Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience

Ronan Pichon et Pascale Sébillot

IRISA
Campus de Beaulieu
35042 Rennes cedex
rpichon@irisa.fr, sebillot@irisa.fr

Résumé

Dans cet article, nous montrons, à travers l'exposé de résultats d'une expérience menée sur corpus, comment la connaissance des thèmes dans lesquels apparaissent des mots et la mise en évidence de similarités et de différences entre les voisinages de leurs occurrences dans les parties de textes abordant ces thèmes permettent de mettre au jour des différences fines dans les acceptions associées aux mots dans chacun de ces thèmes. La méthode proposée pour ce faire est presque entièrement automatique et est basée sur le calcul d'intersections et de différences ensemblistes entre des séquences de mots constituant des contextes.

1. Introduction

Dans (Kilgarriff, 1998), A. Kilgarriff examine les différents types de ressources lexicales existantes, selon les objectifs qu'elles visent et les moyens mis en œuvre pour les constituer. Il constate qu'il existe encore un large fossé entre les ressources établies à partir d'un traitement automatique de corpus, en vue, par exemple, de désambiguïser le sens des mots (Yarowsky, 1995) ou d'extraire des informations (Riloff, 1996), et les lexiques construits manuellement sur des bases plus formelles, par exemple à des fins lexicographiques. Ce fossé se situe surtout au niveau de l'abstraction trop faible du premier type de lexiques, qui sont trop spécifiques ou trop liés à l'utilisation visée. Les informations obtenues par extraction automatique sont encore globalement incomplètes par rapport à une description des mots que donnerait un lexicographe, en ce sens qu'elles n'expriment pas toute la richesse du sens des mots.

Cependant, A. Kilgarriff estime que ce fossé est amené à se combler grâce aux progrès de la recherche, notamment dans les domaines de l'acquisition lexicale et de l'apprentissage automatique. Une des clés pour construire des lexiques pertinents est, tant pour lui que pour D.A. Cruse (Cruse, 1986), B. Pottier (Pottier, 1992) ou F. Rastier (Rastier, 1991), de baser leur élaboration sur une analyse experte des usages constatés des mots dans les textes. L'entrée lexicale associée à chacun des éléments du lexique est construite par une synthèse ou une abstraction de ces usages (complétée si nécessaire par un expert).

Nous partageons ce point de vue et notre objectif, dans cet article, est de montrer, à travers la description des résultats d'une expérience que nous avons réalisée, qu'il est possible de mettre en évidence automatiquement des différences fines entre les significations d'un mot lorsqu'il est utilisé dans différents thèmes au sein d'un corpus, en se basant sur la connaissance des thèmes dans lesquels apparaissent ses diverses occurrences et en utilisant les différences entre les voisinages de ces occurrences dans les divers thèmes. Nos travaux, outre leur placement dans le cadre général de la *Linguistique Harrissienne* (Harris *et al.*, 1989), s'inspirent globalement d'idées proposées par F. Rastier (Rastier *et al.*, 1994; Rastier, 1996).

Afin de préciser davantage notre but, nous nous positionnons, dans un premier temps, par rapport à des travaux dont l'objectif est également d'acquérir automatiquement des éléments de sens sur corpus, et qui se basent aussi sur le contexte des occurrences des mots pour ce faire. Nous ne cherchons pas ici à faire un tour d'horizon complet de ces travaux (cf. par exemple (Habert *et al.*, 1997) ou (Pichon & Sébillot, 1997) pour une vue plus globable, ou (Wilks *et al.*, 1996) pour, entre autres, une comparaison entre les informations lexicales qui peuvent être extraites d'un dictionnaire électronique et d'un corpus) mais à mettre l'accent sur les principales classes de recherches effectuées et à les commenter par rapport à notre objectif. Deux grandes catégories apparaissent, selon les buts visés par ces travaux :

1. *Reconnaître les différents sens ou acceptions d'un mot associé à chacune de ses occurrences.* C'est notamment le cas de tous les travaux effectués dans le domaine de la désambiguïsation du sens des mots (Yarowsky, 1995; Schütze, 1998). En général, la démarche consiste à regrouper les occurrences en fonction de leur contexte, chaque regroupement ainsi obtenu étant ensuite associé à un ou des sens prédéfinis du mot étudié.

   L'objectif de ces travaux n'est pas, de façon précise, de se poser la question du sens des mots, mais plus celui du rattachement d'occurrences à des étiquettes (Wilks & Stevenson, 1997). De plus, il y a parfois des problèmes pour trouver un sens lexicographique à une occurrence de mot dans un texte, car certaines nuances dans l'interprétation peuvent ne pas être représentées dans un dictionnaire.

   Ces travaux se rapprochent de ceux que nous décrivons ici en ce sens qu'ils lient eux aussi la différence de significations des mots à la différence de contextes. Toutefois ils supposent des étiquettes pré-établies, et n'analysent pas de manière plus précise a posteriori les contextes qui ont servi à constituer les regroupements.

2. *Constituer des classes sémantiques.* On aborde ici un problème différent, où il s'agit de constituer des ensembles de mots qui forment autant de classes sémantiques, dans lesquelles les mots peuvent être structurés à l'aide de relations lexicales classiques (Grefenstette, 1993; Riloff, 1996).

   Cependant, pour beaucoup d'expériences relatées dans ces travaux, les mots étudiés sont considérés comme monosémiques, toutes les occurrences d'un même mot étant appelées à appartenir à une même classe sémantique. De plus (et indépendamment du caractère monosémique ou non des mots étudiés), la granularité de la description du sens d'un mot dans une classe est assez grossière, car il n'y a pas de recherche d'information plus précise que cette appartenance à la classe. Enfin, là aussi, les contextes utilisés pour la constitution des classes n'ont pas de rôles ultérieurs.

Notre objectif est donc différent de celui de ces deux familles de travaux. D'une part, nous ne supposons pas la connaissance de sens pré-établis à découvrir et, d'autre part, nous voulons obtenir une description précise des similarités et différences de sens entre les occurrences d'un même mot, et ceci, même au sein d'une classe sémantique. Toutefois, comme la composition

des classes sémantiques varie en fonction des différents thèmes abordés dans un texte (Rastier, 1995; Rastier, 1996), nous avons choisi de nous placer d'abord au niveau de l'appartenance d'une occurrence de mot à un thème (en termes d'apparition dans une partie du texte abordant ce thème) plutôt qu'au niveau de son appartenance à une classe sémantique. Par thèmes, nous entendons les « sujets » abordés dans les textes ou les segments de textes d'un corpus. Nous utilisons donc la connaissance des thèmes dans lesquels apparaissent des mots, ainsi que les voisinages de leurs occurrences dans les parties des textes abordant ces thèmes, pour montrer des similarités et des différences entre les sens de ces mots grâce aux similarités et aux différences entre les voisinages. Pour ce faire, nous comparons les contextes associés aux mots dans un thème et essayons de trouver des éléments de contexte caractéristiques du mot dans le thème, et donc du sens des mots qui y sont utilisés. Notre objectif est donc de mettre au jour des blocs (ou parties) de contextes caractéristiques d'un aspect de sens particulier, et, dans cet article, nous montrons plus précisément, à travers l'exposé des résultats d'une expérience, que de tels blocs existent et qu'il est possible de les déterminer de façon presque entièrement automatique. Le but à plus long terme de ce travail est d'étudier le lien entre ces blocs de contextes et les relations sémantiques entre les mots dans le lexique.

Nous débutons cet article par la description du cadre de l'expérience que nous avons réalisée : corpus, prétraitement, obtention des thèmes abordés dans le corpus, méthode utilisée pour trouver les contextes des occurrences de mots. Nous présentons ensuite les résultats obtenus en étudiant les nuances de sens mises automatiquement en évidence par l'étude des similarités et différences d'une part, entre les contextes des diverses occurrences du même mot au sein de thèmes différents et, d'autre part, entre les contextes de divers mots apparaissant dans le même thème. La conclusion discute la qualité des résultats obtenus et présente les perspectives de ce travail.

2. Contexte de l'expérience

Nous travaillons sur un corpus constitué d'articles du journal LE MONDE DIPLOMATIQUE. Ce corpus comprend environ 7,8 millions de mots. Il a été segmenté, lemmatisé et désambiguïsé[1].

Nous avons effectué une recherche préalable et automatique des thèmes abordés sur un sous-ensemble du corpus (1 million de mots environ). Plus précisément, nous n'avons extrait que les thèmes principaux du corpus, certains d'entre eux pouvant donc être une abstraction de plusieurs sujets effectivement rencontrés dans le corpus. Pour atteindre cet objectif, nous avons étudié à l'aide d'une méthode d'analyse des données, l'AVL[2] (Lerman, 1991), la distribution des noms les plus fréquents du sous-corpus dans l'ensemble des paragraphes. Cette analyse nous a permis de mettre en évidence un ensemble de thèmes différents traités dans le corpus, tels que : PRESSE, INSTITUTIONS, FINANCE, SITUATIONS DE CRISES, etc. Chacun de ces thèmes est associé à un petit ensemble de mots-clés déterminés automatiquement grâce à la méthode AVL. La combinaison de ces derniers permet de caractériser et de nommer les thèmes (par exemple le thème PRESSE est représenté par l'ensemble {*journaliste, journal, presse*}), ainsi que de détecter leur présence au sein d'une unité de texte[3]. Dans la recherche des thèmes du corpus, nous avons considéré chaque paragraphe comme une de ces unités. Nous associons un

1. La segmentation a été effectuée par le logiciel MtSeg, et la lemmatisation et l'étiquetage par MtLex, développés à l'université d'Aix-en-Provence. La désambiguïsation a été faite à l'aide du logiciel Tatoo de l'ISSCO.

2. Analyse de la Vraisemblance des Liens.

3. Nous nous inspirons de la notion d'isotopie développée par F. Rastier (Rastier, 1996).

thème à un paragraphe si au moins deux des mots du paragraphe correspondent à des éléments de l'ensemble qui caractérise ce thème. Nous nous basons, pour ces dernières hypothèses, sur l'observation du corpus utilisé. Le fait de rechercher la co-présence de plusieurs mots-clés d'un thème dans un segment réduit les erreurs qui peuvent être liées à leur ambiguïté. Par ailleurs, cette méthode autorise un même segment de texte à être éventuellement rattaché à plusieurs thèmes.

Dans cet article, nous nous intéressons à l'étude de la variation des interprétations des différentes occurrences des mots selon les thèmes des unités de textes dans lesquelles elles apparaissent. Pour illustrer notre propos, nous avons choisi deux des thèmes détectés automatiquement dans le corpus, à savoir : NÉGOCIATIONS, associé à l'ensemble de mots caractéristiques {*négociation, accord, création, position*}, et TERRITOIRE, qui correspond à la notion d'étendue de terre dépendant d'un groupe humain ou d'une juridiction et est associé à l'ensemble {*autorité, région, territoire*}[4].

Nous repérons les occurrences de chacun des mots que nous voulons étudier dans les unités du corpus relevant de ces thèmes, et nous extrayons leurs voisinages constatés dans les textes. Le voisinage utilisé est défini par les cinq[5] mots qui précèdent et les cinq mots qui suivent chacune de ces occurrences. Pour chaque mot étudié, nous retenons alors un sous-ensemble des mots apparaissant dans l'ensemble des voisinages de ses diverses occurrences au sein d'un thème, chaque mot de ce sous-ensemble étant associé à sa fréquence d'apparition dans l'ensemble des voisinages ; le sous-ensemble correspond à une restriction de l'ensemble complet des mots présents dans les voisinages constatés aux seuls noms et adjectifs. Les verbes ont été volontairement écartés car la forme de notre représentation du contexte ne tient pas compte de la position du mot étudié par rapport aux verbes, ce qui augmente trop fortement leur ambiguïté. Nous ne retenons finalement, pour chaque couple (mot étudié, thème), que les éléments les plus fréquents dans la liste qui décrit son voisinage. Pour ce faire, nous fixons à chaque liste un seuil n tel que $n > 2$ et que le nombre des éléments de la liste des mots dont la fréquence est supérieure à n soit le plus proche possible de vingt[6].

Les voisinages simplifiés ainsi constitués nous servent à étudier les similarités et dissimilarités de sens entre deux occurrences d'un même mot dans deux thèmes différents, ou entre deux mots dans un même thème. Cette étude se fait par calcul de l'intersection ou de la différence ensembliste entre les voisinages simplifiés. Notre but est d'interpréter les ensembles de mots ainsi obtenus en y recherchant des séquences caractérisant une différence entre la signification de mots. Les membres d'une séquence ont la particularité de posséder un élément de sens en commun, ce qui implique leur désambiguïsation implicite. Les différents éléments de sens ainsi mis en évidence sont associés au mot étudié. Les mots éléments de voisinage trop ambigus ou isolés ne sont pas pris en compte. Cette partie interprétative se fait, quant à elle, manuellement.

3. Résultats

Pour illustrer notre propos, nous présentons maintenant quelques exemples de résultats obtenus. Nous comparons tout d'abord les voisinages constatés d'un mot rencontré dans deux

4. Remarque : les paragraphes qui sont reconnus comme étant conjointement dans les deux thèmes parlent de négociations dont l'enjeu est un territoire.

5. Taille fréquemment utilisée par défaut pour ce type de travaux depuis (Church *et al.*, 1993).

6. Cette valeur est entièrement paramétrable. Le nombre actuel a été choisi pour pouvoir juger facilement de la qualité des résultats obtenus, en termes d'interprétation ; cette phase s'effectuant actuellement à la main, nous avons donc voulu n'avoir qu'un nombre restreint de mots.

thèmes différents du corpus. Nous poursuivons par la présentation d'un ensemble de mots différents dans un thème donné.

Le premier mot étudié que nous présentons est ***militaire***. Nous repérons donc, pour les 2 thèmes choisis NÉGOCIATIONS et TERRITOIRE, selon la méthode décrite ci-dessus, les noms et adjectifs apparaissant le plus souvent dans les voisinages des occurrences de ce mot dans les unités de textes portant sur chacun de ces 2 thèmes. On obtient ainsi par ordre de fréquence croissante :

1. pour le thème TERRITOIRE : aide 4[7], *États-Unis* 4, *grand* 4, *moyen* 4, *opération* 4, *régime* 4, *russe* 4, *victoire* 4, *base* 5, *massif* 5, *occupation* 5, *intervention* 6, *puissance* 6, *américain* 7, *économique* 7, *présence* 9, *force* 12, *politique* 13
2. pour le thème NÉGOCIATIONS : *action* 5, *grand* 5, *ordre* 5, *puissance* 5, *effort* 6, *États-Unis* 6, *responsable* 6, *pays* 7, *dépense* 8, *force* 8, *intervention* 8, *atlantique* 9, *Europe* 9, *OTAN* 9, *présence* 9, *aide* 10, *américain* 10, *économique* 10, *organisation* 12, *politique* 22

On voit immédiatement un ensemble de mots communs aux deux voisinages : *force, États-Unis, américain, grand, puissance, économique, intervention, politique, présence, aide.* Il nous reste alors, pour distinguer les usages de ***militaire*** entre ces deux thèmes, les mots suivants :

1. pour le thème TERRITOIRE : *moyen* 4, *opération* 4, *régime* 4, *russe* 4, *victoire* 4, *base* 5, *massif* 5, *occupation* 5
2. pour le thème NÉGOCIATIONS : *action* 5, *ordre* 5, *effort* 6, *responsable* 6, *pays* 7, *dépense* 8, *atlantique* 9, *Europe* 9, *OTAN* 9, *organisation* 12

Dans le thème TERRITOIRE, les éléments de contextes qui lui sont propres indiquent une connotation essentiellement guerrière de ***militaire***, par la présence des mots {*opération, massif, occupation, victoire, moyen*}. Le thème NÉGOCIATIONS met, quant à lui, davantage en avant le côté organisé et structuré attaché au mot, par la présence notamment des mots {*organisation, OTAN, atlantique, dépense, responsable, ordre*}.

De même pour le mot ***guerre***, on obtient les ensembles suivants :

1. pour le thème TERRITOIRE : *américain* 3, *début* 3, *israélo-arabe* 3, *nouveau* 3, *Tchétchénie* 3, *Turc* 3, *Vietnam* 3, *Washington* 3, *interminable* 4, *Irak* 4, *acquisition* 5, *Israël* 5, *jour* 5, *lendemain* 5, *régional* 6, *premier* 9, *froid* 10, *territoire* 10, *civil* 11, *Golfe* 14, *second* 14, *mondial* 17
2. pour le thème NÉGOCIATIONS : *conflit* 3, *début* 3, *Israël* 3, *long* 3, *paix* 3, *premier* 3, *année* 4, *an* 4, *étoile* 4, *Liban* 4, *vainqueur* 4, *nouveau* 5, *commercial* 7, *économique* 7, *second* 7, *lendemain* 8, *mondial* 10, *civil* 16, *froid* 22, *Golfe* 22

Outre les références communes qui se rapportent à des guerres très connues (les guerres mondiales, la guerre du golfe, la guerre froide), on peut aussi mettre au jour des différences entre les voisinages constatés de ***guerre*** dans les deux thèmes :

1. TERRITOIRE : *américain* 3, *israélo-arabe* 3, *Tchétchénie* 3, *Turc* 3, *Vietnam* 3, *Washington* 3, *interminable* 4, *Irak* 4, *acquisition* 5, *jour* 5, *régional* 6, *territoire* 10

7. Le nombre qui suit le mot indique le nombre de fois où ce mot apparaît dans le voisinage étudié. Ici, aide apparaît quatre fois dans le voisinage de ***militaire***.

2. NÉGOCIATIONS : *conflit* 3, *long* 3, *paix* 3, *année* 4, *an* 4, *étoile* 4, *Liban* 4, *vainqueur* 4, *commercial* 7, *économique* 7

Dans le premier cas, on voit que la notion de ***guerre*** est très fortement liée à un lieu ({*Tchétchénie, Irak, acquisition, territoire*}) et à ses acteurs, qui sont fréquemment représentés par des noms de pays également, et, dans ce thème TERRITOIRE, la signification de guerre est essentiellement celle d'un conflit armé. Dans la seconde thématique apparaissent des enjeux plus « abstraits » ({*paix, vainqueur*}) ou bien des types particuliers de ***guerre*** ({*commercial, économique*}). Ce dernier point s'explique aisément en prenant en compte le fait que la fin d'une guerre économique ou d'une guerre (militaire) est souvent l'objet de négociations.

Enfin, si nous étudions le comportement du mot ***économie***, on obtient :

1. TERRITOIRE : *comptoir* 3, *dépendant* 3, *développement* 3, *palestinien* 3, *pays* 3, *région* 3, *israélien* 4, *local* 4, *place* 4, *territoire* 4, *marché* 5
2. NÉGOCIATIONS : *capitaliste* 3, *État* 3, *place* 3, *planification* 3, *développement* 4, *palestinien* 4, *ressource* 4, *secteur* 4, *ministre* 5, *politique* 5, *pays* 8, *marché* 12, *mondial* 12

Si on fait abstraction des quelques mots communs aux deux ensembles ({*développement, palestinien, pays, marché*}), on remarque que le premier thème fait essentiellement appel à l'économie d'un lieu, alors que le second porte davantage sur les mécanismes de fonctionnement et de régulation de l'économie.

Nous présentons maintenant l'étude comparée des voisinages constatés de trois mots différents, mais dont des sens se recouvrent, à savoir ***pouvoir***, ***autorité*** et ***gouvernement***, dans chacun des deux thèmes proposés. Le but de cette seconde partie de l'expérience est, pour des mots candidats à appartenir à une même classe sémantique dans chaque thème étudié, de chercher à déterminer ce qui les rassemble d'une part, mais surtout ce qui pourrait les différencier.

1. TERRITOIRE :
   (a) ***pouvoir*** : *état* 7, *local* 7, *soviétique* 7, *année* 8, *exécutif* 9, *parti* 9, *prise* 9, *public* 10, *économique* 11, *président* 11, *nouveau* 12, *place* 12, *arrivée* 17, *politique* 21, *central* 36
   (b) ***autorité*** : *Pékin* 4, *place* 4, *président* 4, *preuve* 4, *région* 4, *transfert* 4, *chinois* 5, *état* 5, *nouveau* 5, *territoire* 5, *gouvernement* 6, *politique* 6, *israélien* 13, *palestinien* 13, *local* 16
   (c) ***gouvernement*** : *fédéral* 7, *occidental* 7, *président* 8, *français* 9, *ministre* 9, *régional* 9, *union* 9, *formation* 10, *politique* 12, *nouveau* 14, *central* 15, *national* 16, *israélien* 32

   Dans le thème du territoire, on retrouve peu d'éléments de voisinage communs à ces trois mots : {*nouveau, politique, président*}. Toutefois, les mots désignant l'étendue géographique ou institutionnelle sur laquelle l'***autorité***, le ***pouvoir*** ou le ***gouvernement*** exerce son autorité sont très fréquents, sans qu'il s'agisse des mêmes pour chacun de ces trois mots ; ainsi {*local, central*} pour ***pouvoir***, {*région, territoire, local*} pour ***autorité*** et {*fédéral, régional, union, central, national*} pour ***gouvernement*** forment un ensemble cohérent par rapport à cette notion d'étendue géographique (particulièrement pour ***autorité***) ou institutionnelle (particulièrement pour ***gouvernement***).
   Parmi les différences, on peut noter que l'***autorité*** est très associée à *local* quand ***pouvoir*** et ***gouvernement*** sont fortement liés à *central*, ce qui amène à penser que l'***autorité*** est

subordonnée à un ***gouvernement*** ou un ***pouvoir***. Par ailleurs, la co-présence spécifique de {*fédéral, national*} d'une part, et {*ministre, union, formation*} d'autre part, indique que le ***gouvernement*** exerce son autorité dans un cadre institutionnel bien défini et structuré, alors que ***pouvoir*** et ***autorité*** impliquent un cadre plus informel. L'***autorité*** est très liée à la notion de territoire : {*région, territoire, local*}, alors que le ***pouvoir*** s'exerce sur autre chose qu'un territoire : {*public, économique, exécutif*} et semble indiquer une entité plus changeante : {*place, prise, arrivée, année*} que ***gouvernement*** ou ***autorité***.

2. NÉGOCIATIONS :
   (a) ***pouvoir*** : *accession* 8, *an* 8, *armée* 8, *concentration* 8, *pays* 8, *nouveau* 9, *place* 9, *coalition* 10, *contrôle* 10, *gouvernement* 10, *arrivée* 16, *état* 17, *partage* 17, *parti* 17, *achat* 22, *central* 22, *public* 27, *économique* 28, *politique* 50
   (b) ***autorité*** : *américain* 3, *frontière* 3, *local* 3, *nouveau* 3, *pays* 3, *pouvoir* 3, *problème* 3, *provisoire* 3, *armée* 4, *autonome* 4, *Cisjordanie* 5, *élu* 5, *état* 5, *gouvernement* 5, *politique* 7, *palestinien* 8
   (c) ***gouvernement*** : *actuel* 11, *opposition* 11, *position* 11, *premier* 11, *sandiniste* 12, *accord* 13, *membre* 13, *national* 13, *central* 14, *chef* 14, *occidental* 14, *état* 16, *Bonn* 17, *coalition* 17, *formation* 18, *français* 25, *américain* 26, *nouveau* 26, *pays* 26, *fédéral* 27, *européen* 28, *politique* 37, *israélien* 61

   Les mots communs aux voisinages de ces trois mots sont : {*état, nouveau, pays, politique*} ; la présence des premier et troisième éléments montre qu'un *état* et qu'un *pays* sont représentés par un ***pouvoir***, une ***autorité*** ou un ***gouvernement***.
   Le fait que ***gouvernement*** soit présent dans les séquences de mots associées à ***autorité*** et ***pouvoir*** laisse apparaître une idée de hiérarchie entre ces 2 groupes. Par ailleurs, dans les éléments de voisinage propres à ***autorité***, on retrouve des mots qui évoquent les domaines où elle s'exerce : {*frontière, armée, Cisjordanie, local*}, ainsi qu'une notion de précarité : {*problème, provisoire, autonome*}. Le ***gouvernement*** indique lui une notion de représentation de quelque chose : {*sandiniste, occidental, français, américain, européen, israélien*}, d'institution : {*fédéral, national*}, ainsi que de structuration : {*chef, membre, coalition, formation*}. Quant à ***pouvoir***, il indique, comme dans le thème TERRITOIRE, une notion de domaines de compétence : {*public, économique, politique*}, de changement : {*accession, an, arrivée, partage*}, mais aussi, de structuration : {*parti, coalition, armée*}.

Nous pouvons utiliser ces résultats pour bâtir une représentation partielle des significations de ***pouvoir, autorité*** et ***gouvernement*** dans les thèmes TERRITOIRE et NÉGOCIATIONS. Le tableau 1 en est une première vue synthétique, dans laquelle chaque colonne représente une abstraction de l'élément de sens associé aux séquences extraites (cf. ci-dessus) sous la forme d'un sème. Le signe "+" indique la participation du sème à la signification du mot dans le thème abordé, et "-" son absence de participation.

La figure 1 présente, quant à elle, la même information sous la forme d'un réseau sémantique, dans lequel les nœuds indiquent la signification d'un mot dans un thème, et les arcs orientés entre ces nœuds indiquent en quoi chaque signification diffère d'une autre. De façon plus précise, un arc orienté étiqueté /sème/ entre les nœuds A et B indique que le sème /sème/ participe à la signification de A et pas à celle de B. Une telle représentation nous permet, en usant des mécanismes décrits dans (Rastier, 1996), lors de la lecture de la phrase : « Dans le cadre des négociations de l'OMC, le gouvernement X et le pouvoir Y sont parvenus à un accord. », qui se situe dans le thème des NÉGOCIATIONS, de déterminer que la nation X est *représentée* dans

	/représentant/	/domaine/	/changeant/	/institution/	/étendue/	/structure/
gouvern$_{TER}$	-	-	-	+	+	+
pouvoir$_{TER}$	-	+	+	-	+	-
autorité$_{TER}$	-	+	-	-	+	-
gouvern$_{NEG}$	+	-	-	+	-	+
pouvoir$_{NEG}$	+	+	+	-	-	+
autorité $_{NEG}$	+	+	+	-	-	-

TAB. 1 – *Exemple de tableau d'analyse sémique*

ces négociations par son gouvernement, alors que le pouvoir Y est interprété comme une entité plus instable et à l'autorité plus circonscrite.

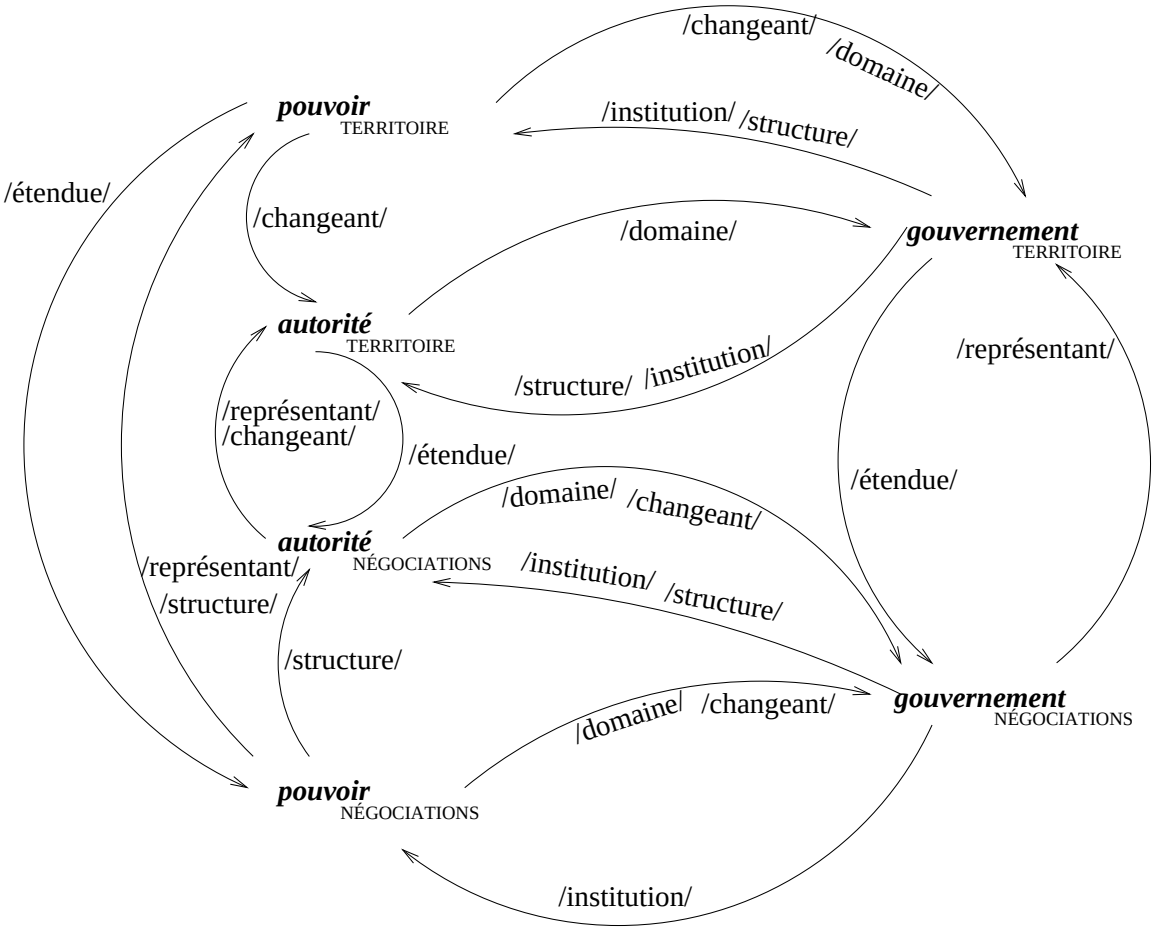

FIG. 1 – *Un exemple de représentation lexicale*

4. Discussion et perspectives

Les différents résultats de l'expérience que nous avons décrite et dont nous avons présenté quelques exemples dans la section précédente montrent qu'il est possible de mettre au jour des séquences de mots ou blocs de contextes caractéristiques des différences de significations entre des mots dans des thèmes, en se basant sur la connaissance des thèmes dans lesquels apparaissent les diverses occurrences de ces mots et en étudiant les différences entre les voisinages

de ces occurrences dans les divers thèmes. En mettant l'accent sur ces différences, tant entre des occurrences d'un même mot au sein de deux thèmes distincts qu'entre des mots différents au sein du même thème, on participe à la construction d'une représentation du sens.

Cette mise en évidence de ce qui rassemble et distingue des occurrences d'un même mot dans deux thèmes est entièrement automatisée et est basée sur le calcul d'intersections et de différences ensemblistes entre les listes de mots formant le contexte de chaque occurrence dans les thèmes considérés. Seule l'interprétation, c'est-à-dire le nommage de la ressemblance ou de la différence mise au jour, est manuelle. Pour ce qui concerne la mise en évidence de points communs et d'éléments distinctifs entre des mots différents dans un même thème, seule la première partie est entièrement automatique et est basée sur les mêmes méthodes que pour le cas précédent. Les sous-listes caractérisant une différence particulière doivent, quant à elles, être extraites manuellement des contextes associés aux mots. L'interprétation est également manuelle.

Nous envisageons de développer le travail présenté dans trois directions complémentaires. La première concerne la recherche d'une amélioration des contextes extraits. La façon dont est effectué le typage des voisinages constatés au sein du corpus peut, en effet, être affinée, en prenant, par exemple, en compte la position de l'élément de voisinage par rapport au mot étudié, en détectant la présence éventuelle d'expressions complexes parmi ces éléments de voisinage, ou en considérant des relations de dépendances syntaxiques locales comme cela est fait dans (Fabre *et al.*, 1997). Ces travaux sont, à notre connaissance, ceux dont la démarche est la plus proche de la nôtre. En effet, les mots étudiés sont aussi mis en relation par les contextes constatés dans des corpus, mais il s'agit, dans ce cas, de contextes constatés au sein de groupes nominaux extraits du corpus. Ces contextes servent à étiqueter les arcs d'un graphe dont les nœuds sont les mots étudiés. Les auteurs cherchent dans ce graphe des cliques ou des composantes connexes qu'ils interprètent manuellement à l'aide d'experts. Parmi les remarques formulées sur leurs résultats, ils expliquent que l'interprétation des ensembles obtenus dans un corpus de langue non spécialisée est fortement dépendante de la connaissance du thème du discours dont les contextes sont extraits. Cette constatation faite a posteriori par ces auteurs est en fait un point central dans nos travaux. Cependant, s'ils expriment la volonté d'enrichir les types de contextes qu'ils manipulent, c'est par l'intermédiaire de structures syntaxiques marqueurs (systématiques) d'une relation sémantique, et en cherchant des relations de dépendances syntaxiques autres que celles exprimées dans les seuls groupes nominaux.

La seconde perspective concerne l'automatisation de l'extraction des sous-séquences marqueuses d'un élément de différence entre des mots dans un même thème. Nous envisageons, pour ce faire, d'utiliser les séquences discriminantes entre les occurrences de mêmes mots dans des thèmes distincts d'un domaine globalement homogène, sur un nombre conséquent de mots, pour repérer des séquences de mots dont la co-présence est caractéristique de chacun de ces thèmes. C'est une démarche analogue que suit l'expert humain dans la méthodologie d'analyse conceptuelle exposée dans (Assadi, 1998) pour préciser et compléter les champs conceptuels mis en évidence par une classification automatique effectuée sur les termes d'un domaine spécialisé. Toutefois, l'objectif de ces derniers travaux, à la différence des nôtres, est de bâtir l'ontologie d'un domaine à partir d'un corpus de textes techniques.

Enfin, la troisième perspective est un peu plus lointaine et concerne la mise en relation entre les blocs de contextes marqueurs de distinctions ou de similarités de sens entre les mots et les relations que cela doit servir à mettre effectivement à l'œuvre dans un lexique sémantique dont on chercherait à automatiser le plus possible la construction.

Références

ASSADI H. (1998). *Construction d'ontologies à partir de textes techniques - Application aux systèmes documentaires*. PhD thesis, Université Paris 6.

CHURCH K. W., GALE W. A. & YAROWSKY D. (1993). A Method for Disambiguating Word Senses in a Large Corpus. *Computer and the Humanities*, **26**, 415–439.

CRUSE D. A. (1986). *Lexical Semantics*. Cambridge Textbooks in Linguistics.

FABRE C., HABERT B. & LABBÉ D. (1997). La polysémie dans la langue générale et les langages spécialisés. *Sémiotiques*, **13**.

GREFENSTETTE G. (1993). Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches. In *Proceedings of Workshop on Acquisition of Lexical Knowledge from Text, SIGLEX/ACL*, Columbus, États-Unis.

HABERT B., NAZARENKO A. & SALEM A. (1997). *Les linguistiques en corpus*. Paris: Armand Colin.

HARRIS Z., GOTTFRIED M., RYCKMAN T., MATTICK(JR) P., DALADIER A., HARRIS T. & HARRIS S. (1989). The Form of Information in Science, Analysis of Immunology Sublanguage. *Boston Studies in the Philosophy of Science*, **104**.

KILGARRIFF A. (1998). Bridging the Gap Between Lexicon and Corpus: Convergence of Formalisms. In *Proceedings of Workshop on Adaptating Lexical and Corpus Ressources*, Grenade, Espagne.

LERMAN I.-C. (1991). Foundations in the Likehood Linkage Analysis Classification Method. *Applied Stochastic Models and Data Analysis*, **7**, 69–76.

PICHON R. & SÉBILLOT P. (1997). *Acquisition automatique d'informations lexicales à partir de corpus : un bilan*. Rapport de Recherche n^{0}3321, INRIA.

POTTIER B. (1992). *Sémantique Générale*. Presses Universitaires de France.

RASTIER F. (1991). *Sémantique et recherches cognitives*. Presses Universitaires de France.

RASTIER F. (1995). *L'analyse thématique des données textuelles*. Didier.

RASTIER F. (1996). *Sémantique Interprétative*. Presses Universitaires de France.

RASTIER F., CAVAZZA M. & ABEILLÉ A. (1994). *Sémantique pour l'analyse : de la linguistique à l'informatique*. Masson.

RILOFF E. (1996). An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *AI journal*, **85**.

SCHÜTZE H. (1998). Word Sense Discrimination. *Computational Linguistics*, **24**(1), 97–124.

WILKS Y., SLATOR B. & GUTHRIE L. (1996). *Electric Words: Dictionaries, Computers, and Meanings*. Bradford.

WILKS Y. & STEVENSON M. (1997). Sense Tagging: Semantic Tagging with a Lexicon. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics"*, Washington, États-Unis.

YAROWSKY D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, États-Unis.