# Création semi-automatique d'une ontologie et des annotations sémantiques pour une liste de diffusion d'une communauté de pratique

Bassem Makni, Khaled Khelif, Rose Dieng-Kuntz et Hacène Cherfi

Projet Edelweiss, INRIA Sophia Antipolis, 2004 route des lucioles,
BP 93, 06902 Sophia Antipolis
{bassem.makni, khaled.khelif, rose.dieng, hacene.cherfi}@sophia.inria.fr

## Résumé

Cet article décrit une approche de création semi-automatique d'une ontologie et d'annotations sémantiques à partir d'informations extraites des textes d'une liste de diffusion dédiée au support informatique. L'idée d'utiliser une liste de diffusion comme corpus de référence pour cette tâche est assez originale vu la qualité des textes échangés par le biais de ce moyen de communication. Ces ressources sémantiques une fois générées permettront de créer une FAQ (Foire aux Questions) répondant aux questions fréquemment posées dans cette liste de diffusion.

**Mot clés :** TAL, traitement de messages électroniques, ontologie, annotation sémantique, CoP

## 1. Introduction

L'extraction d'informations à partir de messages électroniques (mails) n'a pas été beaucoup étudiée dans la communauté du TALN (Traitement Automatique de la Langue Naturelle). Ceci est dû principalement à la présentation informelle des messages mails et à leurs faibles apports d'informations. Cependant, les mails peuvent être parfois la seule source de connaissances pour une organisation ou une communauté de pratique (CoP). C'est le cas d'@pretic[1] qui est une association ouverte à tous les enseignants travaillant en Belgique francophone : instituteurs, professeurs du secondaire, de l'enseignement de promotion sociale et des Hautes Ecoles, exploitant les TIC avec leurs élèves, utilisant l'outil informatique pour préparer leurs leçons ou directement impliqués dans la gestion d'une Cyberécole. La communication dans cette communauté se fait essentiellement par échanges de mails sur une liste de diffusion dont le but est de traiter les problèmes rencontrés par les membres de cette CoP dans le domaine des TIC.

Dans le but de faciliter la navigation dans cette liste et la recherche de solutions pour des problèmes déjà posés, nous avons proposé une approche pour la création d'annotations sémantiques pour cette liste de diffusion, ces annotations sont basées sur une ontologie qui est elle-même extraite en partie à partir du corpus de mails existant.

Ainsi la base d'annotations créée servira pour la construction d'une FAQ structurée dans laquelle la navigation est guidée par l'ontologie en s'appuyant sur un moteur de recherche

1 Association des professeurs exploitant les TIC en Belgique francophone: http://www.apretic.be/

sémantique.

Dans ce qui suit, nous présentons l'ontologie @pretic que nous avons construite en partie à partir des textes, ensuite nous présentons des scénarios d'utilisation de cette ontologie avant de conclure et de discuter notre approche.

## 2. L'ontologie @pretic :

Pour la construction de cette ontologie nous avons opté pour une ontologie modulaire composée de 4 ontologies, chacune est dédiée à une tâche particulière. (i) une ontologie pour les composants informatiques, (ii) une ontologie pour la description des mails, (iii) une ontologie pour la description des problèmes informatiques, et (iv) une ontologie pour les membres de la CoP.

Comme le montre la figure 1, la construction du lien entre les différents modules de l'ontologie a été guidée par l'usage.

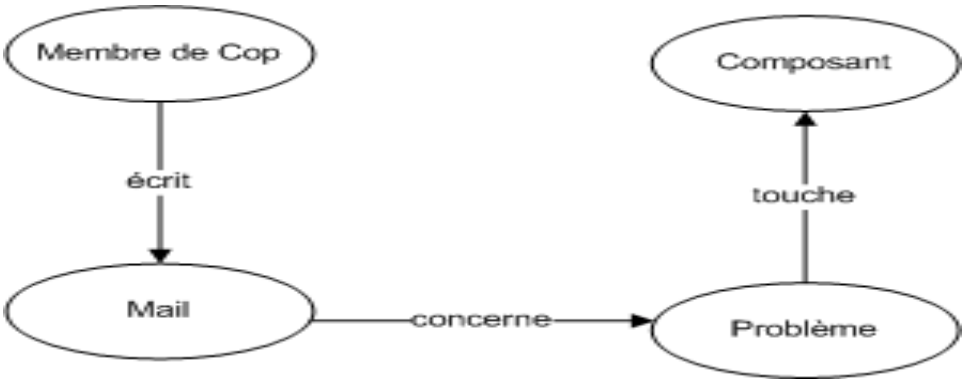


**Fig 1. : Lien entre les différentes ontologies**

### *2.1 L'ontologie des composants : De WeboPedia vers OntoPedia*

Vu que le domaine informatique est très vaste, les messages échangés dans une liste de diffusion ne peuvent pas cerner tous les concepts de ce domaine. Par ailleurs, des travaux ont été réalisés pour fournir des hiérarchies de concepts et de relations décrivant ce domaine. Un exemple de ces hiérarchies est l'encyclopédie en ligne WeboPedia[2]. Nous avons ainsi développé un script permettant de récupérer automatiquement cette encyclopédie, d'en extraire la hiérarchie ainsi que les termes (instances) et de générer une ontologie en RDFS (Mcbride, 2004) que nous avons baptisé OntoPedia.

Cette encyclopédie étant en anglais, nous avons effectué un travail de traduction semi-automatique, et ce en nous basant sur des dictionnaires en ligne, des traducteurs automatiques et des glossaires bilingues tel que le glossaire traduc[3] de la communauté de traduction des projets libres.

[2] http://www.webopedia.com/
[3] http://wiki.traduc.org/

### *2.2 L'ontologie des mails : OeMail*

Nous avons conçu l'ontologie OeMail pour annoter les métadonnées des mails. Elle comporte des concepts génériques comme EmailMessage décrivant un mail et des concepts plus spécifiques comme MimeMessage décrivant un message MIME[4]. Ces différents concepts sont reliés par des relations sémantiques concernant les métadonnées des messages (auteur, date, destinataire, etc.).

### *2.3 L'ontologie des CoP :*

Pour cette ontologie, nous avons réutilisé une partie de l'ontologie proposée dans le cadre du projet européen Palette pour la description des CoP. Cette ontologie (Vidou et al., 2006) fournit un vocabulaire structuré permettant de décrire principalement : la communauté, les acteurs de cette dernière, les compétences des membres d'une CoP et les moyens de collaboration entre ces membres.

Le choix de cette ontologie nous a permis de générer des annotations sémantiques sur les membres d'@pretic échangeant des messages sur la liste de diffusion.

### *2.4 L'ontologie des problèmes :*

Cette ontologie constitue le cœur de l'ontologie @pretic. En effet, le but de cette ontologie est de fournir des concepts et des relations permettant de décrire les problèmes rencontrés par les membres de la CoP et décrits dans les mails échangés. Les annotations fondées sur cette ontologie permettront d'alimenter au fur et à mesure la base d'annotations et d'enrichir ainsi la FAQ des problèmes. Pour construire cette ontologie, nous nous sommes basés sur le corpus de mails fourni par la CoP et nous lui avons appliqué des techniques de TALN qui nous ont aidé à amorcer l'ontologie et ensuite à l'enrichir.

Cette phase du travail est détaillée ci-dessous.

## 3. Construction de l'ontologie des problèmes :

### *3.1 Nettoyage du corpus :*

Comme nous l'avons souligné dès le début, le corpus fourni par la CoP était très dégradé ce qui nous amené à effectuer une phase de nettoyage très lourde afin d'avoir des textes d'une qualité acceptable pour le traitement par des outils de TAL. Cette phase est composée par ces différentes étapes :

- Nettoyage préliminaire : les messages d'une liste de diffusion sont généralement sauvegardés sous format d'un « dump » d'une base de données relationnelle. La première phase de nettoyage consiste alors à extraire les mails sous un format XML, supprimer les « spams » (grâce aux entêtes), supprimer les pièces jointes et restaurer les liens entre les messages d'origines avec leurs réponses (ce lien nous sera utile ensuite pour la construction de la FAQ). Ce nettoyage préliminaire est réalisé à l'aide d'un module que nous avons développé en nous basant sur la JavaMail API.

[4] Multipurpose Internet Mail Extensions

- Filtrage des signatures : d'après les standards, la signature doit être délimitée par un délimiteur standard « -- » sauf que les expéditeurs respectent peu ce standard (les signatures sont collées à la suite du corps par exemple). Pour les filtrer nous avons mis en place un algorithme de détection de signatures qui lors d'un premier passage sur le corpus compare les pieds des messages émis par chaque auteur et considère un pied qui se répète plus que trois fois comme étant une signature.
- Détection de la langue: Bien que notre corpus provienne d'une communauté francophone il était trilingue (quelques messages en anglais et en flamand). Nous avons ainsi utilisé un outil (TextCat (Canvar & Trenkle, 1994)) permettant de détecter la langue afin de ne garder que les messages écrits en français.
- Ré-accentuation : une des origines majeures de la dégradation des textes des mails est la non accentuation. Cette erreur orthographique a un impact négatif sur l'extraction des candidats termes. Pour résoudre ce problème, nous avons utilisé l'outil Reacc[5] d'accentuation automatique.
- Nettoyage répétitif : A l'opposé de corpus formels où le nettoyage fournit un corps de texte non bruité, les corps des mails contiennent beaucoup de bruit même après le nettoyage. Ce bruit n'a pas de forme générique : formes de salutations, remerciements, signatures non filtrées, etc. L'écriture de filtres pour éliminer ce bruit nécessite le parcours du corpus, la localisation des zones de dégradation et l'ajout d'un filtre pour chaque zone détectée. Or cette méthode manuelle est lente et les filtres écrits peuvent être valables seulement pour les zones examinées. Nous avons adopté une méthode de nettoyage semi-automatique dont le but est d'accélérer la détection du bruit. Dans cette approche, la chaîne de traitement automatique de la langue est bouclée par un retour sur le texte. En effet les candidats termes extraits sont parcourus pour détecter les candidats qui n'ont aucun sens ou qui sont très personnels («merci d'avance », « bonnes vacances»…). Les occurrences de ces candidats dans le texte sont utilisées pour générer de nouveaux filtres. Pour illustrer cette approche, nous avons développé un outil d'assistance de nettoyage (cf. figure 2) qui applique l'outil FASTR (Jacquemin, 1997) sur le texte en entrée, garde le lien avec le texte et permet à l'utilisateur de naviguer dans les occurrences de chaque candidat terme et de générer des filtres à partir de ces occurrences.

[5] http://rali.iro.umontreal.ca/Technologies/Reacc.fr.html

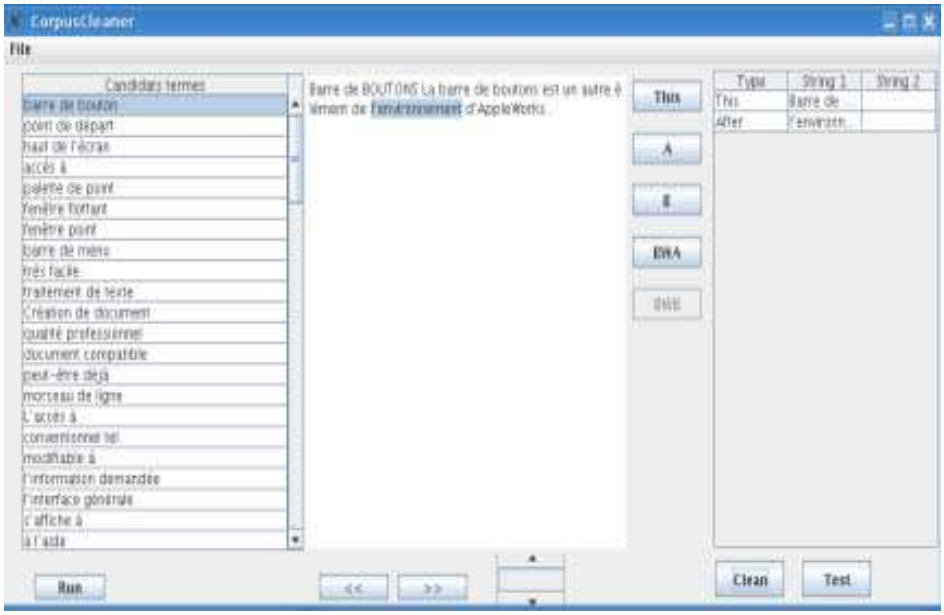


**Fig 2. : L'assistant de nettoyage du corpus**

### *3.2 Extraction des candidats termes :*

Ayant pour objectif d'extraire le maximum de candidats termes significatifs afin de construire une ontologie assez riche et couvrant la majorité des problèmes informatique, nous avons utilisé trois approches différentes de TALN, à savoir : (i) une approche syntaxique par le biais de l'outil FASTR, (ii) une approche statistique en utilisant l'outil Likes (Ouesleti et al., 1996), et (iii) une approche mixte implémentée par l'outil ACABIT (Daille, 1994).

Cependant, vu la taille moyenne du corpus (2105 messages) la deuxième approche a abouti à des résultats peu pertinents et donc nous n'avons pas pris en compte les résultats obtenus lors de cette phase. En effet, les méthodes statistiques, se basant sur des calculs de probabilité et de co-occurrence, nécessitent des corpus de taille conséquente afin de donner de bons résultats.

La Table 1 montre les résultats de ACABIT et de FASTR sur le corpus de mails.

| | |
|---|---|
| FASTR | 9665 termes |
| ACABIT | 11588 termes |
| Intersection des résultats de FASTR et ACABIT | 2198 termes |
| Union des résultats de FASTR et ACABIT | 19055 termes |

**Table 1. : Résultat des outils de TALN sur le corpus**

Le nombre de termes communs est normalement plus important mais le fait que ACABIT ne renvoie que la forme lemmatisée des mots, rend la mise en correspondance entre deux termes identiques assez difficile. Pour la construction de l'ontologie, nous nous sommes basés sur l'union des résultats qui, malgré le bruit, fournit une information plus riche sur la

terminologie du domaine.

### *3.3 Amorçage et enrichissement de l'ontologie :*

En parcourant la liste de diffusion nous avons remarqué qu'une partie des messages partagent une régularité syntaxique par rapport aux termes utilisés pour poser un probléme. Cette régularité consiste à utiliser le mot « problème » suivi du composant informatique concerné par ce problème. Cette étude nous a amenés à amorcer la construction de l'ontologie par la sélection des candidats termes, proposés par les outils de TAL, ayant comme tête le mot « problème ».

Problème de réception - Problème de port -Problème de réseau
Problème juridique - Problème technique -Problème de permission
Problème de câblage - Problème de transfert -Problème de lenteur
Problème de connexion -Problème sur le disque -Problème avec hotmail
Problème de codecs - Problème de licence - Problème de mise à jour
Problème de rapidité - Problème de windows - Problème hardware
Problème d'alimentation - Problème d'accès - Problème de display
Problème d'installation - Problème de configuration - Problème de gestion
Problème de droit - Problème d'alimentation - Problème d'émission
Problème de sécurité - Problème de tabulation -Problème logiciel
Problème matériel- Problème de communication

**Fig 3. : La liste des termes utilisés pour l'amorçage de l'ontologie**

La formalisation d'une partie de ces termes nous a permis d'avoir un premier squelette de l'ontologie qui a été validé par les formateurs de la CoP @pretic. Cependant, cet embryon d'ontologie, quoique intéressant et couvrant une bonne partie des problèmes rencontrés, est assez générique et risque d'induire une ambiguïté lors de la génération des annotations.

Afin d'enrichir notre ontologie et la rendre de plus en plus spécifique, nous avons effectué une analyse manuelle de tous les candidats termes générés par les outils de TAL (l'union des résultats de FASTR et de ACABIT). La figue 4 montre un extrait des termes extraits par les deux outils et utilisés pour l'enrichissement de l'ontologie.

"lenteur de connexion", "lenteur de la liaison", "lenteur pour scanner",
"perte de donnée", "difficulté de connexion", "mauvais embranchement",
"retard dans la réponse", "erreur d'envoi", "erreur de la ligne" "pas de connexion",
"ordinateur contamine", "stress du au matériel", "manque d'efficacité",
"manque de compétence", "panne survenant", "attaque incessante", "message infecte",
"cas de dysfonctionnement", "machine infectée", "cas d'infraction", "poubelle sans ouverture",
"limite de rentrée", "grosse faille", "message irrespectueux", "mémoire insuffisante",
"infection locale", "commande malveillante", "poste litigieux", "courrier avec retard",
"erreur d'envoi ", "erreur d'orthographe", "erreur d'adressage", etc.

**Fig 4. : Extrait de la liste des termes utilisés pour l'enrichissement de l'ontologie**

L'étude de cette liste terme nous a permis, dans une première étape, de :

- détecter de nouveau termes significatifs permettant d'enrichir directement l'ontologie ("manque de mémoire", "lenteur de connexion", etc.).
- détecter des relations de synonymie entre certains termes significatifs ("mémoire insuffisante" = "insuffisance de mémoire" ou "message infecté" = "infection de message"). Cette relation permet de raffiner l'ontologie et de faciliter le processus d'annotation.
- faire émerger des régularités structurelles dans une bonne partie des termes ("lenteur de X", "manque de X", "perte de X", "difficulté de X", "retard de X", "erreur de X", "manque de X"). avec X représentant à chaque fois un composant informatique (i.e. un terme de l'ontologie des composants).

Dans une deuxième étape, nous nous sommes inspirés du travail réalisé dans SAMOVAR[6] (Golebiowska et al., 2001), pour proposer des règles heuristiques qui vont permettre d'alimenter de façon semi-automatique l'ontologie. Ces règles vont détecter des structures prédéfinies dans le texte et enrichir l'ontologie par des candidats-termes qui n'ont pas été nécessairement détectés par les outils de TALN.

```
{ term1.string ='lenteur' }
{ term2.string ='de' }
{ term3.cpt  = 'Composant' } =>
{ term = term1+term2+term3; term.cpt = 'Problème'}
```

**Fig 5. : Exemple de règle de détection de termes : 'lenteur de *composant*'**

Cette règle est écrite en JAPE, langage proposé par la plateforme de développement d'applications linguistiques GATE (Cunningham et al., 2002) pour la représentation de grammaires pour le TALN. Ces grammaires sont appliquées sur le texte à l'aide d'un transducteur implémenté dans GATE.

La structuration finale de l'ontologie des problèmes (cf. figure 2) va reposer essentiellement sur la validation des experts.

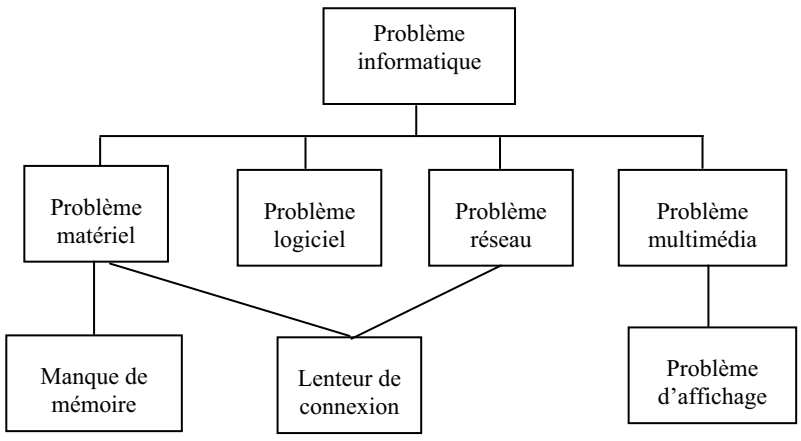


**Fig 6. : Extrait de l'ontologie des problèmes**

[6] Système d'Analyse et de Modélisation des Validations des Automobiles Renault

## 4. Exploitation des ontologies :

Comme nous l'avons déjà signalé, l'ontologie @pretic a pour but de guider la génération d'annotations sur les messages et la recherche d'informations dans la liste de diffusion, pour ensuite faciliter la construction d'une FAQ «intelligente » afin de proposer des solutions à des problèmes déjà rencontrés et résolus.

Les annotations sont générées automatiquement sur les mails existants, et ce en identifiant dans le mail (i) les termes caractérisant le problème, (ii) les composants mis en cause, (iii) les méta-données du mail et (iv) le membre de la CoP qui a posé le problème. En ce qui concerne les nouveaux mails, nous avons développé un système (cf. figure 7) et qui permet de détecter automatiquement leur arrivée et de les annoter en suivant le même processus décrit précédemment.

Une fois la base d'annotations sur les messages existants créée, elle pourra être chargée dans le moteur de recherche sémantique Corese (Corby et al., 2004). En effet, ce moteur permettra de naviguer dans la base d'annotations en tenant compte de la structure hiérarchique des ontologies et de faire des raisonnements avancées (inférences, recherche approchée, application de règles, etc.). La FAQ sera construite autour des fonctionnalités de ce système.

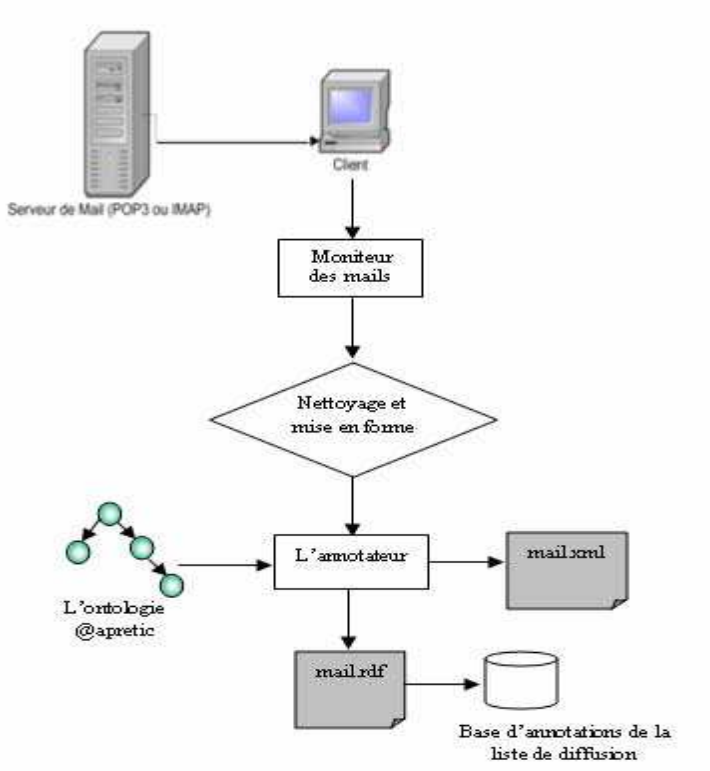


**Fig 7. : Génération à la volée d'annotations sémantiques pour les nouveaux messages**

## 5. Conclusion :

Dans cet article, nous avons présenté une approche pour la génération semi-automatique d'une ontologie et d'annotations sémantiques à partir d'une liste de diffusion de support

informatique. La finalité de ce travail est la proposition d'une FAQ sémantique profitant des avantages de l'utilisation des ontologies, des annotations sémantiques et des raisonnements offerts par un moteur de recherche sémantique (i.e. Corese).

La méthode adoptée pour la construction de l'ontologie des problèmes est généralisable pour d'autres listes de diffusion dans d'autres domaines. En effet :

- la première phase de nettoyage des messages est indépendante du domaine ;
- nous supposons que les termes utilisés pour poser un problème ne diffère pas d'un domaine à l'autre, ce qui rend la phase d'amorçage ainsi réutilisable ;
- des régularités structurelle dans les termes utilisés existent ou naissent au fur et à mesure au sein de la même communauté (i.e. notre cas et celui de Samovar), ce qui facilitera la tâche d'enrichissement de l'ontologie.

L'approche proposée s'inspire en partie de celle de (Golebiowska et al., 2001) et plus généralement de celle de (Aussenac-Gilles et al., 2001), proposées pour la construction d'ontologies à partir des textes. Deux des originalités de ce travail consistent en (i) l'utilisation d'un corpus très dégradé, à savoir le corpus de mails, et (ii) l'utilisation de trois approches/outils de TALN différents ce qui à notre avis rend les résultats plus riches.

Le coté informel des messages échangés sur la liste de diffusion a rendu les tâches de nettoyage/extraction très lourdes et différentes de celles présentées dans (Even & Enguehard, 2002). Dans ce dernier travail les auteurs ont extrait des connaissances à partir de textes dégradés mais plus au moins formel. Le nettoyage de mails a été aussi traité dans (Tang et al., 2005).

Enfin, notons que les travaux sur la construction de l'ontologie des problèmes ont été suivis et validés en grande partie par les membres de la CoP (@pretic). Ces derniers vont aussi être impliqués dans la génération semi-automatique de la FAQ qui est la perspective la plus importante pour ce travail.

## Remerciements

Ce travail est financé par le projet européen Palette (IST-2004-2.4.10). Nous remercions les membres de la communauté @pretic qui nous ont fournit le corpus de mails et avec qui nous avons eu des discussions très intéressantes.

## Références

AUSSENAC-GILLES N. , BIEBOW B., et SZULMAN S., (2000), Revisiting ontology design: a methodology based on corpus analysis, In Proc. EKAW'2000, LNAI 1937, pp. 172-188.

CANVAR W.B. et TRENKLE. J.M., (1994), N-Gram-Based Text Categorization». In Proc. of Third Annual Symposium on Document Analysis and Information Retrieval.

CORBY O., DIENG-KUNTZ R. et FARON-ZUCKER C. (2004), Querying the Semantic Web with the CORESE engine. In R. Lopez de Mantaras and L. Saitta eds, Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), IOS Press, p.705-709

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. et TABLAN V. (2002) : GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02.

DAILLE B. (1994), Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques. Thèse de doctorat en informatique. Université Paris 7.

EVEN F. et ENGUEHARD C., (2002), Extraction d'informations à partir de corpus dégradés, Actes de TALN 2002,Tome 1, Nancy, France, pp. 105-114

GOLEBIOWSKA J., DIENG-KUNTZ R., CORBY O., et MOUSSEAU D., (2001), Building and Exploiting Ontologies for an Automobile Project Memory, Kcpa'01, Canada

JACQUEMIN C. (1997), Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus, HDR, Université de Nantes, France.

MCBRIDE B. (2004), RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, http://www.w3.org/TR/rdf-schema/.

OUESLETI R., FRATH P., ROUSSELOT F., (1996). A corpus-based method for acquisition and exploitation of terms , CLIM-96, Canada

TANG J. LI H., CAO Y., et TANG Z., (2005), Email data cleaning. In Proceedings of SIGKDD'2005. August 21-24, Chicago, Illinois, USA. pp. 489-499

VIDOU G., DIENG-KUNTZ R., EL GHALI A., EVANGELOU C., GIBOIN A., TIFOUS A., (2006), S. Jacquemart: Towards an Ontology for Knowledge Management in Communities of Practice. PAKM 2006: pp. 303-314