou un champ antonymique (travaux avec D. Schwab, [Schwab *et al.*, 2002]). Une fonction de comparaison entre deux vecteurs est la distance angulaire (l'angle que forment deux vecteurs). Celle-ci permet ainsi de retrouver une notion de voisinage et de définir un préordre ou un ordre. Les techniques vectorielles permettent d'obtenir un fort rappel, en particulier pour l'indexation de documents, mais peuvent manquer de précision (par exemple, deux quasi-synonymes peuvent avoir des vecteurs très proches, ce qui n'est pas toujours souhaitable).

Les réseaux lexicaux constituent une approche orthogonale aux approches vectorielles. Ils forment une structure de graphe dont les arêtes sont dotées de valeurs numériques (en l'occurrence et en ce qui nous concerne, un type ou une étiquette, et un poids). Dans ce qui suit, nous considérerons les deux termes de *graphe* et *réseau* comme faisant référence au même type d'objet. L'étiquetage des relations entre les termes permet de représenter autant d'aspects différents que souhaité, qu'ils soient paradigmatiques, syntagmatiques ou ontologiques. Toutefois, s'il existe bien une relation de voisinage entre termes du réseau via la distance entre deux points, celle-ci reste en toute généralité beaucoup plus complexe à calculer qu'entre deux vecteurs.

Pour modéliser des fonctions lexicales [Mel'čuk, 1988, Mel'čuk *et al.*, 1995, Mel'čuk, 1996], le fait de combiner des relations et des vecteurs produit des résultats intéressants (travaux avec V. Prince sur la synonymie relative [Lafourcade & Prince, 2001a] et avec D. Schwab sur l'antonymie relative [Schwab *et al.*, 2002]). La notion d'horizon conceptuel a été introduite afin de rendre compte d'une barrière au-delà de laquelle les vecteurs conceptuels de termes très généraux se ne distinguent plus de termes spécifiques. C'est une des limites des vecteurs conceptuels et de leur application aux fonctions lexicales.

Calculer des vecteurs d'idées

Calculer des vecteurs d'idées peut se faire de multiples façons, qu'il est possible de catégoriser en fonction des données de départ (des textes, des définitions, un réseau lexical, des listes de termes,

etc.) et du type d'algorithme utilisé.

Le corpus d'apprentissage peut être un ensemble de définitions de dictionnaires (à usage humain) qui vient en général en complément à un thésaurus. La méthode de calcul employée peut être, par ordre de complexité croissante : la somme vectorielle des termes saillants (après un filtrage de type TF*IDF), la propagation en remontée sur l'arbre d'analyse, et la propagation en remontée-descente itérée sur l'arbre d'analyse.

Si le corpus d'apprentissage est un réseau lexical, on s'affranchit d'une bonne partie des ambiguïtés présentes dans les définitions. Par contre, selon la ressource utilisée, il n'est pas certain que les termes présents soient systématiquement identifiés comme ambigus. De plus, à partir du réseau lexical, il est possible d'avoir des vecteurs conceptuels étiquetés — c'est-à-dire, qui couvrent une facette particulière (plus seulement les idées associées ou la thématique, mais également les agents ou patients typiques, etc.). Sur les documents, le calcul peut se faire avec un couple de vecteurs en récurrence croisée (travaux avec M. Bouklit, [Bouklit & Lafourcade, 2006]).

La question de la réduction de dimension des espaces vectoriels produits se pose également. Le modèle d'Analyse Sémantique Latente (LSA, [Deerwester *et al.*, 1990b]) la met en œuvre à la fois pour diminuer les structures à manipuler, et pour réduire le bruit. Il semblerait toutefois que l'efficacité voire la pertinence de la décomposition en valeurs singulières et de la réduction de dimension soit de plus en plus remise en cause [Gamallo & Bordag, 2011].

Nous avons entrepris une évaluation sur une combinaison partielle des approches possibles, où il apparaît (sans surprise) que plus la source d'apprentissage est explicite, meilleurs sont les résultats. Pour cela, les réseaux lexicaux sont plus efficaces que les définitions de dictionnaires (nous pourrions argumenter que cela n'est pas étonnant, car le travail d'extraction a déjà été fait). Enfin, la qualité des résultats semble covariante avec la taille des vecteurs, la *quantité relative* de bruit étant globalement constante. Par contre, il est vraisemblable que la *quantité absolue* de bruit générée selon la méthode ou la source d'information puisse aussi être covariante à la taille des vecteurs.

Capturer des relations lexicales et identifier des usages

La possibilité effective d'une acquisition d'informations lexicales via une activité ludique a été démontrée à travers le projet JeuxDeMots ([Lafourcade, 2007], [Joubert & Lafourcade, 2008b], [Joubert & Lafourcade, 2008a]). Cette acquisition prend la forme de la construction incrémentale d'un réseau lexical où les relations sont orientées, typées et pondérées. L'activité ludique est ici la motivation qui pousse les utilisateurs à aboutir à une construction par consensus populaire, sans négociation. Les joueurs n'ont pas besoin d'avoir conscience qu'ils participent à la construction d'une ressource lexicale pour jouer. En effet, lors d'une partie, les joueurs ne sont pas en contact et ne peuvent donc pas *négocier* leurs réponses. PtiCLic (travaux avec V. Zampa, [Lafourcade & Zampa, 2009b], [Lafourcade & Zampa, 2009a]) est une variante de JeuxDeMots mettant l'accent sur la consolidation du réseau via une activité de réattribution de relations pour des couples de termes.

Il est possible de calculer des vecteurs d'idées par émergence à partir du réseau construit dans JeuxDeMots, de façon incrémentale, au fur et à mesure de la construction du réseau. Cette approche est à opposer à celle imposant un recalcul global de l'ensemble des vecteurs (comme dans le cas de LSA).

Contrairement à une approche à partir de définitions, le réseau lexical de JeuxDeMots ne fournit pas directement de sens pour les termes. Toutefois, par identification des sous-cliques maximales ancrées sur une entrée, il est possible de déduire au moins partiellement des usages pour chaque terme. Un usage est la projection d'une acception (au sens classique de la lexicographie) sur un contexte particulier (souvent implicite dans les dictionnaires). L'ensemble des acceptions

est contenu dans l'ensemble des usages (il suffit que le contexte soit général pour qu'un usage corresponde exactement à une acception). Par exemple, le terme *sapin* a comme usage général, entre autres, *sapin>arbre* et comme usage particulier *sapin>Noël*. Disposer des usages de sens semble plus intéressant pour la désambiguïsation lexicale, car ils paraissent souvent plus fidèles aux représentations mentales des locuteurs que les découpages dictionnairiques classiques (travaux avec A. Joubert, [Joubert & Lafourcade, 2008a], [Lafourcade & Joubert, 2010]).

Les usages identifiés pour un terme sont donc inclus dans le réseau lexical, et ce faisant sont indirectement réinjectés dans le jeu. Lors d'une partie, les joueurs peuvent être confrontés à un usage (par exemple *que vous évoque le terme* sapin>arbre *?*) et le renseigner. De plus, ils peuvent dorénavant sélectionner l'usage approprié pour les termes qu'ils proposent durant une partie. Ce bouclage est à l'origine d'un raffinement progressif des termes et des relations dans le réseau. Les champs thématiques associés aux termes deviennent de plus en plus précis au fur à mesure que le réseau se construit, et que ses termes se désambiguïsent.

En environ trois ans de jeu, plus d'un million de relations, entre environ 100 000 termes, ont été capturées. Il semblerait que la distribution des forces d'activation des relations entre termes se conforme à une loi de puissance (pour être précis, sans doute plutôt une loi de Mandelbrot de la forme $f(n) = K/(a+bn)^c$). Il semblerait aussi que la distribution des termes du réseau en fonction du nombre de relations entrantes suive cette même loi. Le réseau construit via JeuxDeMots couvre une partie conséquente de la *longue traîne* de la distribution.

Analyses thématiques et sémantiques

L'*analyse thématique* sera ici vue comme le calcul d'une structure (en général un vecteur d'idées) permettant de représenter le ou les champs lexicaux d'une texte. Pour ce faire, il est possible de procéder de façon statistique (avec des résultats souvent médiocres, d'autant plus que les textes ou segments textuels sont courts) ou bien faire appel à une analyse sémantique, particulièrement utile pour la sélection des acceptions des termes du texte.

L'*analyse sémantique* sera considérée ici comme le calcul qui, à partir d'un texte, produit une structure 1) offrant un support pour traiter un certain nombre de phénomènes linguistiques, et/ou 2) fournissant une ou plusieurs solutions aux problèmes dus à ces mêmes phénomènes. Nous distinguons les deux cas, qui peuvent se manifester simultanément, et souvent se soutenir mutuellement. Par exemple, dans le cadre de la désambiguïsation lexicale, une analyse peut pondérer par ordre de préférence les sens des termes en contexte (premier cas) ou ne retenir que ceux qui sont possibles. La désambiguïsation lexicale peut soutenir le rattachement des groupes, soit thématiquement (*L'avocat a plaidé pour son client à la cour.*), soit sémantiquement (*L'avocat a mangé une pomme dans la cour.*). Dans le premier cas, la connaissance des champs lexicaux majoritaires suffit à faire émerger *avocat>justice* plutôt que *avocat>fruit*. Par contre, dans le second exemple, les thèmes majoritaires sont liés à la nourriture et induisent une interprétation erronée avec la sélection de *avocat>justice*. Des relations prédicatives (concernant le verbe *manger*) ainsi que des opérations minimales d'induction (un *avocat>justice* est un homme ; un homme peut manger ; une *pomme>fruit* peut typiquement être mangée) doivent prendre la suite de l'approche thématique.

Un algorithme, dit de *propagation*, fait circuler des vecteurs d'idées dans une structure de graphe. Il peut s'agir d'un arbre d'analyse morpho-syntaxique : plusieurs variantes sont possibles non seulement selon la nature des ressources lexicales disponibles, mais aussi selon l'application visée (indexation ou traduction). La structure de graphe peut aussi être plus proche de celle des graphes conceptuels, comme c'est le cas pour le projet UNL [1], mais dans ce cas la propagation peut être délicate à mettre en œuvre. En effet, il faut à la fois tenir compte des cycles présents, et aussi de la

1. http://www.vai.dia.fi.upm.es/ing/projects/unl/index.htm

nature sémantique des relations du graphe. Cependant, ce type d'algorithme est 1) trop localiste et produit difficilement des relations à longue portée (entre phrases ou paragraphes) et 2) ne permet pas d'effectuer des modifications structurelles de l'environnement pouvant être lues comme une partie de la solution.

Les algorithmes à colonies de fourmis (travaux avec F. Guinand [Gui2010], et D. Schwab [Schwab, 2005]) permettent de résoudre ce type de problème tout en offrant un certain nombre d'avantages : simplification du contrôle, parallélisation possible, modification possible de l'environnement durant l'exécution, etc. Ici, seuls les vecteurs d'idées sont utilisés comme marqueurs d'information, les fourmis étant des agents transporteurs. La *stigmergie* (communication indirecte par modification de l'environnement) est réalisée via des phéromones artificielles qui sont des marquages se dissipant légèrement à chaque cycle. De plus, sous certaines conditions, les agents peuvent construire des passerelles entre les objets de l'environnement, permettant de proche en proche de sortir d'une démarche strictement localiste. Ainsi, les agents construisent la solution à la fois par renforcement/rejet d'éléments ou par création/destruction de liens entre les objets.

L'adjonction d'un réseau lexical (travaux avec D. Schwab [Schwab, 2005]) permet de dépasser les limites d'un processus fondé uniquement sur des informations thématiques (en plus de celle de l'arbre morphosyntaxique - travaux de J. Chauché avec Sygmart et Sygfran). Il est ainsi possible d'exploiter plus finement des informations prédicatives et valencielles (agent, patient, instrument), des informations de typicalité (lieux typiques : *cheval* → *pré*, moments typiques : *cadeau* → *anniversaire*, etc.) et des fonctions lexicales et ontologiques (synonymie, antonymie, hyponymie, hyperonymie, méronymie, holonymie, etc). Des travaux sur le rattachement de groupes prépositionnels (avec N. Gala, [Gala & Lafourcade, 2007]) ont montré que, dans environ 80% des cas, une information thématique est suffisante pour obtenir un rattachement correct. Dans les autres cas, des relations de typicalité (lieux typiques, instruments typiques, moments typiques) et la prise en compte des restrictions sémantiques sur l'élément régi par la préposition sont nécessaires pour obtenir un rattachement qui correspond à l'intuition.

Applications et ouvertures

Des travaux précédents découlent un certain nombre d'applications qui sortent du cadre strict du TALN, ainsi qu'une ouverture vers quelques pistes de recherche (correspondant à des travaux déjà amorcés).

Le modèle de JeuxDeMots peut trouver des variantes intéressantes permettant à la fois d'acquérir des informations lexicales dont nous ne disposons pas encore (polarité, relation d'inhibition), et de raffiner une relation d'association libre vers une fonction lexicale plus précise. Il s'agit essentiellement de fonctions lexicales ou ontologiques importantes, souvent peu lexicalisées (comme *produit*, *producteur*, par exemple). Le traitement de ces fonctions lexicales est difficile à mettre en œuvre dans le modèle d'origine de JeuxDeMots, car elles présentent un aspect ludique limité : trop peu de réponses, trop immédiates, relativement peu ambiguës. De plus, la sélection automatique de termes intéressants, ou tout simplement valides, pour ces fonctions lexicales peut être difficile et relativement bruitée. Par contre, la reconnaissance d'intrus en contexte peut fournir des informations susceptibles d'aider à l'identification d'usage de termes (cf. partie sur les cliques d'usage avec A. Joubert, [Lafourcade & Joubert, 2010]), ainsi qu'à l'établissement de relations à valeur négative (correspondant à une impossibilité pertinente, par exemple : *autruche agent* voler*).

L'évaluation qualitative du réseau doit également être considérée. Elle peut se faire de façon classique par une approche manuelle via un échantillonnage. Toutefois, nous avons opté dans un premier temps pour une évaluation indirecte via un *jeu de devinette* du type *trouver le mot sur le bout de la langue* [Joubert *et al.*, 2011]. À partir d'un nombre réduit d'indices, est-il possible de faire retrouver

un mot donné ? Des algorithmes extrêmement simples fondés sur l'intersection de vecteurs d'idées calculés à partir du réseau lexical de JeuxDeMots permettent d'obtenir au tout venant un taux de réussite de 70-75%.

Tous les algorithmes d'analyse présentés postulent la pré-existence d'une structure morpho-syntaxique (arbre de constituants ou de dépendance) ou encore une forme approchante de graphes conceptuels (travaux en rapport avec le projet UNL). Il est possible de fournir des pistes de recherche pour que cette partie de l'analyse soit aussi effectuée à l'aide d'algorithmes à fourmis par l'exploitation d'informations disponibles sous formes de réseaux lexicaux. Le texte de départ prend alors la forme d'une chaîne de termes, à partir de laquelle une analyse globale est effectuée (d'où le terme d'holistique). Nous cherchons à nous affranchir de la notion de phase d'analyse (classiquement morphologique, syntaxique, sémantique, logique, etc.) et à viser davantage la résolution de microphénomènes qui, globalement, permettrait de résoudre tout ou partie des problèmes rencontrés. Enfin, des expériences préliminaires ont semblé démontrer que l'adjonction de relations négatives, induisant des phénomènes d'inhibition, augmente de façon très significative à la fois la qualité du résultat et la vitesse de convergence.

Les algorithmes à fourmis exploitant le couplage réseau/vecteurs peuvent être exploités en *ingénierie des modèles* afin de faire du calcul de similarité entre classes (et attributs, ou méthodes) (travaux de R. Falleri avec M. Huchard, [Falleri *et al.*, 2010] et [Falleri *et al.*, 2009]). Il s'agit d'adjoindre des informations lexicales et ontologiques à des processus qui en disposent d'assez peu traditionnellement, et ainsi d'être capable d'effectuer automatiquement des fusions partielles de modèles par identification des objets similaires. Les processus en jeu ici font aussi intervenir des bouclages entre les corpus (ici des modèles de classes) et les sous-réseaux lexicaux construits. Ces travaux se poursuivent dans la direction de la construction d'ontologies de domaines de spécialité, de façon non négociée (dans l'esprit du consensus populaire de JeuxDeMots).

La taille du réseau lexical de JeuxDeMots (plus de 1 200 000 relations, entre 100 000 termes) semble constituer un matériau intéressant afin de mettre à l'épreuve et concevoir de nouveaux algorithmes de visualisation et/ou de classification de lexique (travaux avec M. Hascoët et G. Artignan, [Artignan *et al.*, 2009]). Du point de vue de la clusterisation de termes, les algorithmes doivent faire face à la polysémie des termes du réseau, et reconstituer une ontologie acceptable, à défaut d'être identique à une ontologie classique, ne semble pas trivial.

Un certain nombre d'idées, pour certaines transversales aux sujets abordés, est développé au long de ce document :

> ***Complémentarité***. *Les structures vectorielles, les structures de graphes et les structures ensemblistes pour la représentation en sémantique lexicale sont complémentaires.*
>
> ***Consensus populaire***. *Acquérir des informations lexicales et en particulier des relations entre mots, à l'aide de jeux de consensus populaire non négocié, est une approche opérationnelle.*
>
> ***Acquisition permanente***. *L'acquisition d'information lexicale peut et a intérêt à se faire de façon itérée au sein d'un processus permanent.*
>
> ***Raffinement des structures par bouclage***. *L'identification des usages de termes et le raffinement des relations peut et a intérêt à s'inscrire au sein d'une boucle entre les utilisateurs et les processus.*
>
> ***Activation/inhibition***. *L'analyse sémantique de texte profite au moins autant de l'activation des relations entre termes que de leur inhibition.*

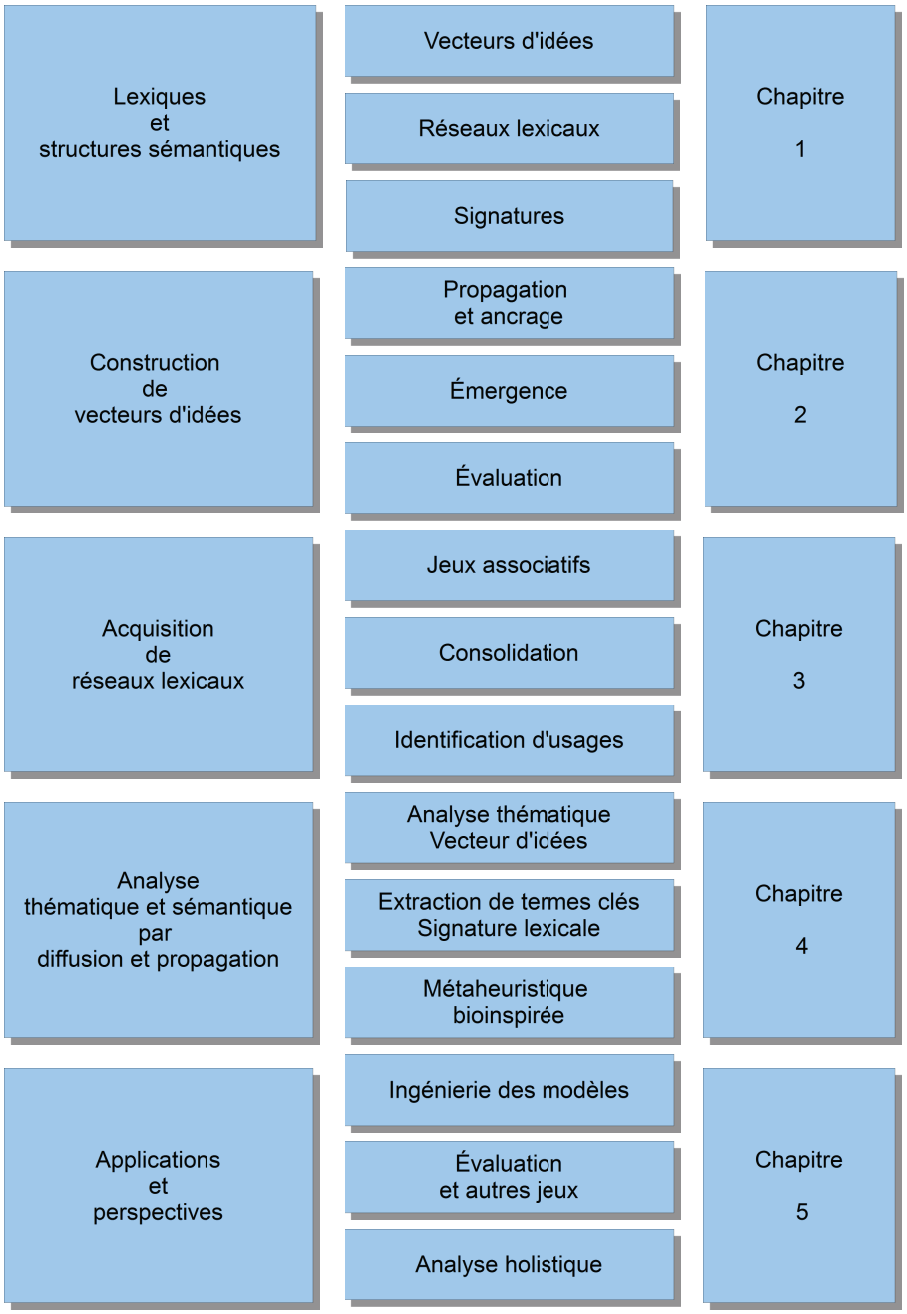

FIGURE 1 – Organisation des chapitres de ce mémoire