

Aide à la conception de méthodes de classification pour la construction d'ontologies : l'atelier Mo'K

Gilles Bisson
Laboratoire LEIBNIZ
CNRS-INPG-UJF
Gilles.Bisson@imag.fr

Claire Nédellec
Equipe Inférence et Apprentissage,
LRI, Université Paris-Sud
Claire.Nedellec@lri.fr

Résumé. Le système Mo'K est un atelier configurable d'aide à la conception d'algorithmes de classification conceptuelle utilisable dans le cadre de l'apprentissage d'ontologies. Il vise à faciliter le travail exploratoire que doit effectuer tout concepteur d'algorithme pour rechercher les méthodes d'apprentissage les plus aptes à accomplir une tâche donnée. Pour ce faire, Mo'K fournit non seulement des services permettant d'évaluer, de comparer et de caractériser les méthodes de classification, mais il offre également une approche générique permettant une implémentation aisée des mesures de similarité et opérateurs de classification proposés dans la littérature. Nous présentons ici quelques résultats expérimentaux mettant en lumière les possibilités du système pour valider des méthodes de construction de classes sémantiques à partir de corpus analysés.

Abstract. This paper describes Mo'K, a configurable workbench that supports the development of conceptual clustering methods for ontology building. Mo'K is intended to assist ontology developers in the exploratory process of defining the most suitable learning methods for a given task. To do so, it provides facilities for evaluation, comparison and elaboration of conceptual clustering methods. Also, the model underlying Mo'K permits a fine-grained definition of similarity measures and class construction operators, easing the tasks of method instantiation and configuration. This paper presents some experimental results that illustrate the suitability of the model to help characterize and assess the performance of different methods that learn semantic classes from parsed corpora

1. Introduction

Dans cet article, nous proposons un atelier de mise au point de méthodes de classification permettant de former de manière semi-automatique des ontologies de type hiérarchie conceptuelle, à partir de corpus analysés. De manière générale, la définition d'une méthode de classification conceptuelle repose sur deux éléments : d'une part, la définition d'une distance permettant de comparer les objets à classer, et d'autre part la définition d'un algorithme de classification qui construit la structure arborescente proprement dite. Dans le domaine de la formation de hiérarchies conceptuelles, la notion de distance entre mots a été étudiée en Traitement Automatique de la Langue (TAL) pour former les classes sémantiques qui forment les nœuds de ces hiérarchies, tandis que les algorithmes de classification ont été plus largement étudiés en apprentissage automatique ainsi qu'en analyse des données.

Il existe de nombreux outils destinés à l'acquisition automatique ou semi-automatique de classes sémantiques visant à regrouper de termes «proches». Cette notion de *proximité sémantique* est basée sur des mesures de distance entre termes en fonction du degré de ressemblance des contextes dans lesquels ils apparaissent. Les descriptions des contextes des termes, (formant les exemples d'apprentissage), et les régularités recherchées diffèrent selon les approches. Ainsi, les contextes peuvent être purement «graphiques» (co-occurrences dans une fenêtre de mots)

tels que dans les travaux décrits par [Sparck Jones et Barber, 71], [Church & Hanks, 89], [Brown et al., 92], cette fenêtre pouvant s'étendre au document entier (voir par exemple [Qiu et Frei, 93]), comme ils peuvent être syntaxiques. Dans ce cadre, citons parmi d'autres, [Hirshman L. et al., 75], [Grishman et al., 86], [Grishman et Sterling, 94], [Hindle, 90], [Grefenstette, 92], [Pereira et al., 93], [Dagan et al., 94], [Dagan et al., 96], [Sekine et al., 92], [Hogehout et Matsumoto, 97], [Bouaud et al., 97] qui ont permis l'élaboration de notre modèle.

Cependant, le choix de la distance appropriée pour un corpus et une tâche donnée est un problème ouvert et peu étudié comme souligné par [Roland D. et Jurafsky D., 98], [Habert, 99]. Dans la majorité des travaux, les auteurs ne s'intéressent qu'à la seule évaluation d'une seule tâche finale, celle la même pour lequel l'apprentissage est réalisé ; c'est le cas dans [Grefenstette, 93 et 94]. De plus, les critères d'évaluation proposés «à la sortie» de l'apprentissage restent purement quantitatifs et donnent rarement lieu à des études comparatives [Dagan et al., 96]. Pourtant, la caractérisation des effets des méthodes donnerait les moyens méthodologiques de choisir la plus appropriée ou d'en mettre au point de nouvelles. Cette observation vaut également pour les algorithmes de classification. Il n'existe pas de méthodologie, ni d'outils, permettant de mettre au point un algorithme de classification conceptuelle pour former des ontologies en fonction de la tâche. Par ailleurs, les travaux effectués en classification conceptuelle, citons parmi d'autres [Michalski et al. 83], [Fisher 87], [Gennari et al. 90], [Bisson 92], [Vasco et al. 96], [Talavera et al. 99], n'ont pas trouvé d'échos en apprentissage à partir de corpus. Il faut cependant reconnaître que l'exploitation de ces travaux dans le domaine du TAL exige une première phase d'adaptation des algorithmes existants. Il en est de même des travaux plus anciens effectués en analyse des données autour de la classification ascendante hiérarchique. De même que pour les distances, la mise au point et le choix de l'algorithme approprié à un corpus et une tâche nécessitent le développement de nouveaux moyens méthodologiques et d'outils.

Pour aider à la comparaison, à l'évaluation et à la mise au point de méthodes de formation de hiérarchies conceptuelles ainsi qu'à la diffusion et à l'application de résultats de classification conceptuelle, nous proposons un atelier configurable appelé Mo'K.

Le modèle de classification qui sous-tend cet outil permet de décrire finement les composants d'une distance et les opérateurs de construction de classes, de telle sorte que l'instanciation en des méthodes existantes et la mise au point de nouvelles méthodes soient simples pour le concepteur. L'ensemble du modèle a été élaboré à partir d'une synthèse de travaux en TAL sur la formation de classes sémantiques à partir de corpus, à partir de travaux en apprentissage sur la classification conceptuelle et de ceux issus de l'analyse des données. Ce modèle est complété par un ensemble de variables spécifiques à l'apprentissage à partir de corpus qui permettent de spécifier certaines caractéristiques de constitution du corpus d'apprentissage comme l'élagage, les «stop-lists», etc. Enfin, des critères d'évaluation des résultats d'apprentissage obtenus pour une configuration de paramètres donnés sont également proposés ; ils sont basés sur une tâche

de prédiction d'attachement syntaxique. Ainsi, outre le modèle et l'atelier Mo'K, cet article présente des résultats expérimentaux sur deux corpus très différents qui illustrent la pertinence de notre approche pour répondre aux objectifs que nous nous sommes fixés. Ces résultats concernant uniquement la formation de classes, l'algorithme de classification n'est pas détaillé dans cet article.

2. Le cadre

2.1 Apprendre des classes sémantiques à partir de contextes syntaxiques

Dans le cadre de l'apprentissage de classes sémantiques, l'apprentissage à partir de contextes syntaxiques présente l'intérêt d'exploiter les relations syntaxiques entre mots pour en déduire des relations sémantiques, dans le cadre de la sémantique distributionnelle et suivant l'hypothèse harissienne [Harris et al., 89] selon laquelle l'étude des régularités syntaxiques dans un corpus de sous-langage (ou langage spécialisé) permet d'identifier des schémas syntaxiques formés de combinaisons de classes de mots qui reflètent les connaissances du domaine traité. Le choix de traiter des corpus spécialisés simplifie la tâche d'apprentissage en limitant le vocabulaire, la polysémie et la variabilité des formes syntaxiques. C'est donc dans ce cadre que nous étudions l'application du modèle d'apprentissage proposé.

Les résultats de l'apprentissage dans les approches syntaxiques citées ci-dessus sont, selon les méthodes, des distances entre termes qui représentent leur degré de similitude (dans un cadre d'«instance based-learning»), [Hirshman et al., 75], [Grishman et al., 86], [Grishman et Sterling, 94], [Sekine et al., 92], [Dagan et al., 94], [Dagan et al., 96], [Resnik, 95], des classes de termes basées sur ces distances à l'aide de méthodes de plus proche voisin [Grefenstette, 92], [Hindle, 90], des degrés d'appartenance à des classes de termes [Ribas, 94], des hiérarchies de classes [Pereira et al., 93], [Hogehout et al., 97], [Bouaud et al., 1997] ou des cadres de sous-catégorisation ou des schémas prédictifs dont les restrictions de sélections sont des concepts [Basili et al., 96], [Thompson, 95], [Gomez, 98], [Faure & Nédellec, 98].

Dans tous les cas, la notion de distance est centrale. Elle permet de quantifier le degré de proximité de deux objets, ici des mots (verbes, termes, adjectifs, etc.)¹, en fonction du degré de ressemblance des contextes syntaxiques dans lesquels ils apparaissent. Les classes formées par l'agrégation d'objets proches selon la distance sont ensuite utilisées dans différents buts selon les travaux, tels que la désambiguïsation syntaxique, [Resnik et Hearst, 93], [Resnik, 94], [Ribas, 94], [Ribas, 95] ou la recherche documentaire, [Grefenstette, 92]. Les moyens de calculer ces distances sont pourtant les mêmes et notre modèle s'appuie donc sur cet ensemble des travaux indépendamment de la tâche finale.

¹ Par la suite, à des fins de simplification, nous n'utiliserons que le mot terme bien qu'il soit très restrictif.

2.2 Classification conceptuelle

Dans le cadre de l'atelier Mo'K, l'ontologie prend la forme de hiérarchies multiples formant un graphe acyclique dont les nœuds sont des catégories de termes, décrites en extension, et la relation du graphe est la relation d'inclusion, vue comme une relation de généralité. L'apprentissage par classification hiérarchique d'un ensemble d'objets peut s'opérer essentiellement de deux manières : descendante, par spécialisation progressive des classes, ou ascendante, par généralisations successives. Nous avons fait le choix d'une approche ascendante à la fois pour des raisons de complexité algorithmique et d'explicabilité pour la validation par l'utilisateur, en cours d'apprentissage (voir [Faure et Nédellec, 99] pour plus de détails sur la justification). Nous nous concentrerons dans cet article sur les moyens qui permettent d'obtenir et d'évaluer *les classes de base* de ce graphe, à savoir les critères de constitution du corpus d'entrée, les distances, et les critères d'évaluation des résultats. Nous ne détaillerons donc pas ici pour des raisons de place, l'algorithme générique de construction des classes. On peut souligner cependant que l'application à notre problème, des algorithmes de classification hiérarchiques, soit conceptuelle [Michalski & Stepp, 83], soit numérique [Day & Edelsbrunner, 84], n'est pas immédiate parce qu'il s'agit ici de construire des graphes acycliques et non des hiérarchies strictes, avec un faible nombre de niveaux d'abstraction.

3. Présentation du modèle et de l'outil

3.1 Les exemples et les résultats d'apprentissage

Les contextes syntaxiques pris en compte dans les travaux existants, c'est aussi le cas dans notre modèle, sont des relations grammaticales binaires. Les exemples sont donc représentés par des triplets, Tête – Relation Grammaticale – Expansion tels que Expansion représente l'objet à classer et Tête – Relation Grammaticale représente l'attribut. Le nombre d'occurrences du triplet dans le corpus caractérise l'attribut pour un exemple. Par exemple, si nous nous intéressons qu'aux relations verbales, les deux fragments suivants de phrases analysées,

- Cela entraîne (COD) une diminution de [...].
- Ce taux élevé résulte de (CC) l'augmentation de [...].

permettent de produire les deux triplets suivant : <entraîner_COD, diminution, 29> et <Résulter_de, augmentation, 2>. Ils sont de la forme <Verbe - Relation grammaticale, Nom, nombre total d'occurrences du triplet dans le corpus> où Nom est la tête du groupe nominal, le nom principal. Dans la suite, nous appellerons «Action», le couple Verbe – Relation_grammaticale. Les actions Entraîner_COD et Résulter_de(CC) peuvent être vues comme les objets et les noms diminution et augmentation comme des attributs de valeurs 29 et 2 respectivement. En classification ascendante, l'algorithme général consiste à réunir successivement par paires, les objets proches, ou les objets et les classes objets proches, afin de former des hiérarchies ou des graphes de classes d'objets. Dans la pratique habituelle en

TAL, les classes exploitées ensuite sont les classes d'objets, c'est-à-dire la représentation des classes en intension (Figure 1.a). Dans la partie expérimentale de l'article, nous *n'utiliserons pas* cette stratégie pour comparer les classes apprises, c'est la représentation en *extension* qui va nous intéresser : ce seront les classes formées par la réunion des attributs des objets proches qui formeront les classes recherchées (Figure 1.b)

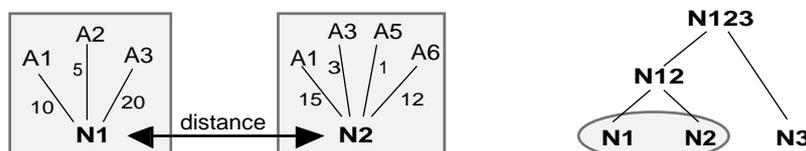


Figure 1.a. Classes d'objets (= noms).

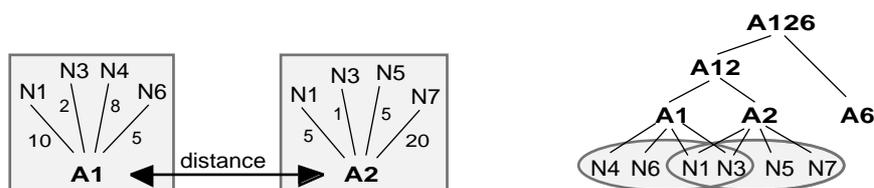


Figure 1.b. Classes d'attributs (= noms).

Prenons un exemple, si les deux objets `Entraîner_COD` et `Résulter_de (CC)` sont sélectionnés pour former une classe, les deux ensembles de leurs attributs seront réunis. Par exemple, `Entraîner_COD` est décrit par l'ensemble d'attributs `{diminution, augmentation, modification, perte, etc.}` et `Résulter_de(CC)` est décrit par `{diminution, augmentation, composition, évolution, etc.}`. La classe de noms apprise comprendra les noms en commun (en gras), mais aussi la réunion des complémentaires (en italique), ainsi 4 nouveaux triplets auront été «induits». Dans la suite, nous utiliserons la stratégie «classes d'attributs» car ainsi les classes «feuilles» que nous évaluons dans la suite sont plus larges. Pour des raisons de place, nous ne développerons pas ici plus avant les différences entre ces deux points de vue, intension et extension, et leur intérêt dans le domaine de la formation d'ontologie.

3.2 Paramètres de constitution du corpus d'apprentissage

Les paramètres de constitution du corpus d'apprentissage dans Mo'K comprennent la sélection des relations grammaticales à étudier, le niveau d'élagage, le «nettoyage» du corpus ainsi que la détermination du nombre d'exemples : la proportion du corpus de départ à utiliser. Nous allons présenter brièvement ces différents paramètres.

- **Sélection des relations grammaticales**

Un des objectifs que nous attachons à notre modèle est de permettre à l'utilisateur de comparer les résultats d'apprentissage en fonction des relations grammaticales sélectionnées en entrée. Les objets à classer et les contextes syntaxiques pris en compte diffèrent totalement selon les approches. Ainsi dans [Hindle, 90], Hindle propose de calculer des similarités entre deux noms

sur la base de leurs contextes verbaux communs. Sont extraits du corpus d'apprentissage les verbes et les têtes (noms principaux) des sujets et des compléments d'objets directs associés. La proposition décrite dans [Grishman et Sterling, 94] consiste à extraire du corpus d'apprentissage tous les triplets tête - relation syntaxique - tête du complément, quelle que soit la tête (verbe ou nom) et la relation syntaxique. Les relations syntaxiques considérées dans [Grefenstette, 92] entre têtes et arguments sont les relations nom-adjectif, nom-groupe prépositionnel, verbe-sujet, verbe-objet direct et verbe-objet indirect (plus la relation nom-nom, pertinente pour l'anglais). Par contre, aucun présupposé n'est fait dans [Dagan et al., 96] sur le type de contexte et d'objets. Enfin, la méthode Asium proposée par [Faure et Nédellec, 98] utilise tous les attachements verbaux, arguments et ajouts et en conserve des triplets (verbe - préposition et relation grammaticale - tête).

Aucun de ces travaux ne mentionne d'étude comparative des résultats en fonction des relations grammaticales sélectionnées dans le corpus d'entrée. Notre modèle permet aisément de les spécifier. Les expérimentations décrites ci-dessous paragraphe 4.1 sur les relations verbales illustre ce point, et montrent, par exemple, des différences significatives entre les résultats selon que les objets et les attributs sont des verbes ou des noms et selon les corpus.

- **Élagage du corpus**

Un deuxième paramètre, mentionné dans la majorité des travaux, et inclus dans notre modèle concerne l'élagage du corpus en fonction du nombre d'occurrences. Le rôle de l'élagage est de supprimer les occurrences trop rares qui représenteraient «a priori» du bruit, au risque de supprimer des cas rares mais néanmoins importants, et de supprimer les occurrences trop fréquentes qui n'apportent aucune information sur le lien privilégié entre un objet et un attribut. Notre modèle permet de spécifier également le nombre minimum d'exemples que caractérise un attribut et le nombre minimal d'attributs pour un exemple ainsi que, pour chacun de ces deux paramètres, le nombre minimal total d'occurrences concernées dans le corpus. Les expérimentations décrites ci-dessous au paragraphe 4.2 confirment que le niveau d'élagage a un effet majeur sur les résultats d'apprentissage et montrent que le niveau optimal diffère notablement selon les corpus, justifiant ainsi notre proposition de l'étudier comme un paramètre de notre modèle.

- **Gestion des «Stop-lists»**

Les travaux existants font référence au «nettoyage» du corpus avant apprentissage incluant la suppression des mots vides, la réorthographe pour des raisons de standardisation et la sélection de certaines formes parmi celles qui apparaissent dans le corpus. Pour répondre à ce besoin, notre modèle prend en entrée des listes de sélection, de suppression et de remplacement.

- **Taille du corpus**

La taille de la portion du corpus utilisée pour l'apprentissage affecte considérablement les résultats de l'apprentissage. Par ailleurs, le volume utilisable est souvent limité par les moyens

de calcul. Notre modèle inclut donc une fonction de partition du corpus en ensembles d'apprentissage et de test.

3.2 Modélisation de la distance

Le but de notre modèle n'est pas de rendre compte de toutes les méthodes possibles qui peuvent être utilisées pour mesurer la similarité entre exemples. Au contraire, notre approche se concentre sur les méthodes qui présentent les caractéristiques suivantes :

- Elles prennent en entrée un texte analysé syntaxiquement,
- Elles n'utilisent pas de ressources externes telles que WordNet, [Miller, 90],
- Elles sont basées sur la comparaison des distributions des attributs décrivant les couples d'objets à classer.

De nombreuses méthodes ont été proposées en TAL qui respectent ce cadre, parmi d'autres mentionnons, [Hindle, 90], [Grefenstette, 92] [Grishman et Sterling, 94], [Dagan 94] et [Faure et Nédellec, 98]. Nous avons développé un modèle générique de ces méthodes qui a été implémenté dans Mo'K, dans le but de proposer une méthodologie d'évaluation et de comparaison dans le cadre de ce modèle. Pour obtenir cette implémentation générique, nous avons identifié les étapes partagées par ces méthodes, comme présenté ci-dessous. Plutôt qu'une librairie de méthodes, Mo'K est plutôt un atelier qui implémente un ensemble de méthodes génériques instanciables dans une représentation orienté objet.

Dans notre modèle, la similarité est basée sur la comparaison de profils distributionnels. En d'autres termes, deux objets seront considérés comme proches si la fréquence relative de chacun de leurs attributs (c'est-à-dire des contextes syntaxiques) sont proches. Les exemples d'apprentissage peuvent donc être représentés par des tableaux de contingence dont les exemples sont les lignes et les attributs les colonnes. Rappelons que ces exemples peuvent représenter différentes choses selon l'objectif d'apprentissage. Par exemple, dans la section 4.2, ils représentent des actions ou des noms selon l'expérimentation. Les cases du tableau représentent le nombre d'occurrences de l'attribut pour l'exemple. Ce tableau est habituellement très creux puisque les exemples sont représentés par un petit nombre d'attributs très variables d'un exemple à l'autre (Figure 2).

En pratique, le calcul de similarité peut être décomposé en deux étapes principales, la pondération et la comparaison des profils.

• La phase de pondération

Elle a pour objectif de remplacer chacune des valeurs brutes de co-occurrences du tableau de contingence par un coefficient, souvent normalisé, que l'on peut interpréter comme une pondération de l'importance de la co-occurrence des exemples et des attributs dans le corpus. Son calcul comprend le plus souvent deux étapes, une initialisation de la pondération des exemples et des attributs, le plus souvent par leur nombre d'occurrences, puis le calcul d'une pondération normalisée de l'importance de chaque attribut pour chaque exemple. Par exemple,

dans [Hindle, 90], chaque paire Nom-Action (NA) du tableau de contingence est pondérée par la mesure de l'information mutuelle qui est calculée à partir du ratio entre le nombre de co-occurrences de NA dans le corpus divisé par le produit du nombre d'occurrences des Noms et des Actions dans le corpus. D'un point de vue technique, cette phase de pondération est implémentée au moyen des 5 fonctions suivantes, décrites ici en pseudo code :

<i>Intitulé de l'étape</i>	<i>Méthodes</i>
Initialisation de la pondération de chaque exemple E Initialisation de la pondération de chaque attribut A Pour chaque exemple E Pour chaque attribut A de l'exemple Calcul de la pondération de A dans le contexte de E Mise à jour de la pondération globale de E Pour chaque attribut A de l'exemple Normalisation de la pondération de A par le poids de E	Init_Weight_Example Init_Weight_Attribute Eval_Weight_Example Eval_Weight_Attribute Init_Similarity

- La phase de comparaison des profils

Elle construit la matrice de similitude entre les couples d'exemples. Il s'agit de comparer l'ensemble des coefficients attribués aux attributs à l'étape précédente, pour chaque paire d'exemples. La ressemblance est évidemment fonction croissante du nombre d'attributs communs, mais les façons de mesurer la ressemblance entre ces distributions diffèrent selon les approches. Par exemple, la mesure proposée par Dagan, [Dagan et al., 94] utilise la divergence de Kulback-Leibler qui est basée sur le rapport entre les coefficients des attributs communs. Dans Mo'K, la décomposition qui est faite de cette étape est minimale :

<i>Intitulé de l'étape</i>	<i>Méthode</i>
Pour chaque paire d'exemple E1 et E2 Calcul de la similarité entre E1 et E2	Eval_Similarity

En pratique on pourrait facilement décomposer l'unique fonction qui est utilisée ici en plusieurs sous-unités. Nous ne l'avons pas fait car le code est généralement de petite taille (~20 lignes de code) mais assez divers. Sous une forme atomique, il améliore la généricité de l'approche tout en donnant la possibilité d'effectuer des implémentations plus spécialisées et donc plus efficaces du calcul. Ce second point est à considérer si l'on songe que le nombre d'exécutions de cette étape est en $O(\text{nombre d'exemples})^3$ pour un processus de classification conceptuelle.

Ainsi, par le biais de ces 6 fonctions, il est possible d'implémenter, en quelques dizaines de lignes de code, la plupart des mesures de ressemblances qui suivent un schéma basé sur la comparaison de paires de profils distributionnels. Ajoutons que nous ne faisons pas ici d'hypothèse plus spécifique sur les propriétés que la mesure doit vérifier : il peut s'agir d'une similarité ou d'une dissimilarité, elle peut être symétrique ou non (c'est ainsi le cas de celle décrite dans [Grishman et Sterling., 94]) et les informations calculées peuvent être d'un type quelconque. Cette approche facilite donc grandement la comparaison des méthodes existantes mais aussi la mise au point de variantes de méthodes existantes ou, plus radicalement, de

méthodes totalement originales. Une fois intégrée à Mo'K, une méthode accède à l'ensemble des services de tests et de classification conceptuelle du système.

3.3 Évaluation de la distance

Dans la construction de hiérarchies de concepts, la notion de distance est centrale, c'est la raison pour laquelle nous avons choisi de l'étudier indépendamment de l'autre paramètre central, la stratégie de classification. Notre étude porte sur l'évaluation de la pertinence de la distance pour la formation des classes sémantiques et l'analyse de son comportement en fonction du domaine d'application et des paramètres de constitution du corpus d'apprentissage. Des moyens d'évaluation sont ainsi proposés dans Mo'K basés sur les N meilleures classes construites par agrégation binaire, soit en d'autres termes, sur les N premiers couples d'exemples obtenant les plus forts scores dans la matrice de similitude.

- **Mesure du taux de rappel**

Comme il a été décrit au début du chapitre 3 la constitution d'une classe conduit à l'induction de nouveaux triplets non observés dans le corpus initial. Dès lors, la démarche d'évaluation va être des plus classiques : elle consiste à partager le corpus en deux parties, un corpus d'apprentissage et un corpus de test. Le premier sert à construire la matrice de ressemblance basée sur la mesure à évaluer, puis le second permet de mesurer le taux de couverture, ou taux de rappel des N meilleures classes, c'est-à-dire leur aptitude à reconnaître les triplets de l'ensemble test. Nous avons sélectionné cette tâche d'évaluation pour deux raisons. D'une part, elle correspond peu ou prou à l'étape élémentaire de tout processus de classification ascendante hiérarchique : l'agrégation des concepts élémentaires. D'autre part, du point de vue du TAL, elle s'apparente à une tâche de désambiguïsation.

- **Mesure du taux de précision**

La précision, c'est-à-dire l'aptitude de la métrique à ne pas reconnaître à tort un exemple négatif, est également une caractéristique importante : à la limite une mesure de similarité qui aurait tendance à regrouper en priorité et de façon «gloutonne», les couples d'objets décrits par un grand nombre d'attributs obtiendrait des taux de couvertures remarquables, sans que les classes produites aient la moindre sémantique. Ce problème est difficile à traiter de façon automatique en apprentissage non supervisé, en l'absence d'exemples négatifs. Dans Mo'K, nous avons envisagé deux mesures décrites ci-dessous (section 4). L'une mesure «la rentabilité» comme le pourcentage de triplets induits qui est effectivement utilisés dans la couverture du corpus de test, l'autre est basée sur un corpus d'exemples artificiels générés automatiquement, partant de l'hypothèse qu'ils seront presque tous négatifs, suivant en cela [Dagan et al, 94]. Les exemples sont construits en sélectionnant aléatoirement pour chacun d'eux un objet et un attribut dans le corpus initial tout en prenant soin que les exemples ainsi formés n'appartiennent pas à l'ensemble d'apprentissage. Nous mesurons le taux de couvertures de ces exemples artificiels par les classes apprises. Plus les exemples artificiels

couverts, réputés négatifs sont nombreux, moins la distance est bonne. Par ailleurs, comme nous le verrons dans la prochaine partie qui présente des expérimentations, d'autres critères sont intéressants à mesurer pour juger de la pertinence d'une mesure de similarité, tel par exemple le «taux d'induction» qui mesure le rapport entre les nombres de triplets appris par cœur et les nombres de triplets induits dans les N classes évaluées.

4. Expérimentations et résultats

Les expérimentations décrites ci-dessous ont pour but d'illustrer les possibilités de paramétrage de Mo'K et l'impact des différentes configurations sur les résultats d'apprentissage, justifiant ainsi l'importance d'un travail exploratoire systématique avec de telles plates-formes dès lors que l'on veut obtenir des résultats pertinents sur un domaine, c'est-à-dire un corpus, donné.

4.1 Les corpus

Les expérimentations rapportées ci-dessous portent sur deux corpus de textes en français de natures différentes. L'un est un corpus de recettes de cuisine, rassemblé sur le Web sur de nombreux sites. L'autre est Agrovoc, un corpus de résumés scientifiques en agriculture, rassemblé par l'INIST. Ces deux corpus ont été analysés par le même analyseur syntaxique de surface (Shallow Parser). Seules les relations verbales ont été conservées pour nos fins expérimentales de la forme <Verbe – relation grammaticale – Nom>. Il convient de souligner que les résultats de ces analyses syntaxiques sont considérablement bruités (30 et 50 % d'erreurs selon le corpus), ceci pour des raisons différentes essentiellement dues au nombre considérable de fautes de grammaire, d'orthographe, de frappe et d'accentuation dans les recettes de cuisine et à l'abondance de termes techniques pris comme des verbes et de compléments de noms emboîtés attachés par erreur au verbe, dans Agrovoc.

Notre expérience avec Agrovoc, concerne un sous-ensemble du corpus contenant 300 verbes pour 18 828 noms, seuls sont conservés les triplets dont le verbe apparaît dans une liste de verbes sélectionnés comme donnant lieu à des nominalisations dans le corpus. Dans le corpus de recettes de cuisine, on trouve 1181 verbes pour 3300 noms, soit un nombre moyen de noms par verbe 20 moins élevé que dans Agrovoc, ce qui reflète la plus grande spécialisation du corpus de cuisine. Par ailleurs, le corpus Agrovoc est quatre fois plus gros que celui des recettes de cuisine (117 156 triplets pour un nombre total d'occurrences de 168 287).

4.2 Qui doit-on classer ?

La première expérience illustre l'importance du choix de l'objet à classer et de l'attribut qui le caractérise. En effet, comme nous l'avons déjà évoqué, du point de vue l'apprentissage, nos exemples sont «symétriques» et l'on peut intervertir lignes et colonnes de la matrice de contingence sans modifier la «nature» du problème d'apprentissage (les hiérarchies produites portent évidemment sur des objets différents). Nous comparons donc ici les résultats obtenus en

considérant les actions comme les objets et les noms comme des attributs, avec ceux obtenus en considérant l'hypothèse inverse : les noms sont les objets et les actions les attributs. La comparaison porte sur les deux corpus, les autres paramètres étant fixes. La distance utilisée ici est celle d'Asium [Faure et Nédellec, 98], l'élagage est léger : pas de nombre minimum d'occurrences pour un triplet, le nombre minimum d'exemples que caractérise un attribut est 2 et le nombre minimal d'attributs pour un exemple est 3. Nous reviendrons sur ce choix au paragraphe suivant. Pour chaque corpus, la couverture des triplets appris et appartenant aux 25 premières classes des objets les plus proches est évaluée sur le corpus de test qui représente 20 % du corpus total et sur un corpus artificiel de 50 000 triplets généré aléatoirement (cf. 3.3). Chacun des tests est effectué 4 fois successivement.

<i>Corpus</i>	<i>Objet classé</i>	<i>%Triplets induits dans les classes</i>	<i>%Taux de rappel</i>	<i>%Taux de faux négatifs</i>
Agrovoc	Action	40 %	4.7 %	45 %
	Nom	38 %	5.3 %	45 %
Cook	Action	34 %	12 %	32 %
	Nom	38 %	9.1 %	52 %

Sur Agrovoc, bien que les classes créées par le classement par action soient environ deux fois plus grandes (en nombre de triplets décrits) que par le classement par nom, les taux d'induction (nombre de triplets induits par rapport au nombre de triplets appris par cœur) sont très comparables (40 % contre 38 %). Le rappel, évalué ici au travers du taux de couverture des seuls triplets induits (sans les triplets appris par cœur) sur le corpus de test, est légèrement meilleur dans le cas du classement par nom (5.3 % contre 4.7 %).

Les taux de précision mesurés à partir de la couverture sur le corpus artificiel et tels que décrits ci-dessus ne donnent pas l'information escompté. En fait, les exemples étant générés aléatoirement, une petite quantité d'exemples positifs est nécessairement générée par la procédure (de l'ordre de 0.5 % sur les corpus que nous avons étudiés). Ce taux est très faible mais malheureusement similaire au taux de couverture qui est de l'ordre de 0.5 % à 2.5 % dans nos expérimentations. Nous avons donc examiné «à la main» ces exemples couverts et constaté que les exemples positifs constituent malheureusement une partie non négligeable des exemples artificiels effectivement couverts par les classes apprises : de l'ordre de la moitié des exemples couverts sont positifs ! Notre conclusion est que la précision ne peut pas être évaluée de façon automatique et fiable à partir des exemples artificiels. La seule alternative est de compter manuellement le nombre effectif de négatifs dans les exemples couverts. Dans les deux expériences décrites ci-dessus, les taux de précision calculés à partir de la couverture sur un corpus artificiel sont similaires : de l'ordre de 45 % des exemples couverts sont effectivement négatifs. Les classes apprises dans les deux représentations présentent donc un rappel et une précision très comparables, le rappel étant marginalement meilleur pour les noms, mais la précision reste très moyenne.

Ce mauvais résultat peut s'expliquer de la manière suivante. L'examen des meilleures paires de noms montre en effet que ces noms sont très généraux par exemple, *technique-méthode*, *influence-effet*, par rapport aux nombreux termes très techniques du corpus. La majorité des actions décrivant ces noms correspondent également à des verbes très généraux "fourre-tout" comme les verbes *présenter*, *observer* ou *rapporter*. Cette observation se vérifie également sur les meilleures paires d'actions (*Réaliser_chez* - *rapporter_chez*, *étudier_cod* - *analyser_cod*, etc.), ce qui explique que les classes apprises ne soient pas d'une grande fiabilité quel que soit le mode de représentation utilisé. Le biais de choix des verbes dans le sous-corpus suffirait à expliquer ce phénomène qui ne se reproduit pas dans le corpus de recette où les verbes généraux n'ont pas la faveur.

Dans le corpus de la cuisine, les classes sont de tailles similaires pour les deux représentations. Le taux d'induction est légèrement plus important pour le classement par nom, 38 % contre 34 %, mais par contre les résultats de couverture sur le corpus de test montrent pourtant un meilleur rappel dans le cas du classement par action, 12 % contre 9.1 %. Les triplets induits sont donc plus utiles dans le cas des actions. De plus, la précision mesurée par le pourcentage de triplets artificiels incorrects et couverts est significativement meilleure pour le classement par action 32 %, contre 52 %. La précision et le rappel sont donc meilleurs dans le cas du classement par action qui pourtant induit moins de nouveaux triplets que le classement par nom et ces deux critères sont notablement meilleurs qu'avec Agrovoc.

L'examen direct des meilleurs couples de noms et d'actions confirme que la surgénéralisation est moins à craindre que pour Agrovoc. Les couples de noms sont plus précis (*réfrigérateur* - *congélateur*, *huile_d'olive* - *huile*, *lamelle* - *dés*, etc.) et décrits par des actions plus techniques, de même les meilleurs couples d'actions (*reposer_au*-*conserver_au*, *piquer_avec* - *percer_avec*, *replier_en* - *plier_en*) sont décrits par des noms incomparablement plus spécifiques que dans Agrovoc. La moins grande variabilité du corpus de cuisine explique ce phénomène, montrant que la plus grande taille d'Agrovoc n'augmente pas signification des régularités observées.

On voit donc dans cette expérience que le choix de représentation peut avoir un impact important sur les résultats et qu'il est donc souhaitable avant de s'attaquer à un nouveau domaine de savoir quel type de représentation est le plus pertinent.

4.3 Les paramètres d'élagage

L'expérimentation rapportée ici illustre l'importance de l'élagage. Deux configurations ont été comparées sur les deux corpus. La distance utilisée est toujours celle d'Asium, les objets sont les actions et les attributs sont les noms. La première configuration est celle qui a été décrite dans le paragraphe précédent. Dans la deuxième configuration, le nombre minimum d'occurrences pour un exemple (triplet) est égal à 2, ce qui élimine les triplets rencontrés à une seule reprise dans le corpus dont on peut penser qu'ils représentent du bruit. Le nombre minimum d'exemples que doit caractériser un attribut est légèrement augmenté, 3 au lieu de 2.

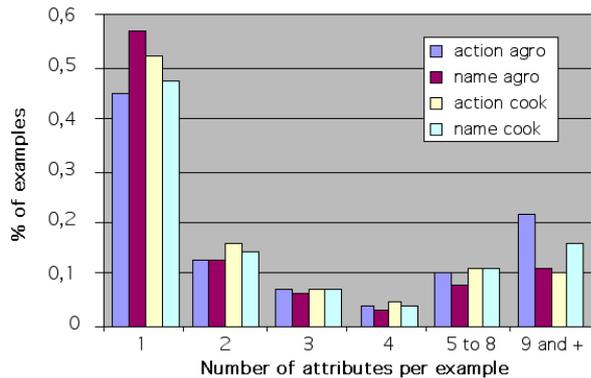


Figure 2 : Répartition des exemples en fonction du nombre d'attributs les décrivant.

Le nombre minimal d'attributs pour un objet est élevé à 5 au lieu de 3, c'est-à-dire que les objets comparés apparaissent dans au moins 5 contextes syntaxiques différents. On voit sur l'histogramme (Figure 2), que ce seuil élimine environ 80 % des deux corpus, quand le seuil de la première configuration élimine moins de 70 % de ces corpus. On peut espérer ainsi un classement plus sûr, mais au risque d'avoir éliminé tant d'exemples que la couverture en soit drastiquement atteinte.

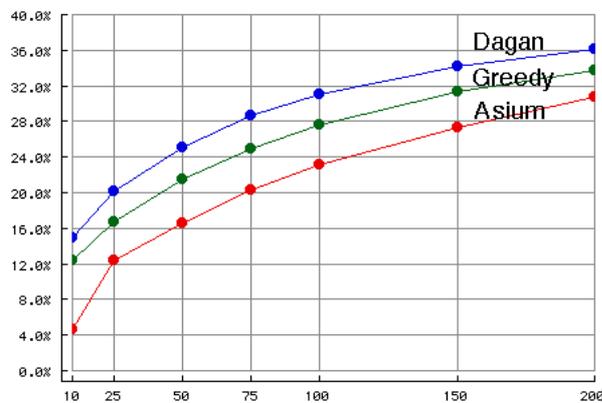
En fait, sur le corpus de recettes, les expériences, que nous ne détaillerons pas ici, montrent que le taux d'induction (32 %) et la couverture (11 %) sur l'ensemble de test ne sont quasiment pas atteints. Par contre le taux de triplets artificiels couverts est multiplié par 3, ce qui est considérable et laisse à penser que le taux d'exemples induits et incorrects augmente à ce taux d'élagage. Avec Agrovoc, le taux d'induction tombe à un tiers du taux précédent, le taux de couverture sur l'ensemble de test tombe de moitié et le taux de triplets artificiels couverts est multiplié par plus de deux. Le rappel est donc plus considérablement affecté par l'élagage que dans le cas des recettes de cuisine. Dans les deux cas, ce nouvel élagage diminue les performances par rapport au premier. Toutefois, cette conclusion basée sur l'évaluation des classes de base doit être nuancée dans le cas de la formation de hiérarchies : l'élagage plus poussé permet en effet d'éliminer nombre de classes de base manifestement peu significatives qui sont issues de la présence d'actions assez similaires, mais décrites par un faible nombre d'attributs. Il est clair que dans un processus de classification hiérarchique, ce type de classe poserait des problèmes en faussant certains regroupements. Le type d'élagage à effectuer sur un corpus dépend donc en partie du type de tâche à effectuer.

4.4 Comparaison de méthodes

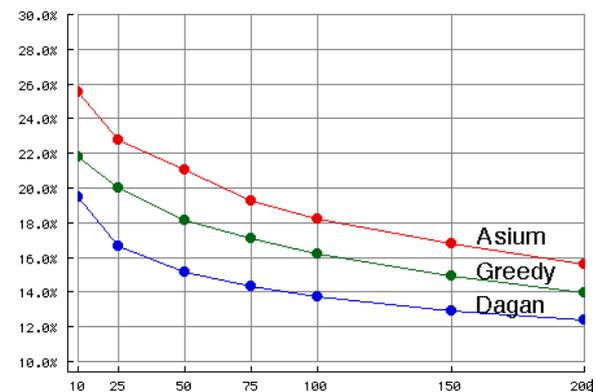
Cette dernière d'expérience illustre quelques aspects de l'utilisation de Mo'K pour comparer les résultats obtenus avec différentes distances. Les distances comparées ici sont celles du système Asium [Faure et Nédellec, 99], de [Dagan et al., 96] et une méthode de référence «Greedy» appliquées au corpus de la cuisine. La méthode Greedy a un comportement simple puisqu'elle est basée sur une mesure de ressemblance inspirée du Khi 2, qui favorise les paires d'exemples décrits par un grand nombre d'attributs, d'où son nom. Nous avons sélectionné ces trois méthodes d'une part, car elles donnent globalement de bons résultats par rapport à d'autres que nous avons testées et d'autre part, elles reposent sur des hypothèses différentes. Elles sont donc une bonne illustration des possibilités de modélisation de distances que nous avons présentées au paragraphe 3.3. Les paramètres d'apprentissage sont les mêmes que ceux décrits au paragraphe 4.1 et les objets sont les actions et les attributs sont les noms, mais nous avons de

plus testé l'influence du nombre de classes disjointes sur lesquelles la couverture est évaluée en faisant varier ce nombre entre 10 et 200. Il apparaît en abscisse des graphiques ci-dessous.

Taux de couverture

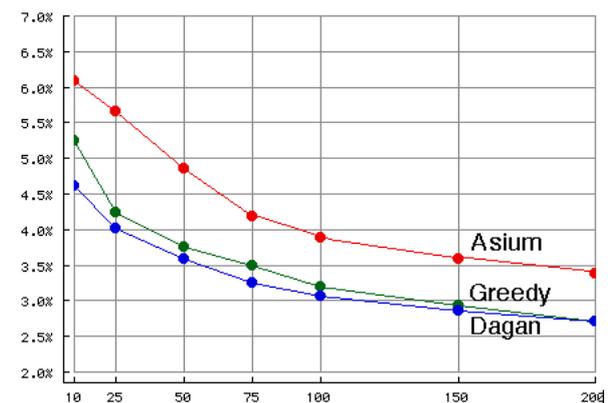


Efficacité des classes apprises



Les deux graphiques présentés ci-dessus montrent respectivement, le taux de couverture de chacune des méthodes sur l'ensemble de test et « l'efficacité des classes apprises ». Cette dernière est mesurée par le rapport entre la somme du nombre de triplets appris par cœur et induits, et le nombre de triplets effectivement utilisés lors du test de couverture. Comme on le voit sur le premier graphique, le taux de couvertures des 3 méthodes croît normalement en fonction du nombre de classes produites, la méthode de Dagan obtenant de meilleurs résultats. Par contre, si l'on analyse l'efficacité des classes sur la base du second graphique, c'est Asium qui exploite le mieux les triplets appris. En combinant ces deux informations, on peut en conclure que ces méthodes ont des comportements différents : l'une, Dagan, conduisant à la construction de classes plus générales (plus de triplets appris dont certains inutiles) et l'autre, Asium, à celle de classes plus spécifiques (moins de triplets appris, mais plus utiles).

On peut affiner ces résultats en étudiant la qualité des inductions effectuées au travers du pourcentage des seuls triplets induits effectivement utilisés lors du test de couverture (figure ci-contre), que nous avons appelé la « rentabilité ». Si la conclusion précédente est confirmée en ce qui concerne Asium et Dagan, on constate que les méthodes de Dagan et Greedy ont, sur ce point, des comportements qui sont identiques ce qui n'était pas le cas précédemment. En fait, il semble que la méthode de Dagan induise des triplets plus intéressants que Greedy, alors que cette dernière a tendance à apprendre par cœur des exemples plus représentatifs du corpus. Les tests sur les exemples artificiels confirment tout à fait ces différents résultats. Ainsi, il semble que les classes apprises par la méthode de Dagan et dans une moindre mesure par Greedy, soient moins robustes et possèdent



une précision moindre pour les paramètres d'élagage et la représentation choisies. On doit souligner que ces enseignements ne sont valables que pour le seul corpus de recettes. Ainsi, avec le corpus Agrovoc, les résultats obtenus sont sensiblement différents, montrant à quel point il est pertinent, lorsqu'on aborde un nouveau domaine, d'avoir une démarche exploratoire préalable visant à déterminer les méthodes et les hypothèses de représentation optimales.

6. Conclusions et perspectives

Avec la multiplication des sources documentaires, les applications issues du Traitement Automatique du Langage deviennent applicables à un contexte de plus en plus large. Nous l'avons vu, de nombreux travaux ont été proposés afin de construire de façon plus ou moins automatique des ontologies de type hiérarchies conceptuelles à partir de corpus analysés. Nombre de tâches du TAL s'appuient, en effet, sur l'existence de telles structures, on peut citer : l'extension de requête, l'extraction d'information, l'indexation de documents, la désambiguïsation syntaxique, ... Malheureusement, à ce jour, les méthodologies et les outils manquent pour mettre au point, évaluer, comparer et mettre en évidence les points faibles et forts des différentes approches pour un corpus et une tâche donnés. Le travail que nous effectuons au travers du système Mo'K va dans ce sens en proposant un atelier configurable de développement de méthodes de classification conceptuelles pour la construction d'ontologies. Le modèle d'apprentissage que nous utilisons est applicable sur des corpus analysés non annotés sans qu'il soit nécessaire de disposer de connaissances additionnelles (terminologique ou sémantique) pour guider l'apprentissage ou évaluer les résultats.

Les premières expérimentations, effectuées dans des domaines scientifiques ou techniques spécifiques, indiquent que la qualité de l'apprentissage effectué est, de façon nette, inversement fonction du degré de généralité du corpus utilisé. Dans ce cadre, il est peu réaliste de vouloir améliorer l'apprentissage en le guidant à l'aide d'ontologies générales, car elles-ci semblent être à la fois trop incomplètes et trop imprécises (notamment du fait de la polysémie) pour apporter une aide efficace. Cette observation est confirmée par, Krovetz et Croft, [91] qui ont montré que 40% des mots, exprimés sous leur forme canonique, qui apparaissent dans les titres ou les résumés des communications de l'ACM sont totalement absents du LDOCE (Longman Dictionary of Contemporary English).

Evidemment, ces conclusions ne concernent que les seuls domaines de spécialité. A contrario, comme le montre amplement la littérature (on peut citer [Yarowsky, 92], [Resnik & Hearst, 93], [Resnik, 94], [Resnick, 95], [Ribas, 94], [Ribas, 95], [Li & Abe, 96]), il est tout à fait souhaitable de s'appuyer sur les ontologies existantes dès lors que l'on souhaite apprendre des classes sémantiques générales. Dans ce contexte, la qualité de l'apprentissage est améliorée en utilisant des ressources additionnelles pour enrichir les données d'entrée, guider l'apprentissage ou confirmer ou en infirmer les résultats. Par exemple, on peut exploiter des terminologies [Grefenstette, 94], des dictionnaires [Krovetz 97] ou des nomenclatures, telles que SNOMED

International dans le domaine médical [Bouaud et al., 1997]. De manière symétrique, il serait également intéressant de pouvoir bénéficier de savoir-faire méthodologique pour intégrer des ontologies spécialisées au sein d'ontologies générales telles que WordNet [Miller, 90].

Bien que nous nous soyons focalisés ici sur les tâches de désambiguation, d'autres méthodes de validation pourraient être intégrées dans Mo'K portant, par exemple, sur des tâches d'extension de requête ou d'extraction d'information. Là encore, cela permettrait de définir, en fonction du corpus étudié, les approches les mieux adaptées pour apprendre les ressources linguistiques nécessaires à la mise en œuvre de ces tâches.

La capacité à apprendre rapidement des ontologies de bonne qualité est un enjeu important notamment pour les applications où, du fait de la rareté de l'expertise et/ou de la rapidité d'évolution du domaine, une acquisition manuelle de ces ressources est trop coûteuse. Dans ce contexte, Mo'K est actuellement appliqué au sein d'un projet en génomique visant à faire de l'extraction d'interaction génique à partir de résumés issus de la base MedLine.

Bibliographie

- Basili R., Pazienza M. T. et Velardi P., "An empirical symbolic approach to natural language processing", in *Artificial Intelligence Journal* 85, pp. 59-99, 1996.
- Bisson G., "Conceptual Clustering in a First Order Logic Representation", in *Proceedings of 10th European Conference on Artificial Intelligence (ECAI'92)*, pp. 458-462, Vienna, 1992.
- Bouaud J., Habert B., Nazarenko A. and Zweigenbaum P., "Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation à deux modélisations conceptuelles", in *Actes des Journées Ingénierie des Connaissances*, Roscoff, Zacklad (Ed.), pp 207-223, May 97.
- Brown P. F., Della Pietra V. J., deSouza P. V., Lai J. C. and Mercer R. L. "Class-based n-gram models of natural language.", in *Computational Linguistic* 18(4), pp.283-298.
- Church K. W. et Hanks P., "Word Association Norms, Mutual Information, and Lexicography", in *proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 76-83, 1989.
- Dagan I., Pereira F., et Lee L., "Similarity-Based Estimation of Word Cooccurrence Probabilities", in *proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, ACL'94*, New Mexico State University, June 1994.
- Dagan I., Lee L., and Pereira F., "Similarity-Based Methods For Word-Sense Disambiguation", in *proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL'96*.
- Day W., Edelsbrunner H., "Efficient Algorithms for Agglomerative Hierarchical Clustering Methods" in *Journal of Classification*. Volume 1. 1-24, 1984.
- Faure D. et Nédellec C., "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition ", in *proceedings of the workshop Adapting lexical and corpus resources to sublanguages and applications of the 1st International Conference on Language resources and Evaluation (LREC)*, pp. 1-8, P. Velardi (Ed.), Granada, Spain, May 1998.
- Fisher D.H. "Knowledge Acquisition via Incremental Conceptual Clustering.", in *Machine Learning Journal* 2, pp. 139-172, 1987. Fisher D., "Model of Incremental Concept Formation." in *Artificial Intelligence Journal*, Volume 40, pp. 11-61, 1989.
- Gomez F., "Linking WordNet Verb Classes to Semantic Interpretation" in *Proceedings of the COLING-ACL Workshop on the Usage of WordNet in NLP Systems*, 1998.
- Grefenstette G., "Use of syntactic Context to Produce Term Association Lists for Text Retrieval", in *proceedings of the 15th International SIGIR'92*, Denmark, 1992.
- Grefenstette G., "Evaluation Techniques for Automatic Semantic Extraction: Comparing Syntactic and Window Based Approaches.", in *Workshop on Acquisition of Lexical Knowledge from Text*, Columbus, OH, June 1993.
- Grefenstette G., *Explorations in Automatic Thesaurus Discovery*, Klüwer Academic (Pub.), 1994.
- Grishman R., Hirshman L., Nhan N. T., "Discovery Procedures for Sublanguages Selectional Patterns: Initial Experiments", in *Computational Linguistics*, Vol 12, (3) pp. 205-215, 1986.
- Grishman R., Sterling J., "Generalizing Automatically Generated Selectional Patterns", in *proceedings of the 16th Int'l Conf. Computational Linguistics (COLING'94)*, 1994.

- Habert B., Nazarenko A. et Salem A., *Les linguistiques de corpus*, Armand Collin, 1997.
- Hindle D., "Noun classification from predicate-argument structure", in *proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 268-275, Pittsburgh, 1990.
- Harris Z., Gottfried M., Ryckman T., Mattick Jr P., Daladier A., Harris T. and Harris S., "The form of Information in Science, Analysis of Immunology Sublanguages", vol. 104 of Boston Studies in the Philosophy of Science. Dordrecht, the Netherlands, Kluwer Academic (Pub.), 1989.
- Hirshman L., Grishman R. et Sager N., "Grammatically-based automatic word class formation", in *Information Processing and Management*, Vol 11, pp. 39-47, Pergamon Press, 1975.
- Hogehout W. R. et Matsumoto Y., "A preliminary Study of Word Clustering Based on Syntactic Behavior", *Proceedings of CNLP, 35th annual meeting of the ACL and EACL'97*, 1997.
- Krovetz R et Croft W.B., W., "Lexical Ambiguity and Information Retrieval", In *Lexical Acquisition: exploiting on-line resources to build a lexicon*, Zernik (Ed.), p 45-65, Hillsdale, New Jersey, Lawrence Erlbaum Associates (Pub.), 1991.
- Krovetz R., "Homonymy and Polysemy in Information Retrieval", In *Proceedings of the 35th annual meeting of the ACL (ACL'97)*, 1997.
- Li H. and Abe N., "Word clustering and desambiguation Based on Co-occurrence Data", in proceedings of Coling - ACL'98, 1998.
- Meersman R. A., "Semantic Ontology Tools in IS Design", in *Foundation of Intelligent Systems, Proceedings of 11th International Symposium, ISMIS'99*, Ras Z. W. et Skowron A. (Eds), Warsaw, Pologne, June 1999.
- G. Miller, "WordNet: an on-line lexical database", in *International Journal of Lexicography*, 3(4), special Issue, 1990.
- Michalski R.S., Stepp E. 1983. "Learning from Observation : Conceptual Clustering". In *Machine Learning I: an Artificial Intelligence Approach*, Tioga, pp. 331-363, 1983.
- Miller G., "WordNet : an On-line lexical database." *International Journal of Lexicography*, 3(4). (special Issue), 1990.
- Pereira F., Tishby N. et Lee L., "Distributional clustering of English words" in *ACL'93*, p. 183-190, 1993.
- Qiu Y. et Frei H. P., "Concept based Query Expansion", in *Proceedings of 16th Annual International ACM SIGIR Conference*, p. 160-169, Pittsburgh, ACM Press, 1993.
- Resnik P., "Using Information Content to evaluate Semantic Similarity in a Taxonomy.", in *Cognitive Modelling*, 1995.
- Resnik P. et Hearst M. A. "Structural Ambiguity and Conceptual Relations", in proceedings of *Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pp. 58-64, Ohio State University, 1993.
- Ribas F., "An experiment on Learning Appropriate Selectional Restrictions from a Parsed Corpus.", proceedings of the *16th Int'l Conf. Computational Linguistics (COLING'94)*, 1994.
- Ribas F., "On Learning More Appropriate Selectional Restrictions.", in *proceedings of EACL'95*.
- Roland D. et Jurafsky D., "How Verb Subcategorization Frequencies Are Affected By Corpus Choice", in proceedings of the *Int'l Conf. Computational Linguistics (COLING'98)*, 1998.
- Schütze H., "Word Space", in *Advance in Neural Information Processing Systems 5*, Hanson S.J., Cowan J.D. et Giles C.L. (Eds), Morgan Kaufmann, 1992.
- Sekine S., Caroll J. J., Ananiadou S. et Tsujii J., "Automatic Learning for Semantic Collocation" in *Proceedings of the third Conference on Applied Natural Language Processing*, pp104-109, 1992
- Sparck Jones K. and Barber E. B., "What makes an automatic keywords classification effective?" in *Journal of the ASIS*, 18, pp. 166-175, 1971.
- Talavera L. and Bejar J., "Efficient construction of comprehensible hierarchical clusterings", in *proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD'98*. pp. 93-101. Nantes, France. J. M. Zytkow and M. Quafafou (eds.) LNAI vol. 1510, Springer Verlag, 1998.
- Thompson C. A., "Acquisition of a Lexicon from Semantic Representations of Sentences", in *33rd Annual Meeting of the Association of Computational Linguistics, (ACL'95)*, pp. 335-337, Boston, M A, July, 1995.
- Vasco J.J.F., Faucher C., Chouraqui E., "A knowledge acquisition tool for multi-perspective concept formation." In proceedings of 9th European Knowledge Acquisition Workshop, EKAW'96, pp. 227-244. Springer Verlag, 1996.
- Yarowsky D., "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods", in.
- Yarowsky D., "Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora", in proceedings of *Colling'92*, p. 454-460, Nantes, 1992.
- Zweigenbaum P., Bouaud J., Habert B. et Nazarenko A. "Coopération apprentissage en corpus et connaissances du domaine pour la construction d'ontologies", in *Actes des Journées Scientifiques et Techniques, Francil, Avignon, Avril 1997*.