Acquisition et structuration des connaissances en corpus : éléments méthodologiques

Chantal Muller, Xavier Polanco, Jean Royauté, Yannick Toussaint

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Acquisition et structuration des connaissances en corpus : éléments méthodologiques

Chantal Muller, Xavier Polanco, Jean Royauté, Yannick Toussaint

N■ 3198

Juin 1997

THÈME 3

ISSN 0249-6399

Acquisition et structuration des connaissances en corpus : éléments méthodologiques

Chantal Muller, Xavier Polanco, Jean Royauté*, Yannick Toussaint†

Thème 3 — Interaction homme-machine,
images, données, connaissances
Projet DIALOGUE‡

Rapport de recherche n° 3198 — Juin 1997 — 45 pages

Résumé : Ce document présente une expérimentation menée dans le domaine de l'agriculture. Les travaux ont été menés dans le cadre du projet ILC sur l'analyse de l'information. L'objectif de cette expérimentation est de montrer comment l'exploitation de modules automatiques de traitement de la langue basés sur la terminologie peuvent être combinés avec des modules de classification pour faire émerger de corpus volumineux de textes, des classes de termes. Ces classes sont interprétables et instancient des modèles abstraits de connaissance du domaine de spécialité que nous avons retrouvés manuellement. Nous avons traité un corpus de 1386 résumés de notices bibliographiques en anglais. La chaîne linguistique opère également sur le français.

Mots-clé : traitement automatique du langage naturel, analyse de l'information, terminologie, classification, réseaux sémantiques, linguistique de corpus, acquisition de connaissance.

(Abstract: pto)

‡ Projet commun au CRIN-CNRS

* Trois premiers auteurs : Équipe PR Infométrie, INIST-CNRS, 54514 Vandœuvre-lès-Nancy Cedex, email : <muller,polanco,royaute>@inist.fr

† INRIA Lorraine, email : Yannick.Toussaint@inria.fr

Unité de recherche INRIA Lorraine
Technopôle de Nancy-Brabois, Campus scientifique,
615 rue de Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY (France)
Téléphone : (33) 83 59 30 30 – Télécopie : (33) 83 27 83 19
Antenne de Metz, technopôle de Metz 2000, 4 rue Marconi, 55070 METZ
Téléphone : (33) 87 20 35 00 – Télécopie : (33) 87 76 39 77

Knowledge Acquisition and Structuration from Corpora

Abstract: This report presents an experiment on the field of agriculture. These results come within the ILC Project on information analysis. Automatic modules for natural language processing are based on terminology and combined with clustering modules. They extract terms from the large corpora and classify them following coocurrence links. We show that experts can explain the belonging of a term to a cluster and we deduced the abstract model related to this knowledge. The corpus was composed of 1386 english abstracts of bibliographical references. The linguistic chain is also able to compute french texts.

Key-words: Natural Language Processing, Information Analysis, Terminology, Clustering, Semantic Networks, Corpora Linguistics, Knowledge Acquisition.

Table des matières

1 Introduction

Le projet ILC[1](Infométrie Langage et Connaissance) vise à développer des outils d'analyse de l'information. Les travaux que nous menons actuellement dans le cadre de ce projet nous permettent de faire émerger une méthodologie d'acquisition de connaissance en corpus. Nous situons donc globalement la portée et la finalité de nos travaux avant de présenter de façon détaillée la spécificité de notre démarche en matière d'acquisition de connaissance.

1.1 L'analyse de l'information

Les progrès dans les télécommunications (INTERNET) et plus généralement dans le mode de diffusion de l'information(CD-ROM...) permettent à l'heure actuelle de collecter un nombre très important d'informations [20]. Les outils de navigation de type hypertexte ont considérablement changé la façon d'appréhender cette information en essayant de concilier deux dimensions opposées : volume de texte important *versus* accès rapide à l'information. Ces techniques n'impliquent aucune *compréhension* des textes et restent au niveau de la manipulation de chaînes de caractères et de pointeurs. Afin d'améliorer ces outils, une analyse plus détaillée des textes, des phrases et des mots, en tant qu'entités linguistiques est nécessaire. Le traitement automatique a atteint un stade de développement suffisant pour qu'il soit envisageable de prendre en compte des contraintes nouvelles venant des domaines de la recherche d'information et de l'extraction d'information à partir de textes. Les outils et méthodologies développés dans un tel contexte doivent être capables de manipuler un volume important de textes, d'extraire l'information essentielle et de la structurer.

Jacobs, dans [8], mentionne que les grands projets américains tels MUC ou TIPSTER ont clairement montré l'intérêt d'accorder une importance toute particulière aux mots et aux relations que les mots entretiennent entre eux. À l'inverse, il semble que des approches basées sur une grammaire ou des modèles du discours très développés, privilégiant donc fortement la syntaxe, handicapent plus le traitement qu'elles ne le favorisent [17].

1. ILC est un projet commun INIST-CNRS/INRIA Lorraine&CRIN-CNRS. La problématique du projet a été à l'origine d'une coopération plus large dans le cadre du Projet ILIAD financé par le GIS Sciences de la Cognition en 1996 et 1997.

1.2 Le projet ILC

Le projet ILC doit permettre à un opérateur humain de traiter l'information sans avoir à lire de façon séquentielle les documents. Son originalité repose sur la combinaison de méthodes linguistiques avec des méthodes statistiques afin de réaliser des outils plus robustes. Il s'appuie sur le fait actuellement bien établi que l'information dans les textes scientifiques et techniques se trouve localisée de façon privilégiée dans les groupes nominaux. De ce fait, notre stratégie d'analyse de l'information se fonde sur les avancées récentes en terminologie, tant sur les aspects linguistiques que sur les aspects connaissance.

ILC est dédié au traitement de résumés de textes scientifiques et techniques d'une base documentaire. Les outils actuellement développés fonctionnent sur les langues française et anglaise.

L'acquisition de connaissance en corpus est une expérimentation que nous avons mené à partir des concepts de base d'ILC. L'objectif de cette expérimentation était de montrer que l'association de méthodes linguistiques et statistiques se révélait pertinente pour extraire de l'information d'un corpus de résumés et faire émerger une structuration de l'information intrinsèque. Cette information structurée, commentée manuellement par des ingénieurs documentalistes spécialistes du domaine d'expérimentation constitue en soi une base de connaissance partielle du domaine.

1.3 Acquisition de connaissance en corpus

Nous présentons dans la suite de ce document, une méthodologie d'acquisition et de structuration de connaissances à partir de l'information présente dans un ensemble de textes scientifiques (les résumés bibliographiques). Cette méthodologie repose sur l'enchaînement de trois types de traitements :

1. des traitements linguistiques informatiques permettant de repérer les termes d'une nomenclature terminologique (thésaurus AGROVOC) ;

2. des traitements statistiques permettant de structurer ces termes en réseaux d'une part et d'en faire une partition en classes d'autre part ;

3. l'intervention des experts du domaine pour verbaliser le contenu des clusters et typer les liens de ces clusters.

Cette méthodologie[2] part de l'hypothèse que les liens de cooccurrence que les termes entretiennent entre eux sont des indicateurs forts de l'organisation conceptuelle de l'ensemble du corpus. Ce sont ces liens que nous cherchons ici à exhiber et typer, et qui doivent nous permettre de modéliser le contenu objectif d'un corpus de résumés dans une perspective d'accès à l'information.

2 Méthodologie

Au lieu de partir d'une modélisation *a priori*, fondée sur la représentation mentale que les experts ont du domaine, l'approche ILC part directement du corpus documentaire définissant le domaine d'application, pour faire émerger à travers le repérage terminologique, les connaissances contenues dans le corpus.

La méthodologie appliquée comporte quatre phases, deux phases entièrement automatiques (1 et 3), tandis que les deux autres supposent une intervention humaine (2 et 4) :

1. acquisition terminologique ;
2. contrôle du vocabulaire d'indexation ;
3. classification des termes en fonction de leurs cooccurrences (constitution des clusters) ;
4. verbalisation des clusters et typage des associations par des ingénieurs documentalistes du domaine d'application.

Les sections suivantes détaillent cette méthodologie. Nous présentons dans la section 3 le domaine d'application, les données et le thésaurus. La section 4 est consacrée aux méthodes et ressources d'ingénierie linguistique. Dans la section 5 nous expliquons les algorithmes de classification propres à SDOC et l'utilisation qui est faite de cette classification, ainsi que la phase préalable d'élimination des termes polysémiques (unitermes), trop génériques et très

2. Les textes que nous traitons sont des résumés bibliographiques, c'est à dire des textes courts d'une dizaine de lignes. Pour des textes plus longs allant d'un article à un livre, il faudrait trouver une unité textuelle pour le calcul de la cooccurrence, qui pourrait être la phrase ou le paragraphe.

fréquents. La section 6 est consacrée à l'intervention active des Ingénieurs Documentalistes : verbalisation de chacun des 40 clusters obtenus à l'issue de la classification ; typage des relations des termes de dix clusters choisis parmi ceux présentant les poids les plus forts ; et analyse détaillée du cluster CATECHOL OXIDASE. Nous faisons état dans la section 6.4 des problèmes rencontrés et des solutions que nous avons pu trouver. Enfin nous concluons dans la section 7 par une évaluation de l'expérimentation, nous indiquons les améliorations à apporter à la méthode ainsi que les voies nouvelles à explorer.

3 Domaine d'expérimentation

Le domaine d'application de l'expérimentation porte sur le traitement des fruits et légumes, de la récolte au conditionnement (sauf les semences et la germination). L'extraction des notices a été réalisée sur les fichiers de la base documentaire PASCAL des années 1993 et 1994.

Le corpus obtenu est composé de 1386 notices ayant un titre et un résumé anglais. Les dates de publication des notices sont : 1991 (1% des notices), 1992 (24%), 1993 (47%) et 1994 (28%). Les notices figurent dans 125 périodiques différents. Les 5 revues qui regroupent le plus de références sont : Journal of agricultural and food chemistry (18 % des notices), Journal of Food Science (12%), Food Chemistry (9%), HortScience (7%) et Lebensmittel-Wissenschaft + Technologie (5%). Ces 5 revues couvrent à elles seules 50% des notices du corpus.

17 pays d'affiliation ont été répertoriés, parmi lesquels les États-Unis occupent une position dominante avec 33% des notices. Dans les 5 premiers pays d'affiliation, l'Espagne vient en deuxième position avec 9% des notices, suivie de la France (6%) , du Canada (5%) et de l'Inde (5%).

Le thésaurus utilisé pour l'extraction terminologique est AGROVOC. C'est un thésaurus trilingue (français, anglais et espagnol) qui compte 14 714 termes préférentiels et 8 495 synonymes pour l'anglais.

L'extraction terminologique sur les titres et les résumés des 1386 notices a permis d'obtenir 2375 termes dont 932 de fréquence 1. Ce qui fait une moyenne d'environ 13 termes par notice.

4 Traitement linguistique du corpus

L'analyse linguistique de la phrase se base, comme nous l'avons mentionné, sur une analyse partielle de la phrase. Les notices bibliographiques collectées sont structurées en SGML. L'objet des traitements linguistiques est de pouvoir intégrer les résultats de ces traitements dans de nouveaux champs à cette structure. Chaque résumé devrait alors être caractérisé par les termes qui ont pu en être extraits automatiquement.

Nous décrivons ici les outils linguistiques et leurs possibilités pour le repérage des termes. Nous précisons le choix du type d'extraction terminologique que nous opérons. La plupart des analyseurs dédiés à la terminologie reposent sur une stratégie d'analyse locale de la phrase.

4.1 Analyse locale des termes

À l'heure actuelle, deux grandes familles d'analyseurs existent pour l'extraction terminologique : les analyseurs travaillant sans référentiel terminologique et cherchant à détecter des groupes nominaux (GN) terminologiques en corpus, et ceux utilisant une nomenclature et cherchant à en repérer les termes dans les textes sous leurs formes normales ou variantes, avec différents degrés de sophistication.

L'avantage de la première méthode réside dans la possibilité de détecter des termes nouveaux qui n'ont pas encore été répertoriés dans des thésaurus parfois mal maintenus. Malheureusement, les termes candidats qui se révèleront être de vrais termes aux yeux d'un expert sont noyés au milieu d'autres non nécessairement pertinents. Cet inconvénient majeur est dû au fait que, sur le plan linguistique, on ne sait pas distinguer un GN terminologique des autres GN. Différentes heuristiques linguistiques [2, 1] parfois combinées avec des calculs statistiques [6] sont utilisées mais elles n'imposent que des contraintes faibles sur la structure du GN qui amènent à retenir un très grand nombre de candidats.

La deuxième méthode – celle que nous utilisons – est fondée sur le repérage de termes appartenant à une nomenclature terminologique [9]. Un ensemble de règles, linguistiquement motivées, et qui définissent les possibilités de variation des termes, permettent d'identifier ces termes sous des formes qui peuvent être

éloignées de la forme d'enregistrement (forme du thésaurus), mais qui permettent de dire qu'il existe un lien conceptuel entre la forme observée dans le texte et le terme. L'identification des termes repose sur un mécanisme plus puissant et plus précis qu'un simple algorithme de collocation qui ne rechercherait ces termes que sur la base de la présence des mots le constituant, dans une fenêtre de n mots. Cette deuxième méthode a l'avantage de ne proposer que des termes motivés, représentatifs du corpus et en nombre limité. Le recours à la variation permet de ne pas s'arrêter au terme tel qu'il a été « figé » au moment de son enregistrement dans le thésaurus, mais à prendre en compte les possibilités langagières qui permettent d'exprimer le même concept sous des formes différentes.

4.2 La plate-forme linguistique

Les traitements terminologiques que nous réalisons s'appuient sur l'intégration de différents outils linguistiques. FASTR [10] est un analyseur robuste qui repose sur les grammaires d'unification et permet l'identification des termes d'une nomenclature terminologique et de leurs variantes dans de gros corpus de textes. Les règles sur les mots, les règles sur les termes et les métarègles sont écrites dans le formalisme logique PATR-II [18]. Il s'agit d'un formalisme linguistique neutre, indépendant de toute théorie linguistique et permettant l'utilisation de traits structurés et hiérarchiques.

Le module d'assignation des catégories grammaticales a été développé à l'INIST et permet l'étiquetage des mots de différents lexiques terminologiques anglais[3] en leur assignant une étiquette; en cas d'ambiguïté le mot reçoit la catégorie la plus probable en fonction de règles prédéfinies[4] [15, 4].

Ces outils permettent, à partir d'une nomenclature terminologique quelconque (dans notre expérimentation le thésaurus AGROVOC), de repérer des termes sous leurs formes de base ou leurs formes variantes. La figure 1 donne l'architecture globale de la plateforme linguistique dont nous détaillons ci-

3. Dans le cadre du projet ILC, une autre stratégie d'étiquetage syntaxique est utilisée pour le français avec l'étiqueteur de E. Brill [3]. Il s'agit d'un étiquetage sans dictionnaire à partir d'une méthode probabiliste à base de règles avec apprentissage.

4. Cet étiquetage se base sur le DELAF anglais, un dictionnaire électronique des formes fléchies dévellopé au LADL (CNRS/ Université Paris VII).

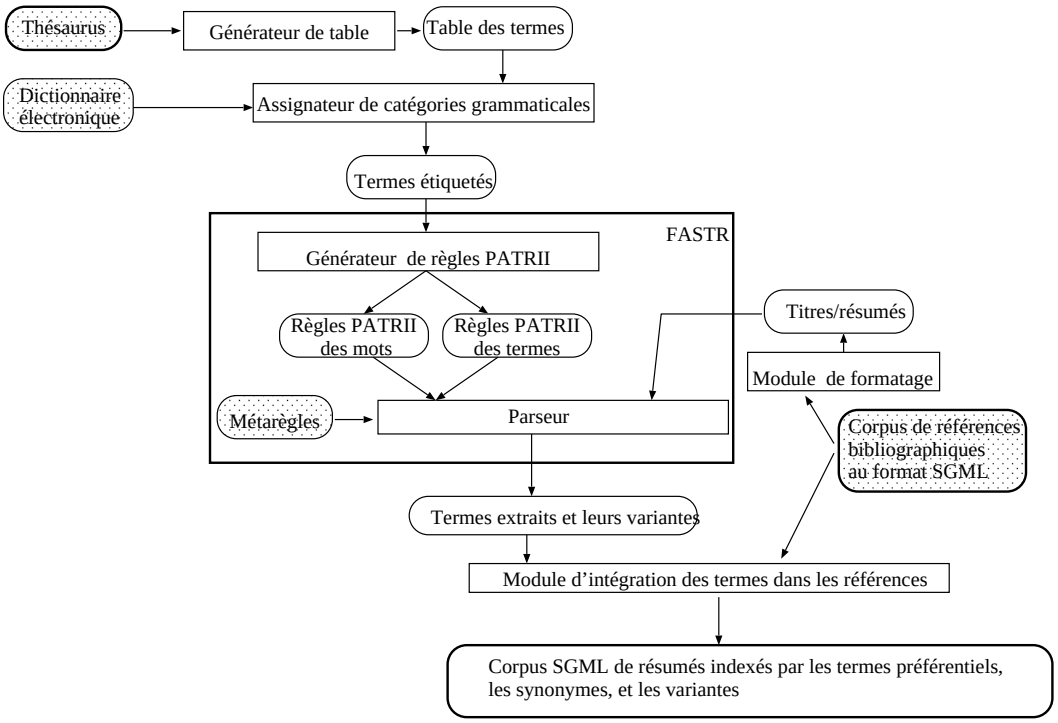

FIG. 1 – *Plateforme d'ingénierie linguistique*

dessous les différentes étapes. Les données en entrées (en grisé dans la figure) sont de quatre types : (1) Les notices bibliographiques desquelles sont extraits les titres et résumés pour l'analyse linguistique. (2) La terminologie de spécialité formée ici du thésaurus AGROVOC. (3) Un dictionnaire électronique permettant de réaliser l'étiquetage en partie du discours du thésaurus. Et (4) l'ensemble des métarègles qui permet de récupérer les termes variants. La phase préalable consiste à générer un fichier compilé de règles sur les mots et sur les termes à partir de la nomenclature terminologique des termes du thésaurus. La seconde phase consiste à extraire les titres et les résumés des enregistrements SGML de chaque référence bibliographique et de repérer à partir d'une analyse linguistique locale les termes de la nomenclature et leurs variantes (*Parseur*). Une dernière étape (*intégration des termes dans les références*) consiste à insé-

rer, dans le corpus au format SGML de départ, des enregistrements supplémentaires. Nous associons de ce fait les termes extraits aux notices bibliographiques initiales et nous gardons la trace du type de variation qui a permis d'identifier le terme quand c'est le cas, ainsi que la séquence textuelle sur laquelle porte cette variation.

4.3 Analyse en corpus

À partir des fichiers de règles compilés, FASTR, dans un second temps, va repérer les termes et leurs variantes en corpus. Il s'appuie sur deux catégories de variations : 1) la variation flexionnelle, 2) la variation syntaxique. Un troisième type de variations (morpho-dérivationnelle[5]) a été expérimenté avec succès sur d'autres corpus, mais n'a pas été intégré pour l'heure dans notre expérimentation.

La **variation flexionnelle** permet d'identifier pour chaque terme, les formes singuliers / pluriels des noms (*property / properties*), et les formes infinitives, participes passés et gérondives des « noms/verbes » (*seed plant : seed planting*). Dans les traitements que nous effectuons, chaque mot est décomposé en son lemme ou racine et sa terminaison. À chaque classe de mots correspond donc un lemme et ses différentes terminaisons.

Les termes sont identifiés selon trois types de **variation syntaxique** :

1. la variation d'insertion concerne tout mot à l'intérieur du groupe nominal, à l'exception de la plupart des mots grammaticaux. Par exemple, *ammonium dihydrogen phosphate* est associé au terme *Ammonium phosphate* ;

2. la variation de coordination concerne toute forme coordonnée de mots (adjectifs ou noms) à l'intérieur du groupe nominal. Par exemple, *apple and pear juice* est associé au terme *Apple juice* ;

3. la variation de permutation implique tous les mots ou les groupes de mots pouvant permuter autour d'un élément pivot (prépositions ou séquences

5. Elle permet d'intégrer les phénomènes de nominalisation et d'adjectivisation pour l'identification des termes en corpus.

verbales). Par exemple, *precipitation of organic acid* est associé au terme *Acid precipitation.*

Toutes ces variantes sont identifiées par des métarègles. Chacune de ces métarègles a pour fonction de décrire un type de variante. Elles s'appliquent sur chacune des règles des termes et les modifient dynamiquement afin de reconnaître ces termes sous leurs formes variantes. Pour reprendre notre exemple avec le terme *Seed planting*, la métarègle suivante :

```
Metarule   Perm (X1 -> X2 X3) =    X1 -> X3 P4 X2:
           <X2 cat>!               P
           <P4 lemma> =            'of'
           <X2 cat>!               A
           <X3 cat>!               P
           < X1 metaLabel> =       'XX'.
```

permet de traiter un cas simple de permutation. Elle comprend trois parties :

- Une partie gauche (à gauche du signe « = ») décrivant la structure d'un terme. Elle permet de reconnaître un terme X1, se réécrivant X2 X3. La catégorie X, définit un mot de catégorie syntaxique quelconque. Dans notre exemple X2 correspond au lemme *seed* et X3 au lemme *planting.*

- Une partie droite (à droite donc du signe « = ») qui définit une structure syntaxique acceptable pouvant être reliée à la forme de base d'un terme. Nous observons que cette partie de la métarègle autorise la permutation des éléments X2 et X3 autour d'un élément pivot, la préposition P4.

- En dessous se trouve la partie décrivant les contraintes imposées sur les mots. Elles sont de deux types : l'égalité (notée par le signe « = ») ou la négation (notée par le signe « ! »). L'égalité impose une contrainte stricte : la préposition P4 ne peut avoir comme lemme que *of.* Si nous n'avions pas imposé cette contrainte, la permutation aurait pu se faire autour de n'importe quelle préposition (ce qui est le cas dans les traitements que nous effectuons pour cette expérimentation). La négation est utilisée quand on veut faire porter l'interdiction sur un trait et autoriser tous les autres. Cela nous amène à ne pas considérer comme linguistiquement correcte la permutation d'un élément en position X2 quand il est adjectif

(A). De même, nous n'autorisons pas la permutation quand un élément d'un terme en position X2 ou X3 se trouve être une préposition.

Cette métarègle permet donc de reconnaître le terme *Seed planting* sous la forme : *planting of seed.*

5 Analyse statistique pour la classification des termes : SDOC

5.1 La méthode

SDOC [14, 7] applique la méthode des "mots associés". Cette méthode considère les mots-clés comme des indicateurs de connaissance (contenu des documents indexés) et se base sur leur cooccurrence pour mettre en évidence la structure de leurs relations (clusters). Un cluster est une classe de mots entre lesquels il existe des associations fortes.

L'idée de « cooccurrence » est essentielle. En effet, si on considère que deux documents sont proches parce qu'ils sont indexés par des mots-clés similaires, alors deux mots-clés figurant ensemble dans un grand nombre de documents seront considérés comme proches. Cependant, la cooccurrence ne permet pas à elle seule de mesurer la force des associations entre mots-clés (leur proximité), car elle avantage les mots-clés de haute fréquence par rapport à ceux de basse fréquence. L'emploi d'un indice statistique permet de normaliser la mesure de l'association entre deux mots-clés.

L'indice utilisé est l'indice d'Équivalence dont les valeurs varient entre 0 et 1: $E_{ij} = {C_{ij}}^2 / (C_i * C_j)$; où C_{ij} est le nombre de cooccurrences des mots-clés i et j, C_i la fréquence du mot-clé i, C_j la fréquence du mot-clé j.

5.2 La génération de clusters

À partir des mesures de proximité entre les mots, l'algorithme de classification hiérarchique du simple lien construit des groupes de mots proches les uns des autres (clusters) n'excédant pas une taille maximale (nombre de mots) fixée par l'utilisateur. Ainsi la figure 2 montre deux clusters $C1$ et $C2$

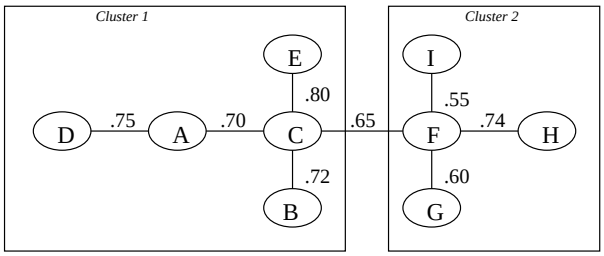

FIG. 2 – *Deux clusters C1 et C2 de 5 mots maximum*

contenant respectivement les mots-clés A, B, C, D, E d'une part et F, G, H, I d'autre part. Un cluster est donc constitué de mots associés les uns aux autres (associations internes). Les clusters peuvent avoir des relations entre eux. Ceci se produit lorsqu'il existe une association entre 2 mots-clés appartenant à 2 clusters différents (association externe) et que la taille du nouveau cluster qui aurait résulté de la réunion de ces 2 clusters dépasse la taille maximum définie par l'utilisateur. Ainsi $C1$ et $C2$ sont-ils reliés par une association externe entre C et F car la taille des clusters ne peut excéder un maximum de cinq mots dans l'exemple présenté.

Après le processus de classification des mots-clés, les documents sont affectés aux clusters. Un document est associé à un cluster, si, dans sa liste de mots-clés, il existe au moins un couple de mots-clés qui pourrait constituer une association interne ou externe du cluster.

La classification est principalement paramétrée par le nombre maximal de mots pouvant constituer un cluster. C'est une variante de la procédure statistique habituelle qui consisterait à utiliser un seuil fixe (une « distance limite » à partir de laquelle aucune agrégation n'est plus effectuée). C'est un moyen pratique pour moduler la coupure dans l'arbre de classification (dendrogramme). En conséquence du critère de taille maximale, les classes résultantes sont très hétérogènes en densité. La première classe obtenue sera constituée des mots-clés les plus fortement liés alors que la dernière sera très lâche, restituant en cela la structure du réseau d'associations. On peut également limiter le nombre d'associations intra ou inter-clusters dans un souci de lisibilité.

5.3 Les caractéristiques des clusters

Un cluster est composé de

- une liste de mots-clés,
- une liste d'associations internes,
- une liste d'associations externes,
- une étiquette,
- une liste de documents affectés au cluster après la classification.

La liste des mots-clés regroupe des mots qui sont proches les uns des autres. Nous distinguons les mots-clés internes (qui apparaissent dans les associations internes) des mots-clés externes (qui apparaissent seulement dans les associations externes car ils ont été rejetés de ce cluster à cause du critère de taille maximale des clusters). Les mots-clés sont triés selon leur nombre d'apparitions dans les associations internes et externes du cluster.

La liste des associations internes décrit la force des associations des mots qui définissent la structure interne des clusters. Plus la valeur de l'association est forte, plus les mots sont fortement associés.

La liste des associations externes décrit les associations existants entre les mots d'un cluster et les mots d'autres clusters. Le nombre d'associations externes peut être limité aux n plus fortes. Dans ce cas, les associations externes ne sont pas nécessairement bidirectionnelles. Dans le cas présent, nous l'avons limité aux 10 plus fortes.

L'étiquetage des clusters: le choix d'un terme représentatif pour nommer le cluster est basé sur une heuristique. Nous choisissons le terme de la liste des mots-clés internes qui apparaît le plus grand nombre de fois dans les associations internes et externes. Le nom proposé est satisfaisant dans plus de 90% des cas.

La liste des documents affectés à un cluster: elle est obtenue après exécution de la classification. C'est la liste des documents qui ont contribué à la formation de ce cluster par la présence dans leur indexation de couples de mots-clés qui pourraient constituer une association interne ou externe du cluster. Un document peut donc figurer dans plusieurs clusters. Un document

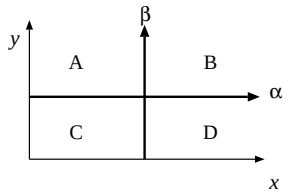

FIG. 3 – *Visualisation des clusters en fonction de la densité et de la centralité*

ne figurant que dans un seul cluster est appelé document propre au cluster. Les documents sont triés selon l'importance de leur contribution à l'élaboration du cluster. À partir des documents sont extraits le titre, les auteurs et la source pour compléter la description du cluster.

Un tableau résumant les caractéristiques structurelles des clusters permet de les catégoriser et d'apprécier la répartition des documents dans les clusters (voir dans les annexes).

5.4 Cartographie : densité (y) et centralité (x)

Deux valeurs structurelles caractérisent les clusters de mots associés : la « densité » et la « centralité ». La « densité » d'un cluster est exprimée par la valeur moyenne des associations entre mots-clés formant le cluster, ou associations internes. La « centralité » d'un cluster est exprimée par la valeur moyenne des associations entre les mots qui le constituent et les mots d'autres clusters, ou associations externes.

Ces valeurs sont ensuite utilisées pour positionner les clusters sur un plan bidimensionnel (y,x), c'est-à-dire une carte (Fig. 3). On peut ainsi repérer les clusters (ou réseaux lexicaux) les mieux structurés du point de vue de leur « densité » (ou cohésion) et les mieux rattachés au réseau (centralité). Sur une telle carte, la proximité entre deux clusters (ou réseaux lexicaux) indique qu'il sont structurellement proches, mais leurs contenus sémantiques ne sont généralement pas voisins. Pour chacune des dimensions exprimées sur cette

carte, il est possible de tracer arbitrairement une droite délimitant deux sous-ensemble de clusters :

- Pour la densité, une limite α entre les clusters très denses et ceux qui sont moins denses, à savoir ceux qui ont une forte cohésion en interne et ceux qui, à l'opposé, sont plus généraux. Ainsi, on séparera les deux zones A et B (très denses) des deux zones C et D (faiblement denses),

- De même pour la centralité, une limite β entre les clusters dont les termes sont fortement liés avec d'autres clusters (zones B et D) en opposition avec les clusters faiblement liés (zones A et C).

De façon empirique, la zone B est celle qui contiendra les clusters les plus intéressants pour notre analyse.

La carte permet à l'analyste d'avoir une appréhension globale (l'ensemble des clusters) et locale (les liens entre certains clusters). La carte correspondant à notre expérimentation est donnée en Fig. 4.

En conclusion : les clusters et la carte (où l'on visualise la position de l'ensemble des clusters et les associations inter-clusters) représentent des structures statistiques pour la représentation de connaissance. Ce support statistique est construit automatiquement à partir de l'extraction automatique des termes des documents (titres et résumés). Il s'agit d'une structure intermédiaire pour que les experts du domaine réalisent les opérations sémantiques de (1) verbalisation des clusters et de (2) typage des associations entre les termes constituant les clusters.

6 Interprétation des résultats

L'intervention des experts a consisté dans un premier temps en un nettoyage des données bruitées dues aux unitermes (polysémies et ambiguïtés, cf. section 6.4). La partie la plus importante de leur travail a été consacrée à une analyse des résultats, guidée par la méthode, et montrant l'organisation conceptuelle des termes. La verbalisation des clusters a permis de se faire une idée globale du contenu du corpus par un décodage de cette structure abstraite que représentent les clusters. Le typage des relations a montré que la cooccurrence met en évidence des liens conceptuels entre les termes du corpus,

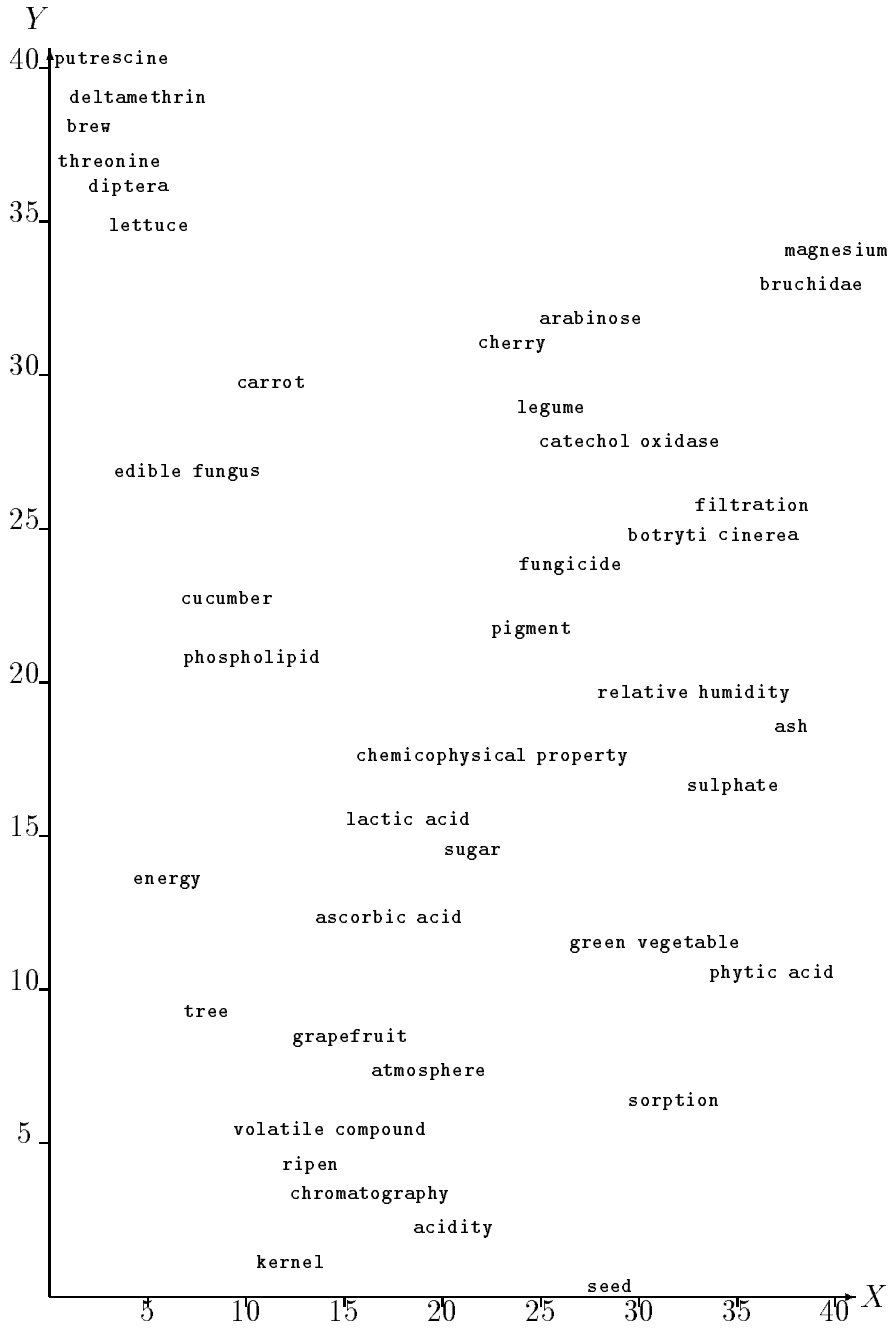

FIG. 4 – *Visualisation des résultats de notre expérimentation*

que ces liens peuvent être nommés et qu'il est possible d'en dresser une typologie. Enfin l'analyse qui a été faite d'un cluster permet de se rendre compte que l'agrégat de relations qu'il représente forme un ensemble structurellement cohérent.

6.1 La verbalisation des clusters

Les 40 clusters issus de la classification ont pu être verbalisés, c'est-à-dire décrits succinctement. Nous dressons un tableau (cf. Table 1) de cette verbalisation en présentant les clusters par ordre décroissant de densité.

6.2 Les types de relations observées

Pour les clusters observés, nous donnons une vue d'ensemble des relations mises en évidence par les liens de cooccurrence. Un grand nombre de ces relations est identifié et reçoit un nom. Dans la typologie que nous proposons, nous identifions deux grandes familles de relations, celles qui entrent dans les paradigmes classiques des relations de classification : synonymie, relation hyponyme / hyperonyme (spécifique - générique), méronymique (partie de, composant de) ; et celles plus complexes et plus difficiles à nommer qui mettent en évidence un processus ou une propriété, que nous nommons relations prédicatives.

6.2.1 Relations de classification

Le tableau 2 donne un aperçu de ces différentes relations. La première classe concerne les équivalences strictes entre un terme latin et son équivalent dans la langue vernaculaire. La seconde permet de regrouper des termes en classes d'équivalence. C'est ainsi que les experts[6] ont pu identifier des classes de termes telles que les FRUITS À NOYAUX, CHAMPIGNONS PATHOGÈNES, etc.

Une des propriétés remarquables de ces termes est que chacune de ces classes d'équivalence forment des cliques[7] dans le réseau de termes. Ils sont regroupés sous un même hyperonyme par les experts du domaine.

6. Nous remercions ici Bénédicte Dimbert et Dominique Vachez (ingénieurs documentalistes à l'INIST) qui ont travaillé à l'interprétation des clusters et de leurs relations.

7. Une clique est un graphe ou sous-graphe où tous les noeuds sont reliés entre eux.

Clusters	Verbalisation
PUTRESCINE	Amine biogène.
DELTAMETHRIN	Insecticide.
BREW	Aliment infantile.
THREONINE	Aminoacide.
DIPTERA	Lutte contre les insectes (fumigation).
LETTUCE	Accumulation biologique dans les légumes.
MAGNESIUM	Composition des fruits/légumes en oligoéléments, éléments minéraux et métaux lourds.
BRUCHIDAE	Infestation des légumineuses par les Bruchidae (coléoptères) et désinfestation.
ARABINOSE	Glucides de structure (fibre et parois cellulaires) des fruits et légumes.
CHERRY	Qualité des fruits à noyaux.
CARROT	Qualité des légumes.
LEGUME	Qualité des légumineuses amylacées.
CATECHOL OXYDASE	Brunissement des fruits et légumes.
EDIBLE FUNGUS	Champignons comestibles.
FILTRATION	Purification préparatoire à l'analyse biochimique.
BOTRYTIS CINEREA	Contamination des fruits par les champignons saprophytes et pathogènes.
FONGICIDE	Lutte chimique contre les champignons.
CUCUMBER	Conservation des fruits et légumes - Saumurage.
GRAPEFRUIT	Agrumes - Jus de fruits et légumes.
PIGMENT	Pigments, pigmentation.
PHOSPHOLIPIDES	Composition lipidique des fruits et légumes.
ASH	Composition chimique des fruits et légumes.
CHEMICOPHYSICAL PROPERTY	Relation composition et propriétés rhéologiques de colloïdales.
SULPHATE	Analyse biochimique par méthode de séparation.
RELATIVE HUMIDITY	Effet des traitements physiques et de l'entreposage sur la qualité des fruits et légumes.
LACTIC ACID	Fermentation.
SUGAR	Composition en glucide et acides organiques des fruits.
ENERGY	Traitement thermique - Tranfert d'énergie.
ASCORBIC ACID	Composition en vitamine C.
GREEN VEGETABLE	Traitement par le froid des légumes et fruits.
PHYTIC ACID	Digestibilité et valeur nutritive des légumineuses, facteurs antinutritionnels.
TREE	Olive, huile d'olive.
ATMOSPHERE	Emballage des fruits et légumes en vue de leur conservation et leur qualité.
SORPTION	Transfert de masse : absorption et désorption d'eau.
RIPEN	Maturation des fruits et qualité.
CHROMATOGRAPHY	Analyse biochimique par chromatographie phase liquide.
ACIDITY	Propriétés organoleptiques et physico-chimiques, effets des traitements appliqués.
VOLATILE COMPOUND	Analyse de la composition en composés volatils par chromatographie phase gazeuse et spectrographie de masse.
KERNEL	Traitement appliqués aux graines et grains.
SEED	Graines et relation sources/puits avec les parties de la plante.

TAB. 1 – *Verbalisation des clusters*

Relations		Exemples
Synonyme		Apricot & Prunus armeniaca
		Cherry & Prunus avium
		Lens Culinaris & Lentil
Classe d'équivalence Sémantique		
Fruits à noyaux		Apricot Plum
Légumes / Fruits		Cucumber, Tomato
Champignons pathogènes		Botritis cinerea, Penicillium expansum
Acides gras		linoleic acid, Linolenic acid
Métaux et éléments minéraux		Copper, Zinc,Maganese,Iron,Magnesium,etc.
Céréales		Hordeum vulgare, Triticum æstivum
Est_un		
Légume	←–	Vigna subterranea
Légume	←–	Groundnut
Légume	←–	Lentil
Légume	←–	Kidney bean
Légume	←–	Phaseolus vulgaris
Fungus	←–	Aspergillus flavus
Partie_de		
Granule	←–	Potato starch
Granule	←–	Amylose
Lentil	←–	Amylose
Légume	←–	Amylose

Tab. 2 – *Relations de classification*

La troisième classe met en relation deux termes dont l'un est est un hyperonyme et l'autre un hyponyme dans une relation de type thésaurus *Est_un*. Nous donnons dans le tableau 1 un exemple de certains termes qui cooccurrent avec l'hyperonyme *légume*.

La dernière classe, portant sur la relation *Partie_de* permet d'apparier des termes dont la cooccurrence traduit une relation d'appartenance.

6.2.2 Relations dénotant des processus

Les relations du tableau 3 mettent en évidence des relations de type processus. Elles sont nommées par des prédicats qui sont, dans la plupart des cas, des nominalisations de verbes. Cette structure prédicative est apparentée à celle que l'on peut rencontrer dans les groupes nominaux (GN) complexes. Sous sa

forme minimale, c'est-à-dire quand le prédicat est relié à un verbe strictement intransitif, elle admet au moins un argument qui est le sujet. Tout comme pour les GN complexes, un ou plusieurs arguments peuvent être omis quand l'information qu'ils portent est très générale ou facilement déductible. Les rôles thématiques que nous utilisons reprennent pour chaque prédicat sous sa forme verbale, les fonctions grammaticales de la phrase : le sujet est généralement l'agent, le rôle thématique thème est l'objet et les circonstants désignent les compléments circonstanciels.

Nous mettons en évidence deux types de prédicats : ceux qui ont comme arguments un couple du type Agent / Thème et ceux qui ont comme arguments un triplet Agent / Thème / Circonstant, mais pour lequel l'argument agent est régulièrement omis. Les prédicats tels que INFESTATION, CONTAMINATION, BIODETERIORATION, entrent dans la première catégorie ; les prédicats tels que TRAITEMENT,PURIFICATION, ANALYSE, etc. entrent dans la deuxième catégorie avec le plus souvent un ou plusieurs arguments omis, notés par le symbole $\emptyset$ (pour catégorie vide).

Les relations de cooccurrences mettent en évidence ces prédicats sous deux formes. La plus fréquente est constituée par des liens de cooccurrence qui relient entre eux Agent et Thème ou Thème et Circonstant. Dans ce cas, c'est l'expert du domaine qui nomme la relation avec un prédicat. Dans le tableau 3 les associations [*Cydia pomonella* & *Nectarine*] et [*Callosobruchus maculatus* & *Cowpea*] ont été identifiées comme une relation pouvant être désignée sous le nom INFESTATION, et où l'Agent est de la classe INSECTE et le Thème appartient à la classe FRUITS/LÉGUMES. De la même façon, les associations [*Methyl bromide* & *Nectarine*] et [*Brine* & *Cucumber*] définissent un processus nommé TRAITEMENT et qui a comme argument un Circonstant appartenant à la classe SUBSTANCES/PROCÉDÉS et un Thème de la classe FRUITS/LÉGUMES. Il faut remarquer que l'Agent est toujours une catégorie vide ($\emptyset$), référant ici à l'expérimentateur. D'une façon générale, pour ce type de relation, quand l'agent appartient à la classe humain, il est toujours omis.

Nous repérons les relations prédicatives de façon plus naturelle quand le prédicat apparaît explicitement dans la cooccurrence (notées par le symbole $\oplus$ dans ce cas, et par un $\ominus$ dans les autres cas). Dans le tableau 3, le terme INFESTATION est relié aux termes *Larva*, *Insecta* et *Coleopteron*, tous trois Agents de ce prédicat ; dans ce cas précis, le Thème est une catégorie vide.

Processus	Agent	Thème	Circonstant
Infestation	Insectes	Fruits/légumes	
⊕	Bruchidæ	Cowpea	∅
⊕	Larva	∅	∅
⊕	Insecta	∅	∅
⊕	Coleopteron	∅	∅
⊖	Cydia pomonella	Nectarine	∅
⊖	Callosobruchus maculatus	Cowpea	∅
⊖	Zabrotes subfaciatus	Kidney bean	∅
Contamination	Champignons/mycotoxines	Fruits/légumes	
⊖	Aflatoxine	Pistachio	∅
Biodéterioration	Champignons/mycotoxines	Fruits/légumes	
⊕	Botrytis cinerea	Malus pumila	∅
⊕	Penicillium expansum	Malus pumila	∅
Production			
⊖	Fungus	Aflatoxine	∅
Effet			
⊖	pH	Cucumber	∅
⊖	Phytate	Biodisponibility	∅
Action			
⊖	Lipoxygenase	Linoleic acid	∅
Traitement		Fruits/légumes	Substances/Procédés
⊖	∅	Nectarine	Methyl bromide
⊖	∅	Cucumber	Brine
⊖	∅	Cowpea	Soak
Purification		Substance	Procédés
⊕	∅	Isozyme	Ion exchange chromatography
⊕	∅	∅	Column chromatography
⊕	∅	∅	Electrophoresis
⊕	∅	∅	Filtration
Analyse			
⊕	∅	Cucumber	Gas chromatography

Tab. 3 – *Relations prédicatives dénotant des processus*

De la même façon le prédicat Purification apparaît ici relié au Circonstant : *Ion exchange chromatography*, *Column chromatography*, *Electrophoresis* et *Filtration*.

Les cas les plus rares sont ceux où la cooccurrence permet de mettre en évidence une relation prédicative complète. On remarque ainsi le lien de cooccurrence entre le terme INFESTATION et son Agent *Bruchida* d'une part, et le lien de cooccurrence entre ses deux arguments (Agent et Thème) : *Bruchidae* et *Cowpea*.

6.3 Analyse d'un cluster : CATECHOL OXIDASE

6.3.1 Description infométrique

Cluster : 28 CATECHOL OXIDASE

Centralité,Densité : 0.05,0.15
Nombre de documents relatifs au cluster : 84
Nombre de documents spécifiques au cluster : 6
Nombre de sources des documents relatifs au cluster : 18
Nombre d'auteurs des documents relatifs au cluster : 241
Coefficient de saturation : 0.07
Indice de cohésion : 0.15
Indice de centralité : 0.05
Nombre de citations par les autres clusters : 8

Termes du cluster :

Poids	Fréquence	Mots-clé
0.27	19	Catechol oxidase
0.27	29	Polyphenol
0.23	74	Enzyme
0.23	41	Phenolic compound
0.23	30	Oxidoreductase
0.19	34	Browning
0.12	17	Polygalacturonase
0.12	24	Oxidation
0.08	7	Cellulase
0.04	4	Maillard reaction
0.12	28	Purification
0.04	63	Pectin
0.04	72	Ascorbic acid
0.04	34	Flavonoid

Associations internes :

Poids	Coocurrences	Association
0.66	19	Catechol oxidase & Polyphenol
0.63	19	Catechol oxidase & Oxidoreductase
0.41	19	Oxidoreductase & Polyphenol
0.21	5	Cellulase & Polygalacturonase
0.14	13	Phenolic compound & Polyphenol
0.13	10	Catechol oxidase & Phenolic compound
0.10	11	Enzyme & Polygalacturonase
0.08	9	Oxidation & Phenolic compound
0.08	10	Oxidoreductase & Phenolic compound
0.08	7	Browning & Catechol oxidase
0.07	6	Cellulase & Enzyme
0.07	3	Browning & Maillard reaction
0.07	12	Enzyme & Oxidoreductase
0.05	7	Browning & Polyphenol
0.05	7	Browning & Oxidoreductase
0.04	7	Catechol oxidase & Enzyme
0.02	4	Oxidation & Polyphenol
0.02	7	Enzyme & Polyphenol
0.02	3	Catechol oxidase & Oxidation
0.02	7	Enzyme & Phenolic compound

Associations externes :

Avec le cluster filtration

0.07	6	Catechol oxidase & Purification
0.04	6	Polyphenol & Purification
0.04	6	Oxidoreductase & Purification

Avec le cluster arabinose

0.06	8	Polygalacturonase & Pectin

Avec le cluster ascorbic acid

0.05	11	Browning & Ascorbic acid

Avec le cluster pigment

0.05	8	Phenolic compound & Flavonoid

Documents relatifs au cluster :

Poids	référence	Titre
0.14	000714	Partial purification of soluble potato polyphenol oxidase by partitioning in an aqueous two-phase system
0.11	000353	Oxidative reactions of caffeic acid in model systems containing polyphenol oxidase
0.10	000715	Cresolase activity of potato tuber partially purified in a two-phase partition system
0.10	000058	Phenolic composition and browning suscepbtibility of various apple cultivars at maturity
0.10	000555	Potential purification and some properties of Monroe apple peel polyphenol oxidase
0.09	000095	Prevention of enzymatic darkening in frozen sweet potaotes Ipomoea batatas (L.) Lam. by water blanching: relationship among darkening, phenols, and polyphenol oxidase activity
0.09	000518	Physiological attributes related to quality attributes and storage life of minimally processed lettuce
0.09	000358	Cysteine as an inhibitor of enzymatic browning. II: Kinetic studies
0.08	001081	Characterization and inhibition of polyphenol oxidase from pears (Pyrus communis L. cv. Bosc and Red)
0.08	001072	Browning phenomena in stored artichoke (Cynara scolymus L.) heads: enzymic or chemical reactions
0.08	000517	Lychee pericarp browning caused by heat injury
0.07	000359	Oxidation of chlorogenic acid, catechins, and 4-methylcatechol in model solutions by apple polyphenol oxidase
0.07	000781	Immunochemical and immunohistochemical study of apple chlorogenic acid oxidase
0.07	000368	Polyphenol oxidase from sweet potato: purification and properties
0.07	000653	Partial purification and properties of plantain polyphenol oxidase
0.06	000554	Studies on inhibition of mushroom polyphenol oxidase using chlorogenic acid as substrate
0.06	001146	Effect of polyphenol oxidase and its inhibitors on anthocyanin changes in plum juice
0.05	000203	Characterization of polyphenoloxidase from Stanley plums
0.05	001104	New approaches for separating and purifying apple polyphenol oxidase isoenzymes: hydrophobic, metal chelate and affinity chromatography
0.05	000626	Production of peroxidase enzyme by callus cultures of Citrus aurantifolia S

6.3.2 Analyse du cluster

L'analyse experte a montré qu'un certain nombre d'artefacts s'étaient introduits dans les relations du cluster. Ces problèmes sont dus pour l'essentiel

à la propagation des relations de synonymie vers des termes dits préférentiels. Ils sont détaillés dans la section 6.4.

Une fois les relations artificielles retirées, le cluster reflète bien le phénomène du brunissement enzymatique des fruits et légumes. Ce phénomène est une réaction d'oxydation des composés phénoliques (polyphénols, flavonoïdes...) catalysée par des enzymes, dont la *catechol oxidase*. L'acide ascorbique peut lui même être un substrat de la réaction, et, pour cette raison, est souvent utilisé comme antioxygène dans la prévention du brunissement enzymatique.

Le cluster, tel qu'il est formé ont amené les experts à faire certaines observations et à se poser des questions par rapport à ce que l'on s'attendait à trouver et que l'on ne remarque pas :

1. Le brunissement est une altération des fruits et légumes, or aucun nom de fruit ou de légume n'apparaît dans le cluster.

2. D'autres enzymes et substrats interviennent dans ce phénomène sans qu'ils apparaissent ici. On remarque par exemple que les substrats de la réaction de Maillard n'apparaissent pas dans le cluster.

3. Enfin les termes Cellulase et Polygalacturonase jouent un rôle marginal dans le cluster, leur présence s'explique uniquement par leur cooccurrence avec Enzyme (relation sorte de).

La principale remarque que l'on puisse faire par rapport à ces observations est que le processus de classification tel qu'il s'opère ne reflète que très partiellement le schéma cognitif des experts. Il met en évidence les régularités propres au corpus qui apparaissent explicitement et agrège dans des clusters les termes qui sont les plus représentatifs. L'examen des titres des résumés des références bibliographiques de ce cluster montrent que les études portent sur une grande diversité de fruits et légumes. Cela a pour effet de rendre le processus de brunissement que reflète le cluster CATECHOL OXIDASE non attaché à des espèces particulières. Si l'on examine les clusters qui impliquent des légumes ou des fruits on observe qu'ils ont été rassemblés dans des classes qui les identifient comme objets d'étude particulier (cluster CHERRY qui porte sur les fruits à noyaux) ou comme sous-ensembles clairement définis faisant l'objet d'un traitement ou d'un problème particulier : le cluster GRAPE FRUIT qui regroupe les fruits permettant la production de jus de fruits, le cluster GREEN

VEGETABLE qui porte sur le traitement par le froid de légumes et fruits, le cluster RIPEN qui concerne la maturation des fruits.

6.3.3 Représentation formelle du cluster

Afin de mettre en évidence la capacité des liens de cooccurrence à décrire des phénomènes de nature complexe dans les clusters, nous avons utilisé le formalisme des graphes conceptuels [19]. Nous montrons que le graphe de co-occurrence du cluster CATECHOL OXIDASE instancie partiellement un modèle.

Le modèle: L'analyse que nous faisons du cluster nous montre que deux types d'entités sont présentes, celles qui réfèrent à des composés chimiques et celles qui réfèrent à des processus. Dans ce cluster, on est en présence de trois ensembles de composés chimiques : les catalyseurs, les substrats d'oxydation et l'oxygène (comme agent d'oxydation).

Les processus appartiennent à la classe du brunissement (Browning) qui se subdivise en brunissement enzymatique et brunissement non-enzymatique.

Les relations sont du type : *Est_un* quand il s'agit d'établir une taxinomie des composés chimiques ou des processus ; ou Agent, Thème et Circonstant, dans le cas de processus.

La figure 5 donne le graphe des relations taxinomiques qui définissent les entités du modèles :

Dans ce graphe construit à la main, les termes notés en majuscules et en français correspondent à des concepts de plus haut niveaux qui n'apparaissent pas dans les associations ou à des concepts de niveau intermédiaire nécessaires à la classification. Il est à remarquer que ce graphe à lui seul décrit la presque totalité des termes du cluster.

Le modèle décrivant le phénomène du brunissement enzymatique est représenté dans le graphe de la figure 6. Les agents, thèmes et circonstants correspondent au trois classes de composés chimiques du graphe taxinomique : les catalyseurs, les substrats d'oxydation, et l'oxygène.

L'instanciation du modèle: Les relations qui unissent les termes du cluster par les liens de cooccurrence, permettent l'instanciation du modèle. Les

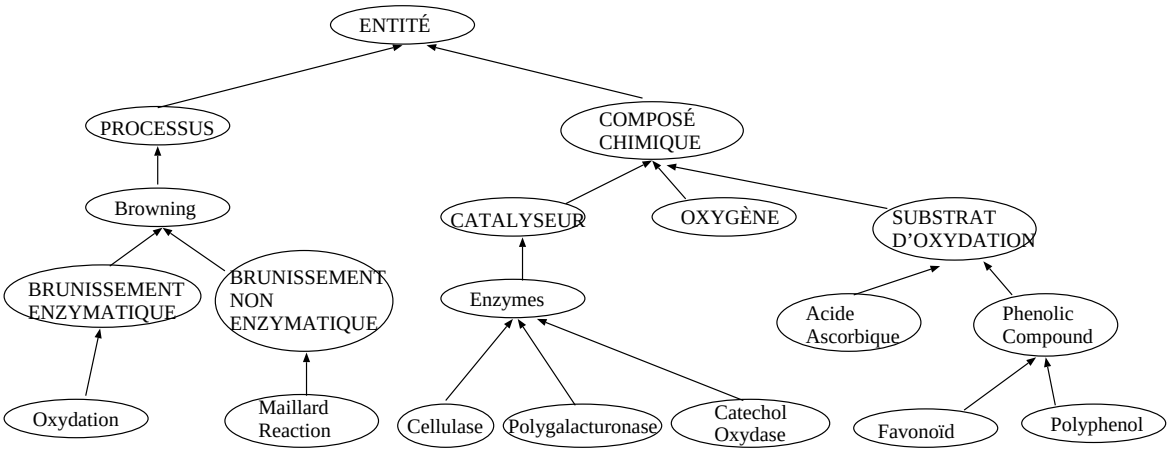

FIG. 5 – *Taxinomie des entités*

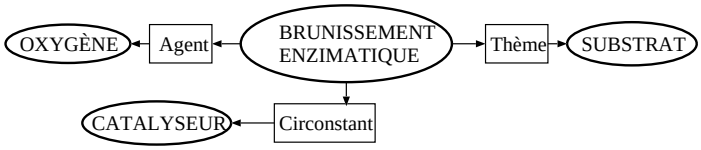

FIG. 6 – *Modèle décrivant le concept du brunissement*

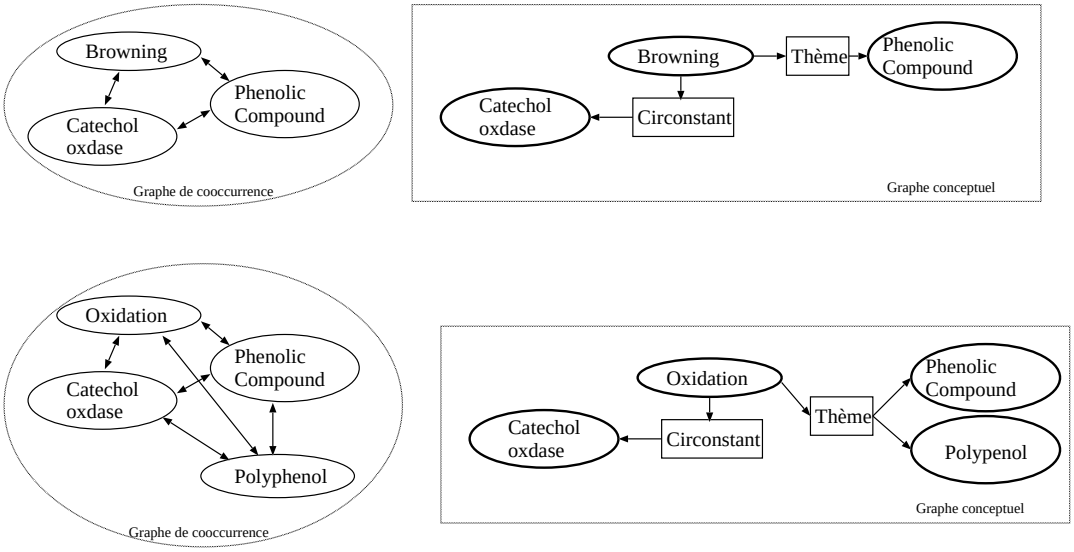

FIG. 7 – *Exemple d'instanciation du modèle : processus browning et oxidation*

graphes de la figure 7 sont des vues de cette instanciation pour la description des processus de brunissement et d'oxydation.

La figure 7 met en évidence une partie du graphe de cooccurrence concernant les processus BROWNING et OXIDATION. Nous montrons comment ce graphe de cooccurrence instancie le modèle de la figure 6. Nous avons extrait du graphe du cluster deux sous-graphes possédant la propriété d'être fermés transitivement. C'est cette propriété de fermeture transitive qui induit la relation prédicative entre le thème et le circonstant.

Si l'on restitue ces deux processus dans le graphe de la figure 5 décrivant la hiérarchie des entités, il apparaît que BROWNING est un terme beaucoup plus générique que OXIDATION. On constate parallèlement que OXIDATION participe à un graphe plus riche et plus spécifique (introduction du terme *polyphenol*).

6.4 Problèmes rencontrés

Un système automatique pour l'analyse de l'information et la structuration des connaissances est nécessairement confronté au **bruit**. Dans le cas présent,

le bruit rencontré est d'une part, de nature linguistique et, d'autre part, structurellement lié au thésaurus. Nous montrons ici les problèmes auxquels nous avons été confrontés, quelles solutions provisoires nous avons trouvé pour faciliter l'analyse et quelles solutions sont envisageables pour l'avenir.

6.4.1 Problèmes de nature linguistique

Nous avons vu dans la section 4.1 que les traitements linguistiques fondés sur l'analyse de gros corpus, et plus particulièrement ceux dédiés à la terminologie, privilégient les traitements locaux, contrairement à ceux fondés sur l'analyse de la phrase dans sa totalité. L'analyse d'une phrase entière suppose une ingénierie linguistique complexe qui ne s'avère vraiment nécessaire que dans des cas particuliers (traduction automatique, génération automatique de textes, etc.). Dans notre cas particulier, le domaine de localité est le groupe nominal. De ce point de vue, tout terme simple (formés d'un seul mot) est une entité libre qui ne subit aucune contrainte terminologique (contrairement aux termes formés de plusieurs mots) et est en soi un facteur de bruit.

Faute de temps pour l'expérimentation, nous n'avons pu focaliser notre attention sur la pertinence de la variation et sur les variantes erronées. Les problèmes linguistiques rencontrés les plus nombreux et les plus difficiles à traiter ont été essentiellement ceux liés aux mots simples (unitermes). Dans la phase initiale, celle qui a précédée la classification, nous avons donc repéré les noms simples par ordre de fréquences décroissantes et nous avons éliminé :

- les plus fréquents,
- les plus génériques,
- les plus polysémiques.

C'est cette dernière catégorie qui est la plus problématique. Elle met en évidence deux catégories de problèmes :

- les noms intrinsèquement ambigus ou ayant des dénominations mauvaises ou mal adaptées,
- les noms présentant des ambiguïtés de type adjectif / nom ou nom / verbe.

Un grand nombre de termes simples sont très ambigus. Nous avons cherché à éliminer les plus fréquents dans la phase préalable à la clusterisation. C'est ainsi qu'un terme comme *skins* désigne la peau des fruits alors que dans le thésaurus il est associé aux « cuirs et peaux » ; le terme *pearls* a la signification de « perles d'huîtres » et est rencontré dans le terme *pearl millet* qui n'existe pas dans Agrovoc. Ces deux termes ont en commun de n'avoir aucun lien avec la thématique du corpus et peuvent donc être repérés avant ou après clusterisation. Le terme *precipitation* est plus problématique car il a le sens météorologique (pluie) et le sens de « précipitation chimique ». Dans le thésaurus, seul ce dernier emploi est retenu.

Beaucoup de noms en anglais ont la particularité d'avoir la même forme graphique qu'un verbe à l'infinitif. Ceci n'est pas problématique quand le nom se réfère à un processus qui peut s'exprimer sous une forme verbale ou nominale. En revanche quand la forme graphique du nom ne partage pas le sens de la forme verbal nous sommes en présence d'une ambiguïté. C'est le cas des termes tels que *lead* (plomb) et le verbe *to lead*, *can* (boîte de conserve) et *can* le verbe modal, etc. Le fait que la plupart des noms ont un emploi adjectif quand ils sont placés à gauche d'un nom rend peu problématique l'ambiguïté adjectif/nom, sauf pour les cas particuliers où l'adjectif et le nom, malgré une graphie semblable, ont des sens différents. Ainsi le terme *Greens* qui désigne les « légumes verts » se trouve être ambigu avec l'adjectif qui désigne la couleur.

6.4.2 Problèmes liés au thésaurus.

Ces problèmes concernent principalement les choix documentaires retenus quand un terme a plusieurs sens d'une part, et la gestion des renvois de synonymie d'autre part. Les ambiguïtés inhérentes aux termes simples conduisent les concepteurs de thésaurus a privilégier un emploi plutôt qu'un autre. C'est ainsi que le terme *Pulp* désigne la « pulpe des fruits » dans le corpus utilisé, alors que dans son usage documentaire il est lié à la « pâte cellulosique » dans l'industrie du papier.

Les problèmes liés au renvoi de synonymie sont parmi les plus importants. Il ne faut pas perdre de vue qu'un thésaurus n'est pas conçu originellement dans le cadre d'une automatisation du traitement de la langue pour la recherche documentaire, mais pour guider un utilisateur dans la formulation de

sa requête. Les synonymes, de ce point de vue, peuvent être divisés en deux catégories :

- ceux qui sont des mots ambigus et pour lesquels on cherche à attirer l'attention de l'utilisateur sur l'emploi qui est fait de ce mot dans les documents indexés. Par exemple, *Skin* renvoie sur *Hides and skin* pour bien préciser qu'il s'agit d'un emploi limité au domaine du cuir ;
- les synonymes qui sont des équivalences d'un même concept.

La première catégorie pose problème. Ce qui est très utile pour assister un utilisateur, va se révéler erratique pour l'automatisation, dans la mesure où l'on substituera, de façon imprévisible, au terme ambigu un terme qui n'a rien à voir avec le contexte d'utilisation.

Une troisième catégorie vient s'ajouter aux deux autres, et montre le manque d'homogénéité de la synonymie. Pour des raisons d'économie des liens hiérarchiques, certains termes désignés comme synonymes entrent en fait dans une relation hyponyme / hyperonyme avec le terme préférentiel. C'est ainsi que les termes *Aspegilic acid*, *Fusarinon* et *Nivalenol* sont considérés comme des synonymes du préférentiel *Mycotoxins*, alors que le terme *Aflatoxins* est considéré comme un spécifique de ce terme. De fait, ces quatre termes, les trois synonymes et le spécifique, sont des spécifiques de *Mycotoxins*.

6.4.3 Problèmes liés aux artefacts

Tous ces problèmes d'ambiguïté liés à la synonymie agissent de façon synergique et créent quelques difficultés pour l'identification de l'origine du bruit. Afin de réduire ces possibilités de bruits, nous avions introduit un traitement permettant de rejeter tout terme inclus dans un autre terme. Ce dispositif permettait, à partir d'un terme tel que *Sweet potatoes* (patate douce), de rejeter le terme *Sweet*, relié par synonymie au terme *Sugar confectionery* (qui dans le cas présent est une mauvaise synonymie) et le terme *Potatoes* (pommes de terre) qui est aussi un artefact dans la mesure ou une « patate douce » n'est pas une variété de « pommes de terre ». Cependant, ce traitement n'avait pas été prévu pour les synonymes. Le problème de ces bruits se révélait donc pour les synonymes (pour les raisons évoquées ci-dessus) plus crucial que pour les

préférentiels, avec en plus des problèmes liés à des ambiguïtés croisées entre synonymes et préférentiels, signalées dans l'analyse du cluster CATECHOL OXIDASE. Pour reprendre un exemple, le terme synonyme *Polyphenol oxidase* renvoie sur le terme préférentiel *Catechol oxidase*. Or ce synonyme se trouve éclaté en deux termes simples *Polyphenol* et *Oxidase,* tous les deux des préférentiels. On se trouve là devant un problème de surgénération artificielle de termes.

6.4.4 Les artefacts du cluster CATECHOL OXIDASE

Nous montrons dans cette section l'effet que peut avoir sur un cluster les bruits liés aux artefacts du thésaurus. Dans le thésaurus Agrovoc le terme *Catechol oxidase* est un préférentiel et a comme synonyme : *Polyphenol oxidase*. Dans le corpus, la forme rencontrée la plus fréquente est *Polyphenol oxidase*. En raison du traitement des termes synonymes, *Polyphenol oxidase* va renvoyer sur *Catechol oxidase* mais aussi sur *Polyphenols* (préférentiel Agrovoc) et sur *Oxidoreductases* (préférentiel Agrovoc d'Oxidases). En découlent alors les relations internes "artificielles" suivantes:

- *Catechol oxidase* et *Oxidoreductases*
- *Oxidoreductases* et *Polyphenols*
- *Oxidoreductases* et *Phenolic compounds*
- *Browning* et *Polyphenols*
- *Browning* et *Oxidoreductases*
- *Enzyme* et *Polyphenols*

et les relations externes suivantes:

- *Polyphenols* et *Purification*
- *Oxidoreductases* et *Purification*

A cela s'ajoutent des relations « semi-artificielles » qui résultent d'une part du problème cité ci-dessus et d'autre part d'une cooccurrence réelle des termes dans le corpus. Ces relations internes sont les suivantes:

- *Phenolic compounds* et *Polyphenols*

– *Enzyme* et *Oxidoreductases*

6.4.5 Solutions retenues

Les bruits (décrits ci-dessus) sont donc dépendants de facteurs linguistiques, de facteurs sémantiques liés au thésaurus, ce qui dans un grand nombre de cas génère par synergie des artefacts.

Le premier travail entrepris a constitué en une épuration *a priori* des unitermes extraits avant classification sur des critères de fréquence, d'une part, et de polysémie ou de trop grande généricité d'autre part. L'examen de la fréquence nous a conduit à rejeter les termes les plus fréquents et particulièrement ceux qui avaient servi à l'équation de recherche. Nous avons également éliminé les termes de fréquence 1, dans la mesure ou ils étaient ignorés du processus de classification.

L'expérimentation a permis de repérer certaines améliorations à apporter aux traitements linguistiques. Elles concernent certaines ambiguïtés particulières. Nous avons vu le cas du terme *Greens* qui peut être un nom ou un adjectif. Le module d'étiquetage automatique affecte par défaut la catégorie nom à ce terme. Afin d'empêcher qu'il soit reconnu en position adjectif il faudrait profiter du fait que ce nom, en tant que terme a toujours un emploi pluriel dans les textes. En contraignant le terme à n'être reconnu que sous la forme pluriel (donc avec un « s » final) on pourrait éliminer tous les emplois adjectifs erronés.

Concernant le problème particulier de la polysémie, il est apparu que la cooccurrence pouvait être une indication précieuse de levée d'ambiguïté dans la mesure où elle permettait de resituer le terme litigieux dans son environnement sémantique. C'est ainsi que le terme *Lead* se trouvait associé en permanence avec des métaux (pour la plupart des métaux lourd). Cette observation a permis de rendre compte que l'usage de ce terme dans le corpus n'était pas ambigu et qu'il convenait de le garder. Nous envisageons à l'avenir, d'utiliser systématiquement la cooccurrence pour juger de la pertinence d'unitermes potentiellement ambigus.

Il est apparu également au cours des différentes manipulations qu'il était possible d'utiliser à l'avenir la cooccurrence dans une perspective de réindexation, quand le terme apparaît non ambigu dans son contexte mais l'est dans

son usage documentaire. Prenons l'exemple de *Pulp*, nous avons vu que l'usage documentaire prévoit que ce terme réfère à l'industrie du papier. Or dans notre corpus il s'agissait bien évidemment de la « pulpe des fruits ». L'examen de la cooccurrence montre que ce terme est associé avec une grande régularité à un terme de la classe « fruit ». Cette observation permet de proposer la règle de réindexation suivante :

1 Si *le terme Pulp est associé à un terme de la classe « fruit »* **Alors** *indexer avec le terme Fruit pulp*

Le problème des artefacts est sans doute le plus difficile à traiter en raison de l'effet de synergie qui y est associé. Nous allons cependant généraliser la possibilité de bloquer la génération d'un terme inclus dans un autre (synonyme et préférentiel confondu). À l'issue de cette phase un nombre significatif d'artefacts entraînant une surgénération ou des relations erronées devrait être supprimé. Nous évaluerons alors l'impact de la génération des synonymes sur le résultat final. Deux possibilités sont envisageables :

- la suppression pure et simple de la génération de synonymes ; il s'agit d'une opération de toute façon intéressante pour des actions de structuration de connaissance et nous y aurons recours;
- le contrôle des renvois de synonymie avant chaque clusterisation, qui s'il est généralisé peut s'avérer efficace et rapide, peut permettre d'améliorer le thésaurus, et donner une indication de qualité.

7 Conclusion

Notre propos était de montrer l'intérêt du couplage de deux ensembles d'outils : linguistique informatique et infométrique, pour l'acquisition de connaissances en corpus. Cela a donné lieu à une méthodologie d'acquisition et de structuration de connaissances en corpus que nous nommons approche ILC. Au lieu de partir d'une modélisation *a priori* fondée sur la représentation mentale que les experts se font d'un domaine, l'approche ILC part directement du corpus documentaire définissant le domaine d'application, pour faire émerger à travers l'extraction terminologique (section 4) les connaissances contenues

dans le corpus. L'économie en temps de travail et d'exécution de cette approche permet d'envisager des applications à d'autres domaines de la base documentaire PASCAL.

Nous avons vu que le traitement statistique basé sur le phénomène de la cooccurrence des termes fournit un support objectif (section 5) pour que les experts réalisent la verbalisation des réseaux lexicaux (clusters) et le typage des associations de cooccurrence entre les termes (section 6). Ces opérations ajoutent une couche sémantique aux deux couches précédentes, à savoir la couche terminologique représentant les connaissances contenues dans les documents, et la couche statistique constituée par les clusters et les associations entre termes valuées par l'indice d'équivalence. Il faut noter que cet indice est analogue aux indices bien connus de Dice, de Jaccard et de Salton ou du cosinus.

Les tableaux 1, 2 et 3 de la section 6 résument la couche sémantique où les clusters construits statistiquement (d'après la méthode signalée dans la section 5) évoluent vers la représentation conceptuelle de faits de langue et deviennent ainsi des quasi-réseaux sémantiques, à cause justement du typage des associations par les experts. A ce niveau, l'emploi d'un formalisme et d'un système opérationnel de représentation des connaissances apparaît donc nécessaire pour l'automatisation de ce typage manuel. L'incorporation à la Plateforme Infométrique Linguistique d'un formalisme et d'un système de représentation de connaissances (du type KL-ONE ou CLASSIC) est le résultat escompté du projet ILC en cours de réalisation.

Partant de ces observations , la première étape de ce travail a donc été de faire une extraction des termes contenus dans les documents analysés. Ces traitements ont servi de point d'entrée aux outils infométriques (SDOC) : 40 clusters ont été obtenus et ont pu être verbalisés par les experts du domaine. La capacité des experts à pouvoir décrire succinctement les structures que sont les clusters est une validation en soi de la pertinence cognitive des clusters. Le travail d'analyse a consisté ensuite à nommer les relations de cooccurrence des termes ce qui a permis d'en établir une typologie. Enfin, nous nous sommes intéressés à un cluster particulier : CATECHOL OXYDASE. Afin de montrer que les réseaux lexicaux internes aux clusters sont des structures cognitives complexes interprétables, nous utilisons le formalisme des graphes conceptuels pour la

représentation ; cela permet d'établir que ce cluster correspond à un modèle abstrait instanciable.

A Verbalisation de quelques clusters

Cluster : CHERRY

Description : Conservation des fruits et légumes - Saumurage.

ASSOCIATIONS INTERNES :

Synonymes ou quasi-synonymes :

		Apricot et Prunus armeniaca
		Cherry et Prunus avium
		Nectarine et Prunus persica
		Peach et Prunus persica
		Cherry et Prunus cerasus

Sorte_de

Fruits à noyaux	←–	Apricot et Plum
	←–	Apricot et Peach
	←–	Peach et Plum
	←–	Nectarine et Peach
	←–	Nectarine et Plum
	←–	Cherry et Plum

Relations d'appartenance géographique

Michigan	←–	Prunus cerasus et de Cherry

ASSOCIATIONS EXTERNES :

Composant_de

Pigment	←–	Nectarine
Flavonoid	←–	Nectarine
Kernel	←–	noyau de Apricot

Relations de type prédicatif

Infestation **de**	Nectarine	**par** Cydia pomonella
Infestation **de**	Nectarine	**contrôlée par** Quarantine
Infestation **de**	Nectarine	**par** Lepidopteron
Infestation **de**	Nectarine	**par** Methyl Bromide
Fumigation **de**	Cherry	
Lésion **de**	Peach	

Cluster : LEGUME

Description : Qualité des légumineuses amylacées.

ASSOCIATIONS INTERNES :

Synonymes ou quasi-synonymmes:

		Lens culinaris et Lentil
		Arachis hypogaea et Groundnut
		Groundnut et Vigna subterranea

Sorte_de

Legume	←–	Vigna subterranea
Legume	←–	Lentil
Legume	←–	Groundnut

Composant_de

Granule	←–	Potato starch
Granule	←–	Amylose
Lentil	←–	Amylose
Legume	←–	Granule
Legume	←–	Amylose

Relations de type prédicatif

Swelling **de**	Amylose	*(sorte de propriété*
Swelling **de**	Granule	
Heat sterilization **de**	Legume	

ASSOCIATIONS EXTERNES :

Sorte_de

Legume	←–	Kidney bean
Legume	←–	Phaseolus vulgaris
Legume	←–	Chickpea
Legume	←–	Cajanus cajan
Chemicophysical property	←–	Swelling

Composant_de

Legume	←–	Granule
Granule	←–	Starch
Starch	←–	Amylose

Autre relation

		Groundnut et Millet

Cluster : CUCUMBER

Description : Fruits à noyaux.

ASSOCIATIONS INTERNES :

Synonymes ou quasi-synonymmes :

		Apricot et Prunus armeniaca
		Cucumber et Cucumis sativus

Composant_de (forme consommable de)

Cucumber	←–	Pickle

Relations de type prédicatif

traitement **de**	Cucumber	**par** Brine

ASSOCIATIONS EXTERNES :

sorte de

Légume-fruit	←–	cucumber et tomato

Composant_de (organe de)

Pericarp	←–	Cucumber

Relations de type prédicatif

Produit de la fermentation **de**	Cucumber	→ Acetic acid
Produit de la fermentation **de**	Cucumber	→ Lactic acid
Fermentation **de**	Cucumber	
Effet **du**	pH	**sur** Cucumber
Analyse **de**	Cucumber	**par** Gas chromatography

Autre relation

		Cucumber et Sodium
		Cucumber et Calcium

Cluster : BOTRYTIS CINEREA

Description : Contamination des fruits par les champignons saprophytes et pathogènes.

ASSOCIATIONS INTERNES :

Synonymes

		Apple et Malus pumila

Sorte_de

Champignons pathogènes	←–	Botrytis cinerea et Penicillium expansum
Nuts	←–	Pistachio
Pathogen	←–	Penicillium
Fungus	←–	Aspergillus
Pathogen	←–	Botrytis cinerea
Pathogen	←–	Penicillium expansum
Fungus	←–	Botrytis cinerea

Composant_de (ensemble de)

Fungus	←–	Species

Relations de type prédicatif

Production **de**	Aflatoxin	**par** Aspergillus flavus
Production **de**	Aflatoxin	**par** Fungus
Contamination **de**	pistachio	**par** Aflatoxine
Contamination **de**	Apple	**par** Penicillium expansum
Biodeterioration(contamination) **de**	Malus pumila	**par** Penicillium expansum
Biodeterioration (contamination) **de**	Malus pumila	**par** Botrytis cinerea
Stade de développement **de**	Penicillium expansum	→ Spore
Stade de développement **de**	Botrytis cinerea	→ Spore
Stade de développement **de**	Fungus	→ Spore
Stade de développement **de**	Pathogen	→ Spore

Autre relation

		Biodeterioration et Pathogen
		Biodeterioration et Spore

ASSOCIATIONS EXTERNES :

Relations de type prédicatif

Lesion **de**	Spore	
Lesion		**par** Botrytis cinerea
Pourissement		**par** Botrytis cinerea
Biodeterioration (contamination) **de**	Botrytis cinerea	**par** Fungicide

Autre relation

		Biodeterioration et Lesion

Références

[1] Bourigault (D.). – *LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquistion des connaissances à partir de textes.* – Thèse de PhD, Ecole des Hautes Etudes en Sciences Sociales, juin 1994.

[2] Bourigault (Didier). – Extraction et structuration de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. *In : RFIA.* – 1994.

[3] Brill (Eric). – *A Corpus-Based Approach to Language Learning.* – Thèse de PhD, University of Pennsylvania, 1993.

[4] Courtois (Blandine). – Un système de dictionnaires électroniques pour les mots simples du français. *Langue Française*, vol. 87, 1990, pp. 11–22.

[5] Daille (B.), Habert (B.), Jacquemin (C.) et Royauté (J.). – Empirical observation of term variation and principles for their description. *Terminology*, 1995.

[6] Daille (Béatrice). – *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* – Thèse de PhD, Université de Paris VII, Computer Communication and Vision, TALANA, février 1994.

[7] Grivel (Luc), Mutschke (Peter) et Polanco (Xavier). – Thematic mapping on bibliographic databases by cluster analysis : a description of the sdoc environment with solis. *Journal of Knowledge Organization*, vol. 22, n 2, 1995, pp. 70–77.

[8] Jacobs (P.S.). – Words, words, words: lexical representation and knowledge acquisition. *In : Proceedings of Language Engineering Convention.* pp. 75–81. – Eur. Network in Language & Speech, july 1994.

[9] Jacquemin (Christian). – Representing and parsing terms with acceptability controlled grammar. *In : Proccedings of Terminology and Knowledge Engineering*, éd. par Verlag (Indeks). – 1993.

[10] Jacquemin (Christian). – Fastr: A unification-based front-end to automatic indexing. *In : Proceedings of Information Multimedia Information Retrieval Systems and Management.* pp. 34–47. – New-York, october 1994.

[11] Jacquemin (Christian). – Recycling terms into a partial parser. *In : Proceedings of the 4th Conference on Applied Natural Language Processing.* – 1994.

[12] Jacquemin (Christian) et Royauté (Jean). – Retrieving terms and their variants in a lexicalised unification-based framework. *In : Proceedings of the 17th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1058–1063. – 1994.

[13] Polanco (X.), Royauté (Jean), Grivel (L.) et Courgey (A.). – Infométrie et linguistique informatique : une approche linguistico-infométrique au service de la veille scientifique et technologique. *In : Proceedings Les systèmes d'information élaborés.* – 1995. abstract.

[14] Polanco (Xavier) et Grivel (Luc). – Mapping knoledge : The use of co-word analysis techniques for mapping a sociology data file of four publishing countries (france, germany, uk and usa). *The International Journal of Scientometrics and Informetrics*, vol. 2, n 1, 1995, pp. 123–137.

[15] Royauté (Jean) et Jacquemin (Christian). – Indexation automatique et recherche de noms composés sous leurs différentes variations. *In : Informatique et Langue Naturelle.* IRIN, équipe "Langage Naturel", pp. 5–26. – 3 rue du Maréchal Joffre, 44041 Nantes Cedex 01, décembre 1993.

[16] Royauté (Jean), Schmitt (Laurent) et Olivetant (Eric). – Les expériences d'indexation à l'inist. *In : Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)*, pp. 1058–1063. – 1992.

[17] Salton (G.), Allan (L.) et Buckley (C.). – Automatic structuring and retrieval of large text files. *Communications of the ACM*, vol. 37, n 2, february 1994, pp. 97–108.

[18] Shieber (S.M.). – *An Introduction to Unification-Based Approaches to Grammar.* – Center for the Study of Language, Stanford, 1986.

[19] Sowa (J.F.). – *Conceptual Structures: Information Processing in Mind and Machine*. – Addison-Wesley, Reading, 1984.

[20] Stephens (Charlotte S.). – The nature of information technology research : a seven year analysis. *Journal of Computer System Information Systems*, vol. 34, n4, Summer 1994, pp. 67–76.

[21] Toussaint (Yannick). – Combining informetrics and linguistics in order to analyse large documentary databases. *In : Knowledge Based Computer Systems, Research and Applications*, éd. par Anjaneyulu (K.S.R.), Sasikumar (M.) et Ramani (S.). pp. 279–290. – Narosa Publishing House, New Delhi, 1996.

[22] Toussaint (Yannick), Royauté (Jean), Muller (Chantal) et Polanco (Xavier). – Analyse linguistique et infométrique pour l'acquisition et la structuration de connaissances. *In : Terminologie et Intelligence Artificielle*. – 1997.

Unit´e de recherche INRIA Lorraine, Technop^ole de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rh^one-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399