Rapport de l'action spécifique ASSTICCOT

Action Spécifique STIC « Corpus et Terminologie » (AS 34)

Rattachée au RTP-DOC (RTP 33)

Janvier 2002 – Juillet 2003

Animatrices : N.Aussenac-Gilles (IRIT) et A.Condamines (ERSS)

Ont contribué à la rédaction de ce rapport :

AMAR Muriel
AUSSENAC-GILLES Nathalie
BIEBOW Brigitte
BISSON Gilles
BOURIGAULT Didier
BOUVERET Myriam
CANDEL Danielle
CHARLET Jean
CHAUDIRON Stéphane
CONDAMINES Anne
DAVID Sophie
DELAVIGNE Valérie
DESPRES Sylvie
DIENG-KUNTZ Rose
HOLZEM Maryvonne
KASSEL Gilles
KAYSER Daniel
LAINE-CRUZEL Sylvie
LALLICH-BOIDIN Geneviève
MINEL Jean-Luc
MOTHE Josiane
NAZARENKO Adeline
NEDELLEC Claire
NORMAND Sylvie
SZULMAN Sylvie
TANGUY Ludovic
TOUSSAINT Yannick
ZWEIGENBAUM Pierre

Rapport Interne IRIT/2003-23-R

RÉSUMÉ

L'action spécifique « terminologie et corpus » ASSTICCOT a vu le jour suite à une double motivation. Il s'agissait, d'une part, de profiter d'une dynamique autour du thème de la constitution de terminologies à partir de corpus (projets de collaboration mais aussi divers groupes de recherche pluridisciplinaires sur ce thème). D'autre part, nous voulions poser les bases d'une réflexion fondamentale permettant de répondre à une demande sociétale très importante en matière de ressources terminologiques, pour permettre une évaluation des réponses qui se font souvent au coup par coup, sans que les compétences propres aux disciplines soient clairement établies. Il a semblé urgent de dresser un état des lieux des compétences des disciplines concernées et de leurs complémentarités afin de donner une assise aux recherches et de définir les lignes forces de ce qui pourrait constituer la recherche sur ce thème dans les prochaines années.

L'action spécifique « Corpus et Terminologie » a démarré début 2002 et a déroulé son activité jusqu'en juillet 2003. Elle a fonctionné en groupe fermé, composé d'une trentaine de chercheurs d'horizons disciplinaires variés : en informatique - recherche d'information, traitement automatique des langues (TAL), apprentissage automatique et ingénierie des connaissances (IC) -; en sciences du langage - terminologie, socio-linguistique et linguistique de corpus -; et enfin en sciences de l'information. Le groupe a choisi de diffuser sa réflexion au sein d'ateliers associés à des conférences (CFD 2002 et plate-forme AFIA 2003). Il a également monté une déclaration d'intention pour organiser un réseau d'excellence européen sur ce thème. L'ensemble des activités et des documents de travail de l'AS est accessible depuis son site web : http://www.irit.fr/ASSTICCOT/ .

Afin de baliser cette réflexion, nous avons proposé une grille de questions à débattre qui s'est organisée autour de quatre thèmes : besoins ciblés, place et nature des corpus, type de ressource utilisée ou produite, méthodes et outils. Discutés, argumentés, élaborés, ces questionnements nous ont permis de dégager quatre axes de prospectives.

1 - Développer et approfondir la notion de « genre textuel »

2 - Prendre en compte les applications et usages pour comprendre la variabilité des méthodes, outils et types de ressources terminologiques

3 - Définir des méthodes pour assurer la maintenance des ressources terminologiques

4 – Se donner les moyens d'évaluer et de valider ces ressources mais aussi ces recherches.

A l'issue de cette Action Spécifique, nous énonçons des recommandations pour la poursuite de la réflexion :

- Proposer et soutenir le financement de projets d'envergure qui prennent en compte la variation dans les ressources terminologiques ;
- Donner les moyens humains, par des recrutements de jeunes chercheurs compétents, formés à cette interdisciplinarité, et les moyens financiers, par des programmes pluridisciplinaires portant sur ces thématiques ;
- Continuer à soutenir des groupes de réflexion interdisciplinaire.

TABLE DES MATIÈRES

1 INTRODUCTION

1.1 MOTIVATIONS ET HISTORIQUE

1.1.1 Identité d'ASSTICCOT

L'Action Spécifique ASSTICCOT a commencé ses activités en janvier 2002, sur l'initiative de plusieurs membres du groupe TIA (Terminologie et Intelligence Artificielle, voir ci-dessous). Leur volonté était d'animer une réflexion pluridisciplinaire autour de la thématique des ressources terminologiques et de leur lien avec des corpus de textes en la focalisant sur la dimension prospective. L'Action Spécifique a été organisée comme un groupe de réflexion réunissant régulièrement des chercheurs sollicités par les animateurs. Le mode de fonctionnement choisi a donc favorisé des réunions thématiques, impliquant tous les participants autour de questions élaborées pour décliner la question des corpus et des terminologies.

Le groupe était composé d'une trentaine de chercheurs issus de plusieurs spécialités de l'informatique (intelligence artificielle, gestion des connaissances, traitement automatique des langues, recherche d'information et apprentissage), de la linguistique (sémantique, linguistique de corpus, sociolinguistique et terminologie) et des sciences de l'information. Ces chercheurs ont été choisis parce que leurs travaux s'appuyaient déjà sur une démarche pluridisciplinaire, via des collaborations ou des compétences personnelles, et que leurs thématiques de recherche offraient un point de vue spécifique sur le thème de l'AS.

Les membres du groupe, son mode de fonctionnement ainsi que le détail de ses activités sont reportés en Annexe 1.

1.1.2 Historique

a. Groupes de travail existant

L'Action Spécifique a pu voir le jour grâce aux échanges, réflexions et travaux préalables des équipes de recherches mais aussi des groupes de travail auxquels elles participent, en particulier les groupes du GDR I3.

Ainsi, le groupe « Terminologie et Intelligence Artificielle » TIA[1], animé par A. Condamines et N. Aussenac-Gilles, réunit depuis 1993 une vingtaine de chercheurs en linguistique (linguistique de corpus et terminologie) et en informatique (intelligence artificielle, traitement automatique des langues et apprentissage pour la fouille de textes). Ils s'interrogent sur les apports mutuels de leurs travaux aux problèmes de la gestion des terminologies, bases de connaissances terminologiques et ontologies. En particulier, ils ont mis au point ou identifié des outils et méthodes utiles à la construction de ces ressources à partir de textes.

Un autre groupe a favorisé les échanges sur les problèmes de modélisation de connaissances à partir de texte : le groupe A3CTE[2], animé par C. Nedellec et A. Nazarenko. Il a organisé de 1998 à 2002 une série de séminaires rassemblant des chercheurs et industriels intéressés par l'étude des apports de l'apprentissage, des analyses statistiques et du traitement automatique des langues à l'étude sémantique des textes et à une représentation des connaissances qu'ils contiennent. Ce groupe fonctionne aujourd'hui

[1] http://www.biomath.jussieu.fr/TIA/
[2] http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/index.html

conjointement avec l'action spécifique ASAB[3]. L'objectif de la nouvelle activité A3CTE/ASAB est d'étudier et d'utiliser conjointement des méthodes issues de l'apprentissage automatique, de l'analyse des données et du traitement automatique de la langue pour l'acquisition de connaissances à partir de documents. Le domaine d'application retenu est celui de la bibliographie en génomique fonctionnelle.

b. Spécificités de la recherche française sur ce thème

L'approche française sur l'exploitation de corpus pour l'acquisition de connaissances prend appui sur une double originalité qui a favorisé le développement de l'interdisciplinarité.

D'une part, un des points singuliers de la recherche en ingénierie des connaissances est de ne pas s'intéresser seulement à l'apport de l'informatique en tant que support technique, et donc à des aspects architecturaux ou formels liés à la représentation des connaissances, mais aussi en tant qu'outil de manipulation de connaissances symboliques pour produire du sens, et donc au problème de la sémantique et de la nature des connaissances à gérer, en adéquation avec des besoins d'utilisateurs. L'ingénierie des connaissances française a donc une tradition de travail interdisciplinaire, elle interroge et se nourrit des travaux de disciplines comme l'ergonomie, la psychologie cognitive et bien sûr, pour ce qui est de l'acquisition de connaissances à partir de textes, la linguistique et la terminologie.

D'autre part, les travaux en terminologie s'appuient, en tout cas en partie, sur une tradition d'analyse de discours dite « à la française ». Ce point de vue a permis à la linguistique de s'interroger, dès les années 70, sur le rôle des outils dans l'analyse de corpus et, plus récemment, de participer à l'élaboration d'une terminologie textuelle, c'est-à-dire d'une terminologie qui s'ancre dans la réalité des usages plutôt qu'une terminologie normalisatrice.

Ces deux spécificités françaises, tout à fait complémentaires, ont renforcé la motivation de mettre sur pieds ASSTICCOT, afin de poursuivre la réflexion pour mieux se situer au niveau européen d'une part, et d'envisager une ouverture vers d'autres disciplines d'autre part.

c. Un noyau de collaborations bilatérales

La réflexion menée au sein de l'action spécifique a pu être possible grâce aux acquis des chercheurs et des différentes équipes participantes. En effet, nous avons commencé par échanger et mettre en commun des expériences ayant fait l'objet de collaborations entre au moins deux disciplines concernées par le sujet. Pour chacune de ces expériences, nous avons dégagé des points forts, des difficultés non résolues et des perspectives d'approfondissement. Parmi les collaborations bilatérales, qui réfèrent souvent à des projets pluridisciplinaires, nous avons identifié trois grands terrains d'étude :

- *Les échanges entre terminologie textuelle et TAL sur le lien entre analyse de corpus et terminologies.* Ces échanges sont de deux types :
  - Mise au point et utilisation de logiciels de TAL pour la construction de ressources terminologiques : il est important de souligner que ces collaborations vont au-delà de la situation d'utilisation « passive » de logiciels de TAL pour construire des terminologies. Elles débouchent sur la spécification conjointe de nouveaux systèmes mieux adaptés à l'analyse du terminologue. Par exemple, le logiciel Syntex permet l'extraction de termes et leur mise en réseau syntaxique. Ce réseau fait l'objet d'une analyse distributionnelle dans le système couplé à Syntex, Upery, qui guide ainsi l'identification de classes sémantiques et de relations entre classes (Bourigault & Fabre, 2000). Ces logiciels ainsi que l'interface de consultation de leurs résultats ont été spécifiés en tenant compte des expériences de leur utilisation pour la structuration de terminologies à partir de textes. De même, le logiciel Caméléon a été conçu en collaboration entre des informaticiens et des linguistes pour assister l'extraction de relations sémantiques à l'aide de patrons lexico-

[3] http://asab.limbio-paris13.org/

sémantiques (Aussenac-Gilles & Séguéla, 2000). Ces échanges renouvellent également les méthodes des praticiens et les fondements des éléments constituant les ressources, ce qui donne une validité et un statut plus forts à ces ressources.

- Données terminologiques comme ressources pour le TAL : De nombreuses applications de TAL sont d'autant plus performantes qu'elles peuvent utiliser des données lexicales ou terminologiques (cf.3.2.2) comme des thésaurus. Ainsi, certains thésaurus permettent de construire des patrons de recherche intégrant des connaissances du domaine en extraction d'information. Des lexiques permettent de lever des ambiguïtés au moment d'analyser des textes. Ou encore des lexiques associant des catégories sémantiques à des termes servent de jeu de données de référence à des systèmes d'apprentissage. Un exemple de ce type de contribution est donné par le projet ONCODOC (Seroussi *et al.*, 2000) où un système de gestion de dossiers patients s'appuie sur un thésaurus pour caractériser le contenu de ces dossiers, permettre de les archiver et de les retrouver. Plusieurs thésaurus du domaine médical ont d'ailleurs été évalués dans le projet MENELAS pour divers types de traitements des textes médicaux, en vue de leur représentation structurée (Charlet, 2002).

- *Les collaborations en ingénierie des connaissances et recherche d'information autour des ontologies* font l'objet de nombreux projets depuis environ 1997.
  - Un rapport d'un projet européen Euréka IKF, rédigé par le LADSEB, fait un état complet de plusieurs de ces projets en 2001 (Masolo, 2001). Il en ressort que l'utilisation d'ontologies est porteuse de nombreux espoirs pour réduire le silence, trouver plus de documents pertinents et trouver des réponses approchées à des requêtes sans réponse, mais que la mise en œuvre technique de solutions efficaces est loin d'être acquise (coût et qualité de l'ontologie, nature des relations exploitées pour étendre les requêtes ou représenter le contenu des documents, etc.). Plusieurs équipes de l'Action Spécifique ont elle-même mené des expériences de ce type.
  - Hiérarchie de termes pour la classification de documents : au sein du système DocCUBE, plusieurs hiérarchies de termes constituent autant de dimensions permettant d'analyser une collection de documents d'un domaine, de former des classes et de favoriser une consultation selon des points de vue. Le système permet de sélectionner 2 ou 3 hiérarchies (points de vue) et de visualiser les documents de la collection selon une présentation en 3D qui met en évidence les documents associés à l'intersection de chaque niveau et selon les différents points de vue.
  - Ontologies pour la reformulation de requêtes : Plusieurs études ont montré les limites d'une approche basée sur une ontologie générale dont on exploite les relations systématiquement : dans ce cas, les ontologies n'améliorent pas les performances d'un système de recherche d'information. Une expérience proposant une méthode plus précise a été menée. Wordnet est utilisé comme lexique général servant à étendre les requêtes selon un principe d'expansion dite « prudente », tenant compte de la nature des relations et favorisant les syntagmes par rapport aux mots simples. Des améliorations du rappel sont alors constatées (Baziz *et al.*, 2003).
  - Ontologies pour l'interrogation de données semi-structurées : l'exemple du projet PICSEL (Reynaud *et al.*, 2002), même s'il n'a pas impliqué des participants de l'AS, est tout à fait illustratif des apports potentiels d'une ontologie formelle pour guider la consultation de bases de données hétérogènes sur le web. Une ontologie unique (dans le domaine du tourisme) sert de langage pivot pour interroger plusieurs sites à l'aide de médiateurs. La formalisation de l'ontologie permet de guider la formulation d'une requête, de l'étendre à d'autres objets proches,

- *Collaboration entre terminologie textuelle, sciences de l'information, IC et TAL pour définir des outils de construction de ressources terminologiques*
  - Pour illustrer ce type de collaboration, mentionnons un projet qui cherche à outiller la mise au point d'un Index d'un livre (dans le domaine de l'ingénierie des connaissances) à partir de son analyse par divers outils de TAL, et d'une représentation structurée de son contenu, par exemple sous la forme d'une ontologie ou d'une terminologie. Ce projet rassemble des chercheurs de TAL,

de l'ingénierie des connaissances et de sciences de l'information et va déboucher sur une plate-forme intégrant ces outils pour guider l'analyste dans sa description et structuration d'un index à partir d'un texte (Aït el Mekki & Nazarenko, 2002).

Enfin, des expériences partagées dans des groupes de réflexions pluridisciplinaires ont permis aux différents participants de bénéficier, avant même le démarrage de l'AS, d'une culture commune. Bien sûr, ces groupes existants ne couvraient que partiellement les disciplines réunies dans l'AS, ce qui justifiait de mettre en place un nouveau lieu d'échange d'expériences pour parvenir ensemble à dresser des perspectives.

- Collaborations entre recherche d'information, IC et sciences de l'information : le réseau Rhône Alpes animé par l'ENSSIB, regroupe une dizaine de laboratoires Lyonnais et Grenoblois en informatique et science de l'information autour des problématiques de la recherche d'information et de la gestion documentaire. Ce réseau a permis des rapprochements très originaux (par rapport aux pratiques francophones) entre des communautés exprimant des besoins en matière de gestion de document, d'indexation, de construction et de maintenance de langages documentaires d'une part, et des communautés d'informaticiens cherchant à réaliser des solutions performantes et ayant du sens pour leurs utilisateurs. La structuration des connaissances et leur lien avec les textes, entre autres, est au cœur de ces réflexions, alimentées par des séminaires, des conférences et des groupes de travail.

A partir de la mise en commun de ces expériences de collaboration ainsi que sur la base des acquis des groupes mentionnés ci-dessus, nous avons voulu mieux identifier les objets et méthodes de recherche, les problèmes théoriques et pratiques abordés. L'objectif était de repérer les convergences possibles et les complémentarités des différentes disciplines. Il nous a donc fallu trouver les moyens de passer d'expériences ponctuelles à une théorisation des problèmes pour aller vers une vraie approche pluridisciplinaire.

Nous avons choisi d'identifier tout d'abord différents aspects de l'analyse de corpus, de la définition et de la structuration de connaissances terminologiques et de leur utilisation. Ainsi, plusieurs thèmes ont été retenus, présentés dans la partie suivante. Pour chacun d'eux, il nous a semblé indispensable que chaque discipline expose ses approches, leurs points forts et leurs limites, comment des complémentarités se dessinent avec d'autres approches et les directions indispensables à poursuivre. Nous avons donc élaboré une « grille » de réflexion comme cadre de présentation des travaux mono-disciplinaires. Cette grille, développée en annexe, a permis de mettre en forme les analyses disciplinaires développées en partie 3 de ce rapport. Nous justifions et détaillons dans le début de la partie 3 les questions retenues pour baliser la réflexion.

1.2 JUSTIFICATION SCIENTIFIQUE

La mise à disposition de corpus spécialisés sous format électronique ainsi que la demande sociale en lien avec le traitement de ces données textuelles a fait émerger un champ de recherches nouveau, visant à modéliser le contenu de ces corpus pour permettre un meilleur accès à la connaissance qu'ils contiennent. Un mouvement symétrique, partant des besoins plus nombreux en structures terminologiques ou conceptuelles (que nous appellerons ressources - ou produits - terminologiques ou ontologiques, RTO), a conduit à fédérer ces recherches de manière à rendre plus rapide la mise au point de ces structures de données et plus pertinent leur contenu. Plusieurs disciplines, dont le matériau d'étude est constitué pour l'essentiel soit de textes, soit de représentations lexicales ou conceptuelles, se retrouvent dans cette problématique:

- la linguistique de corpus, qui s'intéresse au problème du sens en lien avec un contexte, question incontournable lorsque l'étude du fonctionnement est faite sur corpus ;

- la terminologie, qui, depuis toujours, élabore des données lexicales à visée applicative, à partir des savoirs de spécialistes et plus récemment de données textuelles ;
- les sciences de l'information et la recherche d'information, qui ont une problématique à la fois de repérage terminologique pour la définition de langages documentaires ou le repérage de mots-clés, et d'indexation ou de caractérisation, qui suppose aussi des traitements automatiques des textes et l'étude de l'association entre textes et index ou mots-clés;
- le traitement automatique des langues (TAL), qui, rapidement au cours de son histoire, a compris la nécessité de circonscrire son champ d'application à des corpus de sous-langages pour plus d'efficacité et qui se pose à la fois en utilisateur et en producteur de ressources terminologiques ;
- l'ingénierie des connaissances, dans laquelle se développe un courant important qui vise la constitution d'ontologies à partir de textes ; confrontée à de nouvelles applications associant connaissances et textes, elle a pris conscience récemment de l'intérêt d'enrichir les représentations conceptuelles d'informations lexicales.

De manière générale, le thème qui fédère ces disciplines concerne la constitution de produits terminologiques en lien avec différents types d'applications : élaboration d'ontologies ou de thésaurus, recherche d'information, indexation ou annotation de documents, extraction de connaissances de textes, aide à la traduction, aide à la rédaction, aide à l'utilisation ou à la consultation de documents …

Point de départ et objectif assez similaires ne suffisent pas pour que l'interdisciplinarité s'organise. Il convient de travailler sur les histoires, les présupposés théoriques et les méthodes, afin de mieux évaluer les complémentarités qui peuvent exister entre les disciplines concernées et proposer des approches efficaces car adaptées aux besoins identifiés. La mise en commun de ces « cultures » disciplinaires a été une des préoccupations de l'action spécifique.

Les groupes TIA et A3CTE mentionnés plus haut avaient déjà effectué un premier travail de balisage de l'interdisciplinarité. Mais, d'une part, toutes les disciplines évoquées ci-dessus n'étaient pas représentées, et d'autre part, le projet d'ASSTICCOT avait une ambition d'une autre nature. Il s'agissait de faire l'état des connaissances dans les différentes disciplines pour mieux identifier les orientations intéressantes, les difficultés communes ou propres à chacune, les pistes à explorer dans l'avenir.

1.3 JUSTIFICATION TECHNOLOGIQUE

A la suite de l'utilisation généralisée des outils de bureautique, de l'internationalisation des échanges et du développement d'Internet, la production de documents sous forme électronique s'accélère sans cesse. Or pour produire, diffuser, rechercher, exploiter et traduire ces documents, pour les utiliser au sein d'applications de veille, de gestion des connaissances, de recherche d'information ou d'aide à la décision, les systèmes de gestion de l'information ont besoin de ressources terminologiques. Ces ressources décrivent les termes et les concepts du domaine, selon un mode propre au type de traitement effectué par le système. La gamme des ressources à base terminologique et ontologique est aussi large que celle des systèmes de traitement de l'information utilisés dans les entreprises et dans les institutions.

Considérer conjointement des ressources terminologiques et les textes dont elles sont issues ou auxquels elles donnent accès est une question cruciale dans les nombreux contextes où se trouvent manipulés des documents électroniques. Nous présentons dans la partie 2 plusieurs études de cas qui illustrent la diversité des problèmes posés et des applications développées, que ce soit pour organiser, retrouver ou partager des connaissances à partir de textes. Certaines de ces applications se situent dans un cadre délimité comme celui de l'entreprise, et renvoient à la problématique de la gestion des connaissances d'entreprises, d'autres dans le cadre ouvert de l'internet, rejoignant la problématique du Web Sémantique. Ces deux problématiques font aujourd'hui l'objet d'une multitude de projets et de recherches. Face à des propositions opportunistes et à des promesses pas toujours réalistes, une approche

scientifique rigoureuse portant sur les fondements qui peuvent asseoir des solutions pertinentes est un pré-requis indispensable.

Au cœur de la mise au point de ces applications se trouvent deux composantes directement liées à la langue : les corpus, qui sont les sources des connaissances ou les fonds à explorer pour l'application; les ressources terminologiques (thésaurus, terminologies, ontologies …), qui sont des représentations des connaissances associées à la langue permettant à l'application de répondre aux besoins. Le fait de construire ces ressources à partir de corpus puis de les utiliser ensuite pour accéder à des connaissances dans ces textes ou d'autres documents revient à leur faire jouer un rôle pivot. Leur qualité est donc déterminante pour garantir l'usage de l'application. Leur lien étroit avec les textes et l'application ressort également.

1.4 Présentation du rapport d'activité

1.4.1 Plan du rapport

La suite de ce rapport s'organise en 4 parties.

Dans la partie 2, nous illustrons tout d'abord les enjeux de société liés à la définition et à l'utilisation de ressources terminologiques en lien avec des documents et des corpus. Pour cela, nous rapportons quelques études de cas qui nous semblent représentatives. Ensuite, différents points de la problématique « terminologie et corpus » sont précisés.

La partie 3 constitue un reflet du cœur du travail des membres de l'AS. Cette partie présente les quatre thèmes de réflexion retenus pour baliser la problématique : besoins, corpus, types de ressources, méthodes et outils. Nous rapportons tout d'abord l'état initial de la réflexion autour de ces quatre questions. Ensuite, chaque partie présente les développements proposés par chaque groupe disciplinaire en réponse à ces questionnements.

La partie 4 rend compte du fruit de cette réflexion en matière de prospectives. Il s'est dégagé des discussions et échanges quatre orientations : (1) une meilleure caractérisation des genres textuels ; la nécessité d'une étude approfondie de la variation avec, d'une part, (2) une meilleure maîtrise des liens entre type d'application, types de ressources terminologiques, méthodes et outils pour les concevoir et d'autre part, (3) la recherche de véritables solutions au problème de la maintenance de ces ressources et de ces applications ; enfin, (4) se donner les moyens de véritables évaluations, autant des recherches que des ressources, des outils ou des méthodes.

Pour finir, dans la partie 5, une synthèse du document permet d'avoir une vue synthétique des travaux d'ASSTICCOT.

L'ensemble des documents concernant l'AS (transparents, comptes rendus de réunions, documents divers) sont disponibles sur le site web de l'AS (http://www.irit.fr/ASSTICCOT).

2 ENJEUX DE SOCIÉTÉ ET PROBLÉMATIQUE

Avant d'aborder précisément la nature de ces ressources, leurs spécificités en lien avec les disciplines qui les produisent ou les utilisent, nous illustrons la diversité des applications qui peuvent les utiliser, et la variété de la demande sociale concernant ces travaux, par quelques études de cas. Nous les avons choisies parce qu'elles présentent chacune une facette particulière de l'enjeux qu'il y a à accéder aux connaissances dans des textes et des modalités possibles pour y parvenir.

2.1 ÉTUDES DE CAS

2.1.1 Web sémantique d'une entreprise multinationale

1. Les mémoires d'entreprise : des systèmes d'information dédiés à la gestion des connaissances

Une "mémoire d'entreprise" est la matérialisation explicite et persistante des connaissances et informations cruciales de l'entreprise, pour faciliter leur accès, partage et réutilisation par les membres de l'entreprise, dans leurs tâches individuelles et collectives (Dieng-Kuntz *et al.*, 2001). De telles mémoires sont conçues, en pratique, pour différents types d'organisations. On parle ainsi de mémoire de projet, de mémoire d'un département de R&D ou encore de mémoire d'entreprise étendue (à ses fournisseurs et sous-traitants). Différents objectifs motivent leur développement, qui soulignent l'existence d'enjeux socio-économiques importants, notamment : améliorer la circulation des informations au sein de l'organisation, aider à l'intégration de nouveaux employés, accélérer le rythme de l'innovation, permettre de planifier les recrutements.

2. Une gestion de la mémoire guidée par les ontologies

Pour la gestion de ces mémoires (leur création, la diffusion de leur contenu, leur maintenance), différentes ressources sémantiques s'avèrent utiles, parmi lesquelles les ontologies. La mémoire contenant notamment la documentation de l'organisation, gérer cette mémoire passe par une meilleure valorisation de la documentation. Pour ce faire, il s'agit de décrire le contenu des documents et leur contexte de création et d'utilisation. La démarche revient alors à élaborer une spécification des (principaux) concepts contenus dans ces ressources, c'est-à-dire une ontologie. L'ontologie construite peut jouer plusieurs rôles :

- Permettre aux utilisateurs d'accéder à une description non ambiguë du contenu de la mémoire (O'Leary, 1998)(Fortier & Kassel, 2003).
- Servir de référence pour indexer/annoter sémantiquement la mémoire à des fins d'amélioration de la recherche de documents ou d'informations contenues dans les documents (Staab & Maedche, 2001).
- Permettre aux utilisateurs d'exprimer leurs centres d'intérêts de façon à recevoir des informations ciblées (Domingue & Motta, 2000).

Dans le cas d'entreprises internationales, les requêtes, de même que les documents, sont multilingues. Pour permettre d'assurer les services que l'on vient de voir, l'approche courante consiste à coupler l'ontologie à des lexiques dans les différentes langues représentées (Roussey *et al.*, 2001).

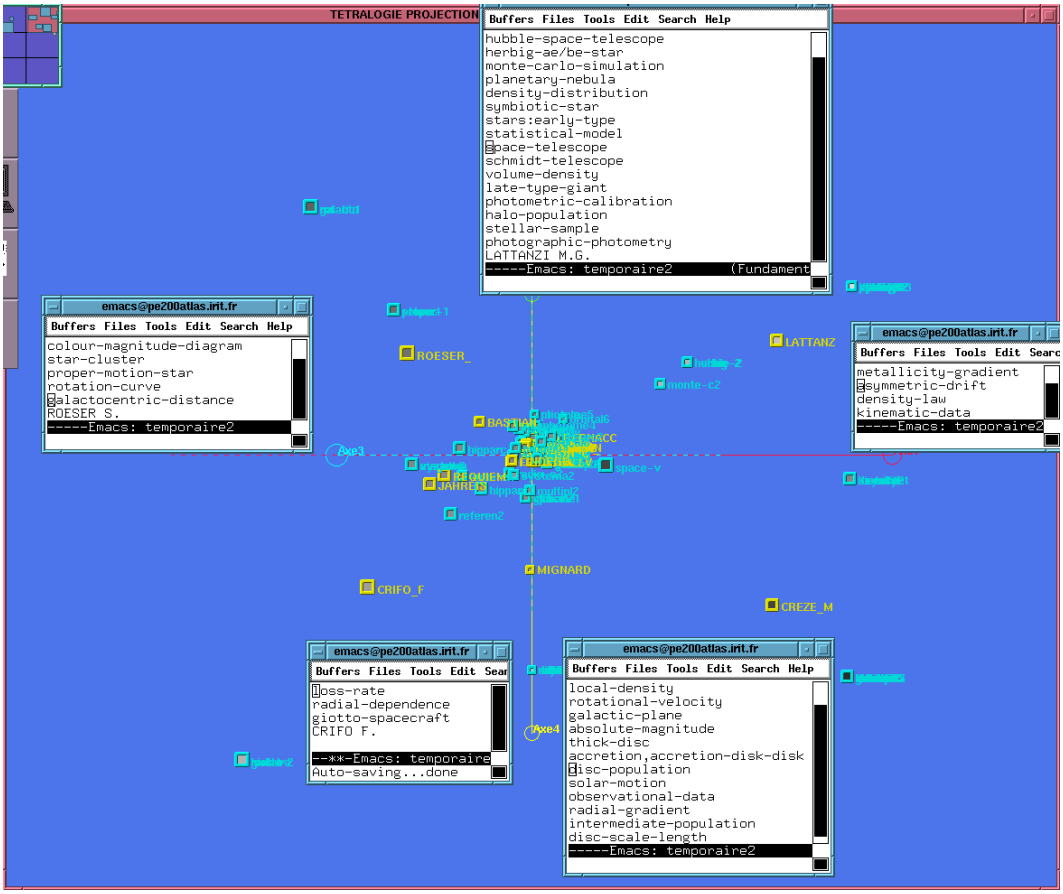
3. Questions ouvertes

Une ingénierie du développement des mémoires d'entreprise passe donc par la maîtrise de la construction d'ontologies. Or, aujourd'hui encore, plusieurs questions touchant aux fondements de la construction et de la maintenance des ontologies se posent, notamment :

- Comment articuler l'analyse textuelle des documents avec la réutilisation d'ontologies existantes (génériques, d'entreprise ou de domaines), de façon à minimiser les coûts de construction ?
- Comment faire évoluer une ontologie pour tenir compte de l'évolution de la documentation de l'entreprise ou pour prendre en compte des ressources extérieures à l'entreprise ? Comment mettre à jour les index et annotations élaborés à partir d'une ancienne version de l'ontologie ?
- Comment permettre à des utilisateurs, ou des groupes d'utilisateurs, d'adapter une ontologie d'entreprise pour tenir compte de besoins spécifiques ?

2.1.2 Fouille de textes pour la découverte de connaissances - Application à la veille

La veille est une activité cruciale dans le monde concurrentiel actuel. Les entreprises doivent surveiller les activités de leurs concurrents, collecter et analyser des informations sur le marché, les technologies et les actions gouvernementales pour définir leur politique d'alliance, d'innovation et leurs stratégies vis à vis de leurs clients. La fouille de textes peut les aider dans cette tâche. Cette activité implique de nombreuses opérations telles que la collecte d'un sous-ensemble ciblé de ressources ou documents, l'homogénéisation, la structuration et la représentation des informations filtrées en vue de leur analyse, la définition de méthodes d'analyse et d'exploration d'information, des principes d'interaction permettant la découverte et l'exploitation des éléments utiles, enfin une présentation des résultats permettant leur interprétation, soit par un acteur humain, soit par un agent.



Tétralogie, système développé à l'IRIT et diffusé dans diverses organisations, intègre ces différentes étapes pour aider les décideurs via l'analyse de sources diverses (brevets, Web, publications scientifiques, etc.). La copie d'écran ci-dessus illustre le type de visualisation permettant de faire ressortir des termes clés des documents, leurs proximités et les documents traitant de ces thèmes.

Par exemple, une analyse basée sur Tétralogie a permis à Idelux d'offrir une analyse du domaine des transports en Europe : les principaux transporteurs, les pays les mieux desservis, ceux pour lesquels l'offre de transports devrait être développée, le type de moyen de transport, en fonction des pays, des sociétés de transports et des produits transportés. Ce type d'information est extrait grâce à l'analyse de corrélations entre éléments (produits, pays, transporteurs, moyens de transport, etc.) et en utilisant des méthodes de classification et des analyses factorielles. Dans le cas d'informations datées, l'analyse de l'évolution des corrélations fournit des éléments stratégiques sur les tendances, les marchés potentiels et ceux à abandonner. Dans le cas d'informations géo-référencées, les outils de visualisation sur des cartes géographiques proposent des représentations faciles à exploiter (Hubert *et al.*, 2001), (Mothe *et al.*, 2001).

2.1.3 Repérage d'interactions entre gènes

1. Les interactions géniques dans la littérature scientifique

La modélisation des interactions géniques présente un intérêt scientifique considérable pour les biologistes car elle constitue une étape fondamentale dans la compréhension du fonctionnement cellulaire. Aujourd'hui, la majeure partie de la connaissance biologique sur les interactions n'est pas décrite dans des banques de données mais uniquement sous la forme d'articles scientifiques dans des bases de données bibliographiques accessibles en ligne telle que MedLine. L'exploitation de ces articles

est donc un enjeu central dans la construction des modèles d'interaction entre gènes. Les projets de génomique ont en effet généré de nouvelles approches expérimentales telles que les puces à ADN, à l'échelle globale de l'organisme étudié, et aujourd'hui, une équipe de recherche est capable de produire très vite des dizaines de milliers de mesures. Ce contexte très nouveau pour les biologistes impose un recours à l'extraction automatique de connaissances textuelles : pour être capable d'interpréter et de donner un sens à ces données élémentaires du laboratoire, il faut les relier à la littérature scientifique.

2. Une extraction d'information guidée par les connaissances

Dans ce cadre, si la recherche documentaire à l'aide de mots-clefs offre des performances intéressantes en termes de rapidité de traitement, ses résultats ne sont pas directement exploitables et nécessitent un travail d'analyse considérable des documents sélectionnés pour extraire l'information pertinente du fait du volume de documents concernés, de plusieurs milliers à plusieurs centaines de milliers de documents.

On constate en effet que les approches d'extraction automatiques appliquées jusque-là sont basées essentiellement sur des comptages statistiques de co-occurrences de mots-clefs (Stapley & Benoit, 2000), (Pillet, 2000), (Nédellec *et al.*, 2001) ou sur des règles ou automates d'extraction définis manuellement, à base de verbes significatifs et de noms de gènes (Blaschke *et al.*, 1999), (Thomas *et al.*, 2000), (Poibeau, 2001). Les résultats obtenus par ces approches présentent, soit une précision très faible, soit une couverture limitée (Nédellec, 2002).

L'extraction automatique de connaissances pertinentes dans les documents sélectionnés nécessite donc la mise en œuvre de méthodes d'extraction d'information plus complexes qui s'appuient sur des ressources spécifiques au domaine étudié, de types lexical, syntaxique et sémantique comme des ontologies. Ces ressources spécialisées sont généralement difficiles et longues à acquérir manuellement (Riloff, 1993), (Soderland, 1999). L'aspect novateur de l'approche proposée dans le projet Caderige (http://caderige.imag.fr) réside dans la définition et dans l'implémentation de techniques informatiques originales permettant leur acquisition automatique ou semi-automatique à partir de corpus textuels.

Le domaine de la génomique fonctionnelle, de par son homogénéité, permet de mettre en œuvre cette approche. Les articles sont écrits dans une langue de *spécialité*, correcte d'un point de vue grammatical et dont le vocabulaire est relativement limité. Les informations recherchées sont locales (exprimées sur quelques lignes au plus) et les critères de pertinence des biologistes sont précis. Comme l'a mis en effet en évidence Harris (Harris *et al.* 89) en immunologie, la variabilité des sous-langages utilisés dans les domaines de recherche spécifiques est limitée à la fois du point de vue du vocabulaire, de la polysémie, des formes syntaxiques et du nombre de concepts représentés. Dans ces conditions, il est réaliste de vouloir acquérir semi-automatiquement les ressources lexicales, sémantiques et conceptuelles nécessaires à une analyse profonde, ceci à partir des régularités observées dans un corpus (Staab *et al.*, 2000). Par exemple, voir Figure 1.b un extrait d'ontologie telle qu'elle peut être apprise par une méthode de sémantique distributionnelle Figure 1.a.

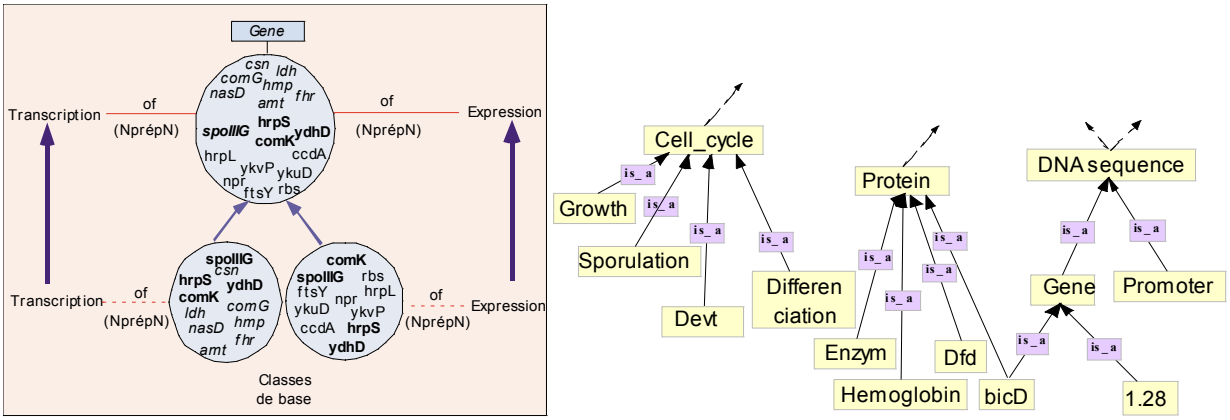


Figure 1. Exemple d'apprentissage d'ontologie en génomique

Ensuite, grâce à ces ressources, par l'intermédiaire de transformations, on peut ramener toute phrase d'un corpus de spécialité à une forme canonique qui représente son contenu informationnel en termes d'opérateurs (ou prédicats) et d'arguments (Bessières *et al.*, 2001) (Figure 2).

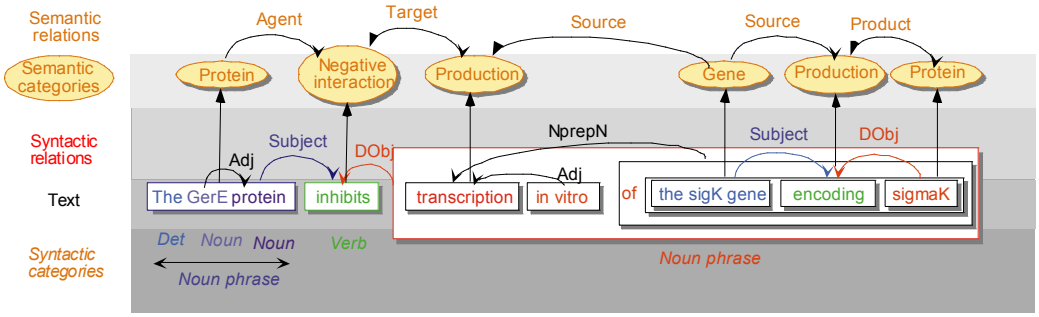


Figure 2. Exemple de normalisation en génomique à l'aide de ressources terminologiques et ontologiques

Cette représentation facilite l'apprentissage, l'exploitation et la maintenance de règles d'extraction concises, générales et peu nombreuses (Figure 3)

Règle d'extraction
interaction_agent(X,Y):-
cat(X, GN), concept(X, protein), sujet(X,Y), cat(Y,verbe), concept(Y, interaction), cat(Z, NP), COD(Z,Y), concept(Z, gene-expression)

Texte
[..] the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK [..]

Formulaire rempli

Interaction	Type : negative Agent : GerE protein		
	Target:	Expression	Source : sigK gene Product : sigmaK

Figure 3. Exemple de règle d'extraction d'information et d'application en génomique fonctionnelle.

Dans le cadre proposé, il faut souligner qu'une approche multidisciplinaire est nécessaire, qui fait collaborer étroitement des biologistes avec des informaticiens respectivement spécialistes du traitement automatique de la langue, de l'extraction d'information et de l'apprentissage automatique.

2.2 PROBLÉMATIQUE

La présentation de ces cas met en lumière la diversité des utilisations possibles des ressources terminologiques spécialisées. C'est en prenant appui sur cette diversité que nous avons cherché à définir des lignes de force autour desquelles s'organise la problématique.

- La connaissance lexico-sémantique pertinente pour tous les types d'utilisation doit être soit construite, soit adaptée à partir des données disponibles (dictionnaires, terminologies, listes de mots-clés, marqueurs linguistiques …) pour convenir à chaque (type d') application.
- Une réflexion doit être menée sur l'intégration des sources de connaissances hétérogènes. En plus de l'analyse de corpus, considérée comme centrale, il faut prendre en compte les données spécialisées déjà construites mais aussi les données plus générales qui contiennent des connaissances morphologiques ou syntaxiques pouvant être pertinentes dans une application donnée. Enfin, la connaissance de l'expert constitue une autre source de connaissances. Il faut donc s'interroger sur la mise en œuvre complémentaire de ces données hétérogènes par leur nature (lexicales vs textuelles) et leur contenu (spécialisé vs général) qui reflètent des points de vues éventuellement différents, particuliers ou consensuels.
- Une des façons de garantir l'homogénéité des données consiste à ancrer les ressources spécialisées construites dans la réalité des usages textuels ; c'est aussi une façon de documenter la méthode de constitution de ces données, et les données elles-mêmes.
- La dimension multilingue ne doit pas être négligée. Bien qu'élaborées en parallèle sur des langues différentes, les données terminologiques doivent pouvoir être utilisées sur des projets multilingues ; il faut donc là aussi garantir la cohérence de ces données et évaluer la réutilisabilité des méthodes d'une langue à l'autre.
- Les différents points de vue d'utilisation des ressources doivent être pris en compte mais aussi les différents points de vue de constitution de ces données. L'analyse de la complémentarité de ces différentes approches est un point crucial.
- L'accès au contenu d'un texte peut être vu comme un processus cyclique. Un première mouvement, qui va du texte vers les ressources, vise la construction de bases de connaissances linguistiques et conceptuelles ; un second, inverse des ressources vers les textes, concerne l'exploitation de ces connaissances pour analyser le contenu des textes, les indexer, les retrouver ou y rechercher des informations. Les bases de connaissances peuvent être mises à jour en fonction de l'évolution de l'utilisation : la connaissance acquise à une étape est réinjectée dans l'étape suivante, etc.

Ces différents points posent les bases d'une réflexion pluridisciplinaire qui permettent de définir ce qui, dans les méthodes proposées, pourrait être généralisable et donc éventuellement réutilisable.

3 Une réflexion pluridisciplinaire sur le thème « Terminologies et corpus »

3.1 Thèmes retenus pour baliser la réflexion

Notre projet visait d'abord à formuler et à explorer toutes les problématiques abordées dès que l'on cherche à rendre compte du contenu de textes et à organiser les connaissances qui peuvent en être tirées dans des structures plus ou moins formelles, que nous avons appelées «ressources terminologiques ». Cette première vision envisagée est donc clairement ascendante puisqu'elle part d'un matériau textuel pour élaborer des modèles. Ce projet préliminaire a été enrichi après les premiers échanges au sein du groupe pour intégrer aussi les problématiques liées à l'utilisation de ces mêmes ressources, leur usage ayant finalement un impact sur leur nature, leur contenu et leur support. Cette vision est la symétrique de la précédente et s'interroge sur le retour des modèles terminologiques vers les textes et la gamme des accès qu'ils autorisent.

Dans cette double perspective, nous avons choisi d'organiser les questions qui se posent en quatre thématiques, théoriques et appliquées, qui ont constitué un maillage de référence de la réflexion. Les réponses apportées, propres à chaque champ disciplinaire, ont été rapprochées et discutées dans une perspective pluridisciplinaire. Ces quatre thèmatiques rendent compte des questions communes aux disciplines concernées par l'action spécifique :

- Identification des besoins,
- Nature des ressources,
- Utilisation des corpus
- Méthodes et outils.

C'est autour de ces quatre axes que se sont organisées les discussions. Pour les traiter, des groupes disciplinaires ont été constitués, car il nous a paru important de bien repérer et échanger les positions disciplinaires avant de s'interroger sur leurs convergences et complémentarités. La suite de cette partie rapporte l'état initial de la réflexion sur ces quatre points alors que les parties suivantes (3.2 à 3.6) correspondent aux développements proposés par chaque groupe disciplinaire en réponse à ces questionnements.

3.1.1 Identification des besoins

Nous avons identifié plusieurs sortes de besoins : des besoins portant sur le lien entre documents textuels et connaissances, les ressources terminologiques en étant une des matérialisations ; des besoins relatifs aux ressources terminologiques elles-mêmes, pour en définir de nouveaux modèles ; et enfin des besoins relatifs à de nouveaux usages, jusqu'ici inédits.

Le lien qui unit documents textuels et connaissances est symétrique. Dans un sens, on constate que les documents et collections de documents sont envisagés souvent comme de possibles sources de connaissances d'un domaine. Se posent alors des questions difficiles, qui méritent d'être examinées dans toutes leurs dimensions :

- Comment accéder à ces connaissances ? bien au delà d'un problème d'outillage et de technique, il s'agit aussi de se demander plus fondamentalement : en quoi un document contient-il des connaissances ? quelle est la nature de ces connaissances ?
- Peut-on rendre compte de « toutes » les connaissances que contiendrait un document ? ou va-t-on y chercher des connaissances particulières en fonction d'un objectif ?

Finalement, se pose la question du statut de ces documents dans des analyses dont la portée est plus large. On retrouve ainsi des problèmes communs à l'ingénierie des connaissances, au TAL et aux sciences de l'information.

Avant même de les considérer comme des sources de connaissances, les documents textuels peuvent être pris comme des moyens d'accès privilégiés à des manifestations linguistiques. Ainsi, l'analyse de corpus complète l'introspection dans une démarche d'étude de la langue. Le document est un précieux révélateur d'usages, de plus en plus exploité automatiquement depuis qu'il est sous forme électronique. En se focalisant sur des corpus spécialisés, on accorde du poids à des formes effectivement mises en œuvre, a priori partagées et stabilisées. Ce passage du document aux usages linguistiques puis à des connaissances lexicales, terminologiques ou documentaires intéresse la linguistique de corpus, la terminologie, les sciences de l'information.

Inversement, les connaissances représentées dans des structures terminologiques ou autres thésaurus sont autant de moyens pour revenir aux documents, pour accéder à leur contenu. Les terminologies peuvent servir à indexer, à annoter, à caractériser par des mots-clés, à classer ou catégoriser, etc. Là encore, il s'agit autant d'un problème de fond, de réflexion sur le sens des mots dans le document et leur lien possible avec une structure de données, que de problèmes techniques :

- Comment accéder aux documents à travers les connaissances ? faut-il poser des liens définitifs ou dynamiques ? s'appuyer sur l'usage des termes ou sur les structures conceptuelles ?
- Comment faire vivre les connaissances et les corpus dans des contextes aussi dynamiques que le web ou celui d'une production documentaire toujours croissante ?
- Quels produits terminologiques intermédiaires pourraient faciliter cet accès ?

On rejoint là la question de la *nature des ressources terminologiques.* On constate aujourd'hui que la plupart des ressources sont statiques, figées, construites manuellement, organisées de manière très linéaire et faiblement structurées, sans lien avec des textes qui en illustreraient le sens ou l'usage. Cette question comporte plusieurs facettes :

- Face à une évolution des besoins, exprimée dans toutes les disciplines, les types de ressources actuels ne semblent pas toujours adaptés. Quels nouveaux types de ressources définir pour une meilleure réponse à ces besoins ?
- Comment ces ressources peuvent-elles être utilisées dans des domaines particuliers, alors que beaucoup d'entre elles sont générales ? Comment ces ressources, souvent prévues pour être utilisées dans un objectif spécifique - associer des mots-clés à un document, le retrouver au sein d'une collection, faciliter l'expression de requêtes, clarifier une terminologie technique - peuvent-elles rester pertinentes pour des usages très différents comme la représentation de connaissances ou la recherche d'information au sein de documents par exemple ?

Cette adéquation entre ressources existantes ou à construire et usages est d'autant plus difficile à assurer que de nouveaux besoins apparaissent régulièrement. Nous avons identifié trois tendances marquées :

- De plus en plus d'applications couvrent des domaines spécifiques, ce qui suppose de constituer des ressources terminologiques spécialisées, avec souvent des contraintes temporelles fortes sur le temps à y consacrer ;
- Lorsque les documents jouent un rôle majeur dans ces applications, il est primordial de construire ces ressources à partir de textes pour mieux tenir compte des usages et rendre plus efficace l'accès aux documents ; plaquer des modèles génériques ou universels n'apporte pas de vraie solution;
- Enfin, les collections de documents autant que les terminologies elles-mêmes présentent deux propriétés à prendre en compte dès le départ : leur volume et leur évolutivité ; des propositions doivent être élaborées, tant des méthodes ou des logiciels que des structures de données, pour gérer au mieux la cohérence entre documents et ressources dans ce contexte.

3.1.2 Nature des ressources

Que ce soit au sein des entreprises, dans le monde de l'édition ou sur l'internet, la production de documents sous forme électronique s'accélère sans cesse et contribue, en parallèle, à une inflation de la documentation papier. Or pour produire, diffuser, rechercher, exploiter et traduire ces documents, les outils de gestion de l'information autant que les utilisateurs ont besoin de ressources terminologiques. On assiste actuellement à deux phénomènes pour s'adapter à cette nouvelle donne. D'une part, des ressources existantes sont utilisées pour des usages nouveaux, différents de ceux prévus à leur construction. D'autre part, de nouveaux types de structures sont définis pour mieux répondre aux nouveaux besoins. Finalement, la gamme des produits à base terminologique possibles pour répondre à ces besoins s'élargit considérablement (Bourigault & Jacquemin, 2000). A côté des bases de données terminologiques multilingues classiques pour l'aide à la traduction, des langages documentaires utilisés pour l'indexation et la recherche de documents, de thésaurus métier, etc. on voit apparaître de nouveaux types de ressources terminologiques ou ontologiques (RTO) adaptées à de nouvelles applications de la terminologie. Nous en mentionnons ici quelques uns à titre d'exemple :

- thesaurus pour les systèmes d'indexation automatique ou assistée,
- index hypertextuels pour les documentations techniques,
- terminologies de référence intégrées à des systèmes d'aide à la rédaction,
- référentiels terminologiques ou ontologiques pour les systèmes de gestion de données techniques,
- ontologies pour les mémoires d'entreprise, pour les systèmes d'aide à la décision ou pour les systèmes d'extraction d'information,
- glossaires de référence et terminologies structurées pour les outils de communication interne et externe.

La multiplication des types de ressources terminologiques et des langages documentaires met à mal le principe théorique de l'unicité et de la stabilité d'une terminologie pour un domaine donné, ainsi que celui de la base de donnée terminologique comme seul type de ressource informatique pour la terminologie. Le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas *une* terminologie, qui représenterait le savoir sur le domaine, mais autant de ressources terminologiques ou ontologiques que d'applications dans lesquelles ces ressources sont utilisées. Un documentaliste pour une activité de veille technologique aura besoin d'une terminologie du domaine différente d'un système d'aide au diagnostic, même dans le même domaine. Selon l'application, ces ressources peuvent différer sensiblement quant aux unités retenues et à leur description.

L'ensemble de ces constats empiriques entraîne des changements en profondeur de la pratique terminologique, et appelle du même coup à un renouvellement théorique de la terminologie. Il entraîne un questionnement nouveau également en recherche d'information, car il donne un statut pérenne à des ressources qui étaient jusque là des listes de termes. Le fait de s'intéresser à la dimension sémantique est récent et soulève des problèmes parfois orthogonaux à la recherche de l'efficacité de traitements massifs de données. De même, en ingénierie des connaissances, cette négation de l'universalité des ressources va à l'encontre de la motivation qui a donné naissance aux ontologies. Par la même, les ontologies véhiculent un malentendu : conçues pour faciliter l'interopérabilité, elles sont présentées comme consensuelles et souvent comme universelles. Or pratiquement, les ontologies ne sont pertinentes dans le cadre qui nous intéresse (le retour aux textes) que si elles sont des modèles des domaines étudiés intégrant le point de vue de l'application. De fait, elles sont alors locales et spécialisées: on parle d'ontologies régionales. De plus, elles accordent jusqu'ici une place minimale à la dimension lexicale et textuelle des connaissances, qu'il semble urgent d'intégrer au niveau de la représentation des connaissances. Enfin, en sciences de l'information, même si les recherches portent avant tout sur le lien entre une ressource donnée et son utilisabilité pour gérer des documents, une collection ou faire de la veille, l'évaluation de l'utilisation a vite un impact en retour sur le choix de la ressource et sur sa

construction. Des exigences nouvelles liées au constat de non unicité des ressources et à leur dynamique fait aussi ressortir le besoin de nouvelles structures.

L'AS s'est donc donné comme objectif de repérer et caractériser les types de ressources existant, le type d'application pour lesquelles ils ont été définis et auxquelles ils conviennent a priori, leur capacité à être utilisés dans d'autres contextes et, enfin, les évolutions nécessaires pour les adapter à des besoins nouveaux.

3.1.3 Utilisation des corpus : la question du genre textuel

Toutes les disciplines concernées par l'AS ont en commun un matériau d'étude : les textes. Mais il s'agit en fait de collections de textes organisées dans le but d'une analyse particulière ; il s'agit donc, plus précisément, de corpus.

Par rapport à l'utilisation des corpus, le TAL a sans doute une longueur d'avance. En effet, la vérification de modèles, qui constituait l'essentiel de la problématique TAL il y a une vingtaine d'années, a été en grande partie remplacée par l'analyse de matériau textuel, pas toujours d'ailleurs avec une réflexion poussée sur la nature de ce matériau (Péry-Woodley, 1995).

Pour les quatre autres disciplines, la prise en compte des corpus pour construire des terminologies, des ontologies (IC) ou des thésaurus (sciences de l'information et recherche d'information) est récente, en tout cas si l'on se place dans la perspective de chacune des approches. En effet, l'évolution épistémologique de ces disciplines s'est faite sur le même modèle. Point de départ : l'idée que les experts étaient les mieux placés pour fournir leur connaissance, dans des domaines qu'ils étaient censés parfaitement maîtriser. Les difficultés liées à ce mode de capitalisation de la connaissance ont été nombreuses : difficultés à trouver un accord entre les experts, sentiment de normalisation qui dénaturait la connaissance, inadaptation des ressources produites aux besoins réels. Une alternative est apparue sous la forme de l'utilisation de corpus. Avec cette perspective, une porte s'est ouverte pour chacune des disciplines. Mais conjointement à cette prise en compte des corpus, est survenu un autre problème majeur : la confrontation à la variation. Une façon de prendre en compte et de maîtriser cette variation consiste à travailler la notion de genre textuel. En effet, cette notion permet de caractériser un ensemble de textes supposés avoir les mêmes critères. Cette notion se décline différemment selon les disciplines. On la retrouve tout particulièrement en TAL, en linguistique de corpus et en sciences de l'information.

Le point de départ de la réflexion sur les corpus dans ASSTICCOT s'est donc fait sur le constat d'une problématique commune sur la question des genres textuels ; cette question est revenue fréquemment dans les discussions. Le paragraphe 4.1 montrera comment la réflexion, grâce aux convergences qui ont émergé, a permis de dégager des pistes de réflexion pour l'avenir.

3.1.4 Méthodes et outils

Dès que l'on se pose la question des méthodes et outils définis et utilisés pour aborder les recherches de chaque discipline, il ressort des manières très différentes d'appréhender le lien entre terminologie et corpus. Même si, initialement, l'objectif de l'AS était de se focaliser sur la construction de ressources terminologiques à partir des corpus, il est clair que les préoccupations inverses, soulevées par l'utilisation des terminologies pour organiser ou accéder à des textes, ont fait partie intégrante de la réflexion. En ce sens, l'apport des sciences de l'information et de la recherche d'information a permis d'élargir le point de vue retenu et la manière de poser les problèmes en intégrant fortement les usages et les pratiques.

Le thème des méthodes et des outils a été décliné sous trois angles :

- l'exposé des méthodes : comment, à partir d'un besoin d'utilisateurs, identifier un procédé qui va permettre de construire la bonne ressource, de la sélectionner ou de la faire évoluer ? mais aussi

comment organiser une collection en fonction d'une ressource ? comment associer ressource et collection ?

- l'étude d'outils et techniques d'analyse de textes : au delà d'un inventaire et des caractérisations des outils disponibles et utilisés, il s'agit de s'interroger sur l'articulation entre méthodes et outils car tous les participants à ASSTICCOT conviennent de promouvoir une approche supervisée, dans laquelle l'intervention humaine est indispensable pour organiser ou donner sens aux résultats des outils.
- La référence aux outils de modélisation et de structuration des ressources : les questions à aborder portent alors sur les structures de représentation que proposent ces outils et leur adéquation aux besoins en matière de ressources terminologiques, ou encore sur leurs connexions possibles avec les outils d'extraction.

Pour orienter notre confrontation d'expériences, nous avons choisi les méthodes de travail suivantes :

- s'appuyer sur des études de cas ;
- identifier et caractériser, pour chacun de ces cas, des points d'impact de l'application visée sur la démarche de construction de ressources terminologiques ; parmi ces points, nous avons d'ores et déjà listé les suivants :
  1) Profil du « constructeur »
  2) Construction du corpus
  3) Choix de la structure de données
  4) Utilisation des outils de TAL, de fouille de textes
  5) Utilisation des outils de modélisation
  6) Validation, évaluation
- dresser des perspectives pour une meilleure maîtrise et adéquation du processus.

Du point de vue des méthodes, nous constatons, comme point de départ, l'existence de résultats théoriques, de méthodes et des outils, qui aboutissent à des résultats prometteurs. Les convergences dégagées au sein des différents groupes de travail que nous avons mentionnés plus haut constituent finalement des cadres relativement unifiés au sein duquel se déclinent un éventail de pratiques. Parmi les bases de ce cadre unifié, un point fort est que toutes les approches rendent compte de l'usage spécialisé de la langue, ce qui pose la question de la complémentarité entre, d'une part, ressources existantes, qu'elles soient linguistiques (dictionnaires, WordNet) ou conceptuelles (ontologies de haut niveau) et, d'autre part, des ressources spécialisées à construire.

3.2 Prise en compte des thèmes retenus en Terminologie et linguistique de corpus

La prise en compte des corpus en linguistique (c'est-à-dire, de données réelles et pas seulement d'exemples forgés) commence à devenir une question essentielle pour la discipline, au point que la façon dont elle est considérée peut constituer une méthode pour présenter ses différents points de vue (3.2.1). La terminologie textuelle a, elle aussi, émergé récemment pour des raisons différentes mais, ce faisant, elle s'est nettement rapprochée de la linguistique (3.2.2). Cette émergence est allée de pair avec la définition du concept interdisciplinaire de base de connaissances terminologiques (3.2.3).

3.2.1 Les corpus en linguistique

Trois modes de prise en compte des corpus vont être examinés: pas de prise en compte du tout, prise en compte d'un corpus « introspectif », prise en compte d'un corpus qui sert de référence.

a- Aucun corpus n'est pris en compte

Mener une étude sur un corpus oblige à une confrontation avec la réalité des usages, ce qui ne va pas sans poser de question à une linguistique qui voudrait s'inscrire dans une perspective scientifique et, pour cela, décrire un système stable, c'est-à-dire un système dont on maîtrise les éventuelles variations. En effet, les textes ne sont pas seulement des attestations de la mise en œuvre d'un système ; ils s'inscrivent nécessairement dans une situation particulière, qui engage des locuteurs réels et qui se caractérise par une certaine fluctuation par rapport à la norme. Une des façons de contourner cette difficulté consiste à ne pas avoir recours à des productions réelles mais, au contraire, à se donner un objet déconnecté de tout contexte afin d'établir la distance qui permettra à l'analyste de repérer les régularités inhérentes au système. C'est le choix qui sous-tend le structuralisme et le générationalisme, le premier pour des raisons essentiellement méthodologiques, le second parce que la langue est considérée comme relevant de facultés psychologiques innées et universelles.

On ne peut que constater que ces deux courants ont permis une évolution majeure dans la connaissance du fonctionnement linguistique ; le structuralisme a en particulier introduit une rupture avec une vision référentielle qui, jusqu'au début du XX° siècle biaisait considérablement les études. Toutefois, les limites de ces approches semblent maintenant atteintes, à la fois parce que la variation fait irruption dans la problématique théorique de la linguistique et parce que la demande sociétale de résultats d'analyse de corpus est très importante.

b- Recours à un corpus «introspectif »

Beaucoup de linguistes, conscients de la nécessité de moduler les descriptions, essaient de modéliser les phénomènes de variation sur la base du recours à leur propre intuition. Leurs propres attestations constituent ainsi une sorte de corpus dont ils essaient de décrire les régularités. La plupart des chercheurs en syntaxe travaillent de cette manière mais aussi la plupart des chercheurs en sémantique lexicale et beaucoup de chercheurs en énonciation. La notion de contexte, qui permet de contrôler la variation, est alors intégrée dans le modèle et donc parfaitement maîtrisée.

Les résultats obtenus présentent bien sûr des intérêts : les analyses sont souvent très fines et appuyées sur de très nombreux exemples qui, pour être forgés, n'en sont pas moins acceptables. Ce qui fait problème dans ce type d'approche relève de deux ordres.

1) Les exemples sont proposés dans le cadre du test linguistique ; ainsi, ils ne sont pas irrecevables mais ils sont coupés de toute situation langagière réelle, c'est-à-dire d'une situation qui est d'abord une situation d'échange. Par ailleurs, comme le soulèvent de nombreux linguistes, cette méthode accorde une grande importance au jugement du linguiste qui a tendance à ériger en règle générale sa propre acceptation des phénomènes. Tout linguiste a pourtant fait l'expérience d'entendre ou de lire un fait langagier que, quelques jours auparavant, il avait jugé comme impossible.
2) Ces exemples sont tous mis sur le même plan ; ainsi, des phénomènes rares sont considérés au même titre que des phénomènes fréquents. Cette situation s'avère particulièrement problématique lorsque l'on prend en compte des textes réels dans lesquels la répartition chiffrée des phénomènes à l'intérieur de chaque texte ou bien d'un texte à l'autre, prend un relief tout à fait significatif et participe à la construction du sens.

De plus en plus de linguistes essaient de palier ces difficultés en recourant à des attestations réelles, soit à travers l'utilisation de corpus constitués ou bien en faisant des recherches sur l'internet. Mais dans un cas comme dans l'autre, on considère souvent ces attestations comme autant de manifestations du système langagier sans tenir compte de leur origine. Dans le cas de l'utilisation de Frantext par exemple, le risque est grand de proposer des généralisations à partir d'une étude qui ne se base que sur des textes littéraires, souvent du XVIII° ou XIX° (les plus représentés dans Frantext) qui constituent pourtant un

usage langagier bien particulier. Quant à l'internet, pour qui veut essayer de comprendre la variation linguistique, il constitue une véritable énigme tant les variations de genres semblent importantes et pour l'instant, incontrôlables.

c- Corpus comme référence

L'utilisation des corpus en linguistique est loin d'être un phénomène nouveau. Cependant, si les corpus permettent de mieux rendre compte du phénomène de la variation, ils sont loin de résoudre toutes les questions qu'elle pose. Toute la problématique de la linguistique est traversée par un questionnement (une tension) qui s'établit entre la nécessité de définir ce qui fait système, qui est stable et ce qui peut varier, collectivement ou individuellement, parfois jusqu'à l'infini. Dès que l'on quitte le domaine idéal du parfaitement stable, on est confronté au phénomène de la variation et à la nécessité d'en définir les contours. La prise en compte des corpus permet de poser les bases d'une réflexion sur cette problématique mais, dans l'état actuel des recherches, elle est loin d'avoir donné des réponses définitives. Dans une première approximation, on peut considérer deux points de vue selon le mode de prise en compte des corpus : l'un qui considère le corpus comme manifestant la compétence langagière et qui permet d'étudier la langue, l'autre qui considère le corpus comme la référence d'une étude particulière, les résultats ne concernant que ce corpus particulier.

Corpus comme représentatif de la compétence des locuteurs

Trois grands domaines sont concernés par la description d'une langue à partir de corpus puisque c'est de cela qu'il s'agit : la lexicologie (par exemple, constitution du dictionnaire Cobuild), la description de la grammaire, et enfin l'apprentissage d'une langue étrangère. L'utilisation des corpus pour ce type de perspective est nettement plus développée dans la tradition anglo-saxonne que dans la tradition francophone. On la retrouve cependant pour le français pour la constitution du Trésor de la Langue Française (même si sa construction est bien moins systématisée que celle du Cobuild) ou encore dans la mise en place du « français fondamental », destiné à l'apprentissage du français par des étrangers et élaboré à partir d'un corpus d'extraits attestés à l'oral ou à l'écrit.

Ces types de projets, qui ont le mérite de prendre en compte la réalité des usages (de certains usages) posent tous la même question de la représentativité. En effet, s'il s'agit de décrire le fonctionnement tel qu'il se manifeste dans les corpus étudiés, il faut donc que ces corpus, puisqu'ils ne peuvent rendre compte de tous les usages, soient au moins représentatifs de tous ces usages. Or, non seulement, nul n'a les moyens de vérifier cette représentativité mais, et c'est plus ennuyeux, on voit nécessairement réapparaître l'introspection, que l'on pensait évacuer par le recours aux corpus. Des projets d'envergure essaient de baser le rapprochement des textes sur des similitudes linguistiques (Biber, 1988) ; cela suppose de distinguer une caractérisation réalisée *a priori* et une caractérisation réalisée sur la base de régularités linguistiques avérées.

Ces méthodes semblent prometteuses mais elles ne suppriment pas, pour l'instant, l'étape de définition intuitive de genres. C'est en effet sur des bases introspectives que sont définis les différents registres : discours journalistiques, textes littéraires, lettres … Le recours aux corpus permet ainsi de se rapprocher de l'usage réel de la langue mais il ne parvient pas à constituer un objet définitivement maîtrisé.

Corpus comme objet d'étude

Certaines disciplines considèrent qu'une fois élaboré, le corpus constitue la référence à leurs travaux ; ce sont en quelque sorte les régularités propres à ce corpus qu'il faut mettre au jour. Il faut dans tous les cas que le corpus soit construit d'une manière cohérente, soit qu'il émane d'un groupe de locuteurs identifié *a priori* comme c'est le cas dans la sociolinguistique, l'analyse de discours ou même la théorie

des sous-langages, soit qu'il soit constitué dans une perspective particulière comme dans le TAL ou la terminologie textuelle. En limitant la portée des résultats aux corpus auxquelles elles s'intéressent, ces disciplines ont le mérite de percevoir les limites de leur approche. Mais, en s'interrogeant peu sur les modes d'élargissement de ces résultats, elles laissent de côté des questions fondamentales pour la linguistique, qui permettent d'expliquer le fonctionnement même de la langue par la mise en œuvre de connaissances partagées (en tout cas nécessairement supposées partagées). Lors de ces analyses de corpus, quelles connaissances, déjà présentes, sont convoquées et quelles connaissances construites pourraient être réutilisées lors de nouvelles analyses ?

Ancrées dans la réalité des usages, ces approches à partir de corpus font émerger un élément majeur, particulièrement pour la sémantique. En effet, la plupart d'entre elles considèrent que le sens n'est pas un donné mais un construit et que le corpus est soumis à une interprétation. Dans une telle perspective, la question qui se pose est de savoir quels sont les modes possibles de contrôle de l'interprétation. Comme le dit Pécheux :

> « L'analyste de discours ne prétend pas s'instituer en spécialiste de l'interprétation maîtrisant « le » sens des textes […]. L'enjeu crucial est de construire des interprétations sans jamais les neutraliser ni dans le " n'importe quoi " d'un discours sur le discours, ni dans un espace logique stabilisé à prétention universelle. (Pécheux, 1984) » (Maingueneau, 1987, 6).

Dans ce cadre général qui voit de nombreux courants de la linguistique s'intéresser aux corpus, comment peut-on situer la terminologie textuelle et en quoi peut-elle éclairer des problématiques comme la systématisation des résultats obtenus sur un corpus ?

3.2.2 Terminologie et textes

Dans la vision wustérienne, la prise en compte des usages manifestés dans les textes ne peut être la base de la constitution de terminologies. En effet, le discours, dans ses possibilités créatrices peut menacer les fondements mêmes des terminologies.

> «[…] jusqu'à une date récente, la linguistique n'a fait valoir que l'évolution libre, non dirigée, de la langue. C'est l'usage effectif de cette dernière qui, dans la langue commune, sert de norme. On peut appeler cette norme la norme descriptive. En revanche, en terminologie, fertile en notions et en termes, cette évolution libre de la langue mène à une confusion inacceptable… » (Wuster, 1981, 65).

Pour Wüster en effet, la terminologie est normative, par essence et/ou, par objectif. Comme l'a montré Slodzian entre autres (Slodzian, 1995), Wüster croyait en l'existence d'une langue scientifique épurée (en tout cas épurable) de ce qu'il considérait comme les éléments qui nuisaient à une communication transparente.

La réalité de la pratique terminologique se révèle tout autre. En effet, les textes, entendus comme des productions langagières effectives, sont nécessairement pris en compte parce que les terminologues ne peuvent s'appuyer sur leurs seules intuitions linguistiques dans des domaines où ils n'ont pas de compétence. Pour contourner cette « non-compétence », les terminologues font appel à des experts qu'ils interrogent mais aussi à des productions de toutes natures : manuels, documents d'entreprises, listes de termes existantes… Les praticiens et plus encore les chercheurs se heurtent alors à une double difficulté. D'une part, cette prise en compte des réalisations possibles n'est pas systématisée ; or, il est bien évident qu'il y a une difficulté à considérer comme autant d'attestations des textes qui s'adressent à des non-experts et d'autres qui sont très spécialisés ou encore, des textes qui ont une visée principalement injonctive et d'autres qui ont une visée descriptive. D'autre part, la tâche qui est demandée au terminologue revient à construire une norme à partir d'usages attestés. Très rapidement, il est confronté

au manque de balisage que constitue cette situation si elle n'est pas accompagnée d'une réflexion sur l'objectif de la normalisation.

3.2.3 Les bases de connaissances terminologiques

Les bases de connaissances terminologiques (BCT) sont un concept récent (Meyer *et al.*, 1992). Leur apparition a manifesté deux évolutions importantes, d'ailleurs reliées : l'une concerne l'affirmation d'une relation entre problématique de la terminologie et problématique de l'intelligence artificielle (dans le terme de bases de connaissances terminologiques, on retrouve en effet, celui de base de connaissances, qui provient de l'intelligence artificielle), l'autre a concerné la nécessité de donner une représentation relationnelle aux définitions terminologiques jusqu'alors existant sous une forme uniquement discursive. La réflexion qui est menée sur les BCT est ainsi d'emblée pluridisciplinaire et la création du terme de BCT ne vient en fait que consommer un lien qui s'est mis en place plusieurs années auparavant. Bien que très récent, ce concept de BCT a déjà beaucoup évolué et on peut penser que cette évolution vient de ce que la confrontation interdisciplinaire a conduit chacune des disciplines à éclaircir ses postulats et à développer la réflexion. Ainsi, pour la terminologie, c'est la réflexion sur le recours aux corpus qui a bénéficié de la rencontre avec l'informatique ; et pour l'intelligence artificielle, la réflexion sur les ontologies et leur éventuelle généricité a certainement été enrichie par l'apport de la vision de la terminologie textuelle. Si bien que, si on les replace dans leur chronologie, les BCT peuvent être vues à la fois comme un point de jonction qui concrétise une réflexion pluridisciplinaire et comme une ouverture vers des analyses mono-disciplinaires enrichies.

Ce thème des BCT a constitué le principal élément de réflexion qui a permis de fonder la constitution du groupe TIA, en 1993 (Bourigault & Condamines, 1995), (Aussenac-Gilles, 1999), (Aussenac-Gilles & Condamines, 2001).

3.3 Prise en compte des thèmes retenus en TAL

Le TAL est particulièrement concerné par le thème de la terminologie en lien avec les corpus et ce, de deux manières complémentaires. D'une part, le TAL peut être producteur de ressources terminologiques ; de très nombreux outils sont ainsi développés pour aider à extraire la terminologie à partir de corpus électronique (3.3.1). D'autre part, le TAL est aussi consommateur de ressources terminologiques, dont la nature peut être diverse en fonction des tâches à faire réaliser par les outils (3.3.2).

3.3.1 TAL comme producteur de ressources terminologiques

Soutenue par une demande sociétale forte, la définition d'outils d'aide à la constitution de ressources terminologiques à partir de textes est un des domaines les plus productifs du TAL. Il faut toutefois noter que très peu de ces outils sont commercialisés ; pour la plupart en effet, ils sont développés dans des laboratoires de recherche, le plus souvent dans le cadre d'une thèse. Par ailleurs, souvent faute de résultats d'études poussés sur des corpus spécialisés, qu'on pourrait attendre de la linguistique, les outils proposés ne donnent pas toujours les résultats de très bonne qualité. Sans entrer dans les détails de ces nombreux outils, on peut toutefois les présenter en trois catégories selon leur objectif :

- *Outils d'extraction de termes candidats.* Ces outils visent à proposer des mots ou, le plus souvent des groupes de mots susceptibles d'être des termes. Ces listes sont ensuite proposées à des experts, cogniticiens ou terminologues qui doivent les valider ou les invalider. Parmi ces outils on peut citer Lexter (Bourigault, 1994) et Nomino (David et Plante, 1990) pour le français ou ANA (Enguehard et Pantera, 1995) qui s'adapter à différentes langues.
- *Outils d'extraction de « relations candidates ».* Comme pour les termes mais cette fois-ci pour les relations, ces outils proposent des listes constituées d'extraits de textes susceptibles d'être

représentés sous forme de relations. Selon les cas, la possible nature de ces relations (hyperonymie, méronymie, cause…) est proposée ou non. Des outils comme Seek (Jouis, 1993), Caméléon (Séguéla & Aussenac-Gilles, 1999), Prométhée (Morin, 1999), Likes (Rousselot et al., 1996) ont pour objectif de proposer des relations candidates en s'appuyant sur des techniques différentes : exploration contextuelle, utilisation de marqueurs lexico-syntaxiques de relations ou encore repérage de contextes partagés.

- *Outils d'exploration de textes.* Ces outils sont les plus généralistes. Ils ne visent pas particulièrement la recherche terminologique mais plus généralement l'exploration de textes, pour des objectifs très divers. Leur noyau commun est constitué par un concordancier mais ils peuvent être enrichis de connaissances linguistiques (syntaxiques, voire sémantiques) et proposer des fonctionnalités assez diverses. Citons par exemple Sato (Daoust, 1992) et Yakwa (Rebeyrolle et Tanguy, 2001), qui retournent des fragments de textes contenant des concordances lexicales ou grammaticales, ainsi que Syntex (Bourigault et Fabre, 2000) qui produit un réseau terminologique relié au texte et Upéry, basé sur une analyse distributionnelle (Bourigault, 2002)

Les outils d'aide à l'extraction proprement dite, que ce soit l'extraction de termes ou l'extraction de relations, reposent sur deux grands principes qui révèlent deux façons de concevoir le fonctionnement de la langue. Dans le premier type d'outils, tout texte est considéré comme la mise en œuvre d'un système très stable. Ainsi, les termes sont considérés comme respectant des patrons très récurrents (par exemple, *Nadj de N*). Quant aux relations, on considère que des marqueurs, identifiables par introspection, les mettent en œuvre de manière régulière dans les textes, par exemple *tous les N1 sauf N2* permettrait de repérer systématiquement une hyperonymie entre N2 et N1 comme dans *Paul aime toutes les fleurs sauf les roses*. Le second type d'outils considère au contraire qu'on ne peut pas prévoir tous les phénomènes langagiers qui peuvent apparaître dans les textes et qu'il faut s'attendre à découvrir des éléments (formes de termes ou marqueurs de relation en l'occurrence) qui n'auraient pas pu être prédits par introspection. Ces outils s'inscrivent plutôt dans une tradition distributionnelle : c'est la récurrence des contextes et des distributions qui fait sens.

3.3.2 TAL comme utilisateur de ressources terminologiques

Ce paragraphe va permettre de dresser un catalogue assez complet des différentes utilisations actuelles des ressources terminologiques.

Systèmes à base de connaissances

Les systèmes à base de connaissances ont besoin pour fonctionner d'une représentation explicite des connaissances du domaine. Il s'agit d'une représentation des objets du domaine, de leurs propriétés et de connaissances heuristiques (règles ou axiomes) relatifs aux définitions et raisonnements qui puisse être mis en œuvre par un système informatique. La représentation doit donc être intelligible par un système de résolution de problème (ou moteur d'inférence). Ces connaissances forment une *ontologie*. Elle comprend, généralement et dans les grandes lignes :

- un inventaire des notions du domaine
- un système de relations

Une ontologie se distingue des autres ressources terminologiques par *un besoin de formalisation plus grand*. Ce besoin provient en partie du fait qu'un système à base de connaissances fonctionne sur des entrées qui ne sont généralement pas textuelles mais opérationnelles. Il a donc besoin de s'abstraire des énoncés textuels et donc de travailler avec un *modèle* du domaine dont l'interprétation informatique soit possible et non ambiguë. Les ressources terminologiques peuvent servir à des outils de TAL utilisés pour construire ces ontologies. CE point sera développé dans la partie 3.4.

Recherche d'information

Dans les grandes lignes, le schéma de base du processus de recherche d'information est le suivant :

- *requête* en texte libre
- retour d'une liste ordonnée de *documents*

Les étapes de ce processus incluent l'indexation (ou représentation du document), le traitement de la requête (simplification, expansion), la comparaison de la requête avec l'index et le classement des résultats. Différents paramètres conditionnent le lien entre système de recherche d'information, ressources et logiciels de TAL. Ce processus sera détaillé et commenté en 3.5.

On peut distinguer système de recherche d'information « générique » (pas de spécialisation de domaine) et système de recherche d'information « spécialisée » (le domaine est fixé). Le second a plus de chances de pouvoir bénéficier de ressources terminologiques et d'analyses de textes.

On peut situer un système de recherche d'information sur un axe qui va de la recherche d'information de type Internet (fonds ouvert), à travers un moteur de recherche existant dont on ne contrôle pas le processus d'indexation à la recherche d'information sur une base construite (fonds fermé), à l'aide d'un moteur de recherche spécifique, pour lequel on contrôle de la tâche d'indexation (on peut effectuer une analyse des documents avant l'indexation). Les ressources et les outils de TAL peuvent améliorer les performances dans le 2è contexte.

Extraction d'information

La tâche principale de l'extraction d'information est le *remplissage de schémas prédéterminés* (extraction d'informations ciblées) à partir de textes d'un genre également prédéterminé (par exemple, détermination des actants d'un attentat à partir de dépêches de presse). Les informations extraites de chaque texte ne sont qu'une petite partie des informations que contient ce texte (représentation non exhaustive). Une sous-tâche de l'extraction d'information est la reconnaissance d'entités nommées.

Les ressources terminologiques peuvent servir ici à mettre au point des patrons de recherche utilisés par les logiciels de TAL chargés de l'extraction. De plus, d'autres outils de TAL plus élémentaires peuvent être intégrés dans ce processus : analyseurs syntaxiques, logiciel de repérage d'entités nommées ou de désambiguïsation.

Questions-réponses

L'utilisateur pose une question. Le système recherche des extraits de documents qui contiennent la réponse à cette question. Ces systèmes poussent plus loin la problématique de la recherche d'information en retrouvant, à partir d'une requête qui est en fait une question, non pas des documents entiers, mais des extraits, les plus précis possible, éventuellement reformulés, qui répondent au plus près à la question posée. De nombreux traitements TAL sont nécessaires pour retrouver l'information exacte, et l'utilisation de ressources terminologiques et linguistiques est indispensable pour caractériser la nature de la question et passer de la formulation de l'utilisateur à celle présente dans les textes.

Veille technique et scientifique

La veille scientifique se distingue de la recherche d'information ou l'extraction d'information par le fait que les éléments recherchés sont censés relever d'une connaissance ou d'un fonctionnement inconnu. Les systèmes de veille se focalisent sur des domaines délimités, dont la terminologie peut être identifiée. Les ressources terminologiques sont de précieux moyens de caractériser les documents trouvés ou

recherchés, de définir des profils d'utilisateurs ou de juger de la pertinence ou de la nouveauté des informations retrouvées.

Catégorisation de documents

Un document étant donné, il s'agit de déterminer dans quelle(s) catégorie(s) le ranger. Les méthodes principales n'utilisent pas de ressources terminologiques (au mieux, une simple lemmatisation, voire une racinisation des mots). Une ressource terminologique incluant des liens de proximité entre termes (synonymie, hyperonymie, voir-aussi) permettrait de tester l'emploi d'une similarité entre termes dans les calculs de similarité entre documents et catégories.

Classification de documents

Ici, la liste des catégories n'est pas donnée a priori. Il faut faire émerger une organisation de l'ensemble de documents traité en les regroupant par similarité (en anglais, « clustering »). Le même principe que pour la catégorisation pourrait s'appliquer (relations de proximité entre termes prise en compte dans le calcul de la distance entre deux documents) pour la caractérisation des classes de documents trouvées à l'aide de ces termes.

Navigation dans des bases textuelles

Il s'agit de mettre à disposition de l'usager (lecteur) d'une base documentaire un index dans lequel il navigue (liens hiérarchiques, liens transversaux aussi ?). Les feuilles de l'index pointent vers des documents. Le principe est similaire à l'index d'une documentation technique. Cet index peut être vu comme une aide à la formulation de requêtes pour des recherches au sein de cette base. Le Tal est un des moyens de constituer ce type d'index.

Traduction automatique et variantes

Nous incluons ici les outils d'aide au traducteur (qui proposent des mémoires de traduction, des terminologies bilingues, etc.), les systèmes de traduction automatique, la traduction (expansion) de requête en recherche d'information trans-langue. Une part importante des ressources terminologiques générales (comme la BD lexicale WordNet) disponibles sur support informatique ont été produites pour cet objetif.

Indexation automatique contrôlée

Les documents de certaines bases documentaires (par exemple, Medline : résumés d'articles du domaine bio-médical) sont indexés manuellement par les termes d'un thesaurus (hiérarchisé : le MeSH, Medical Subject Headings). Il s'agit d'assister, voire d'automatiser cette tâche. Pour que ce soit plus facilement faisable, le *thesaurus d'indexation* doit avoir une forme particulière (ses termes doivent être proches d'énoncés en langue naturelle) pour faciliter la recherche de variantes, etc.

Indexation multimédia

Cette indexation (d'images, de sons, de vidéos) se fait généralement sur un texte (paratexte, épitexte?) qui accompagne et décrit le document non textuel. On se ramène alors à la problématique de la recherche d'information textuelle. Elle peut se faire aussi l'aide de méta-données (mots-clés) qui peuvent être

puisées dans des ressources terminologiques décrivant le domaine concerné. On retrouve alors la recherche d'information et l'indexation par mots-clés.

Résumé automatique

Les méthodes se fondent généralement sur l'*extraction des phrases les plus saillantes*. Cet objectif requiert la mise au point de marqueurs lexicaux et d'heuristiques de reconnaissance de structures textuelles. Une sous-tâche est la *résolution d'anaphores*, qui peut être aidée par des variantes de termes (recherche de chaînes co-référentielles : pour l'analyse des textes et le collage des phrases extraites).

Constitution de glossaire (dictionnaire)

Un glossaire contient des mots (termes) et leurs définitions. On peut envisager d'extraire les définitions de corpus (travaux toulousains, projet Deffinder à Columbia, etc. par exemple en utilisant des marqueurs de définition) ou de les construire par observation des emplois de chaque terme. Le résultat de cette tâche (le glossaire) est précisément une ressource terminologique. Il est fait pour être consulté par un être humain, contrairement à beaucoup d'autres ressources terminologiques examinées jusqu'ici (une exception est la terminologie bilingue présentée à un traducteur humain).

Segmentation thématique

Cette tâche consiste à délimiter des parties d'un texte qui parlent chacune d'un *thème uniforme*. Un réseau lexical / terminologique est utile pour cette tâche (voirtravaux du LIMSI). Nous considérons qu'il s'agit généralement d'une tâche auxiliaire plutôt que d'une tâche finalisée, qui prépare une utilisation plus efficace d'autres outils d'analyse de textes pour une objectif finalisé.

Diagnostic de langue

Un document étant donné, il s'agit de déterminer sa langue ou, pour un document plus long, comportant des passages de langues différentes, d'identifier chacune des langues présentes dans les différents paragraphes. Il faut alors découper le document selon les langues employées, ce qui permet par exemple de les regrouper pour les traiter ensuite de manière homogène. Il s'agit ici aussi d'une tâche "auxiliaire".

Ce repérage des principales applications utilisant du TAL permet de mesurer la diversité des besoins en matière de ressources terminologiques ou ontologiques, la nature de ces ressources étant liée à chaque type d'application. Il est probable que la nature des ressources nécessaire est différente voire très différente en fonction des usages. Ce lien, entre nature des besoins et nature des ressources, est encore insuffisamment identifié. Il a été un des thèmes qui a alimenté la discussion au sein de l'AS.

3.4 Prise en compte des thèmes retenus en ingénierie des connaissances

3.4.1 Présentation et objet d'étude

L'ingénierie des connaissances (IC) est un domaine de recherche étroitement lié à l'informatique et à l'intelligence artificielle. Elle s'intéresse à la conception de systèmes nécessitant de gérer ou manipuler des connaissances, qu'elles correspondent à des savoir-faire, des pratiques, à des informations écrites ou

structurées pour assister un opérateur dans sa tâche. Ce domaine cherche à définir des techniques, des langages de représentation des connaissances et des logiciels de modélisation pour acquérir les connaissances nécessaires à la réalisation de ces systèmes, c'est-à-dire à leur identification, à leur recueil et à leur structuration avant leur formalisation. Depuis les années 90, les recherches montrent l'intérêt d'organiser les connaissances au sein de modèles conceptuels, qui en sont des représentations non opérationnelles, avant de les formaliser. Ces modèles distinguent connaissances du domaine et connaissances du raisonnement ou de la tâche. Ils sont devenus le principal objet d'étude en IC, et avec eux les méthodes et outils permettant de les construire. Le processus de modélisation est vu comme un processus supervisé par un cogniticien qui s'appuie sur des techniques et des logiciels.

La problématique de l'ingénierie des connaissances (IC) rejoint les thèmes de recherche abordés par ASSTICCOT essentiellement sur trois points :

- en prenant les textes comme sources de connaissances, partagées et stabilisées à l'écrit, complémentaires de l'expertise humaine, l'IC se pose aussi la question des corpus ;
- en cherchant des méthode et des outils pour un dépouillement systématique et efficace de ces textes, elle est amenée à étudier les apports du TAL, de l'apprentissage et de résultats linguistiques ;
- en proposant des structures de représentation des connaissances pour un objectif finalisé avec une composante lexicale ou non, comme les ontologies ou les bases de connaissances terminologiques, elle contribue à la définition de ressources.

Vers 1995, les recherches en IC ont pris un nouveau virage suite à deux évolutions marquantes : d'une part le succès de la notion d'ontologie et d'autre part, le déploiement de méthodes et d'outil d'acquisition de connaissances à partir de textes. Au delà des ontologies, l'IC s'intéresse à une gamme de plus en plus large de structures de données pour rendre compte des contenus des textes. Ces structures viennent souvent d'autres disciplines et sont par exemple les terminologies, thésaurus, base de données lexicales, taxinomies, lexiques, langages documentaires ou bases de connaissances terminologiques.

3.4.2 Les ontologies en ingénierie des connaissances

a. Définition

Pratiquement, les ontologies correspondent à une représentation informatique (en général formelle) des concepts, des relations sémantiques et des heuristiques d'un domaine. Conçues initialement comme des représentations de connaissances consensuelles et interopérables (faciles à échanger entre systèmes informatiques), les ontologies héritent d'une tradition qui privilégie leur réutilisabilité plus que leur fondement sémantique. Représentées à l'aide de langages standards (comme DAML ou OWL), elles sont de plus en plus mises à disposition comme ressources sur le Web (CHA,02). Par contre, leur contenu a tardé à faire l'objet de recherches autres que celles issues de la philosophie ou de la représentation des connaissances en Intelligence Artificielle. La construction d'ontologie est donc restée longtemps une tâche manuelle, partant de connaissances individuelles, fastidieuse et au résultat difficile à évaluer.

b. Historique

Les ontologies telles qu'elles sont conçues en intelligence artificielle et en ingénierie des connaissances ont pour origine le projet ARPA Knowledge Sharing Effort (Neches et *al* ., 1991). La proposition était la suivante : « *Building Knowledge-Based Systems today usually entails constructing new knowledge bases from scratch. It could be done by assembling reusable components. Systems developers would then only need to worry about creating the specialized knowledge and reasoners to the specific task of their system, using them to perform some of its reasoning.* ». Ainsi, la construction de systèmes intelligents devait être plus performante si, au lieu de créer entièrement les bases de connaissances nécessaires à ce type de

systèmes, on s'appuyait sur des composants réutilisables, pour n'avoir plus à considérer que les connaissances et les modes de raisonnement spécifiques au système construit. Les ontologies furent par conséquent conçues pour répondre à un premier besoin de réutilisation et de partage des connaissances. Le besoin d'assurer une interopérabilité entre les systèmes s'appuyant sur des connaissances a également contribué à l'origine de la construction d'ontologies. Pour communiquer, ces systèmes doivent faire appel à des représentations du monde compatibles et cohérentes à défaut d'être identiques. L'ontologie vient par conséquent répondre au besoin de recherche d'invariants dans le domaine, d'une description générique ou au minimum, consensuelle des connaissances.

c. Construction d'ontologies

Jusqu'en 1996, les premières ontologies ont été développées de façon *ad hoc*, sans suivre de méthodes prédéfinies. Les grandes lignes méthodologiques, qui sont associées aux premiers projets de construction d'ontologies (TOVE,1995 ; Entreprise Ontology, 1995), sont apparues en 1995 et ont été affinées par Uschold et Gruninger en 1996. La démarche générale, commune à ces projets, comprend trois étapes : identifier le but de l'ontologie et délimiter le domaine d'intérêt ; construire l'ontologie (extraction de connaissances ; formalisation et codage ; intégration éventuelle d'ontologies existantes) ; évaluer l'ontologie. Depuis 1998, des cadres méthodologiques plus élaborés, comme METHONTOLOGY (Fernandez-Lopez et *al*., 1999), sont proposés. Ils s'appuient sur les ébauches des méthodes précédentes et s'inspirent de celles du génie logiciel. Une démarche générale commune se dégage de ces différents cadres qui reprend, dans un cycle de vie, les trois grandes étapes de : planification ; construction ; évaluation. En parallèle, des méthodes ont été fondées sur des d'approches linguistiques (Szulman et al., 2002), différentielle (Bachimont, 2000), semi-informelle (Kassel, 2002), et formelle (Guarino & Welty, 2000; Gangemi *et al.,* 2002). L'articulation de ces différentes approches reste à accomplir.

d. Ontologies et textes

Deux approches issues de travaux sur les Bases de Connaissances Terminologiques (Aussenac-Gilles & Condamines 2001) et les Ontologies Régionales (Bachimont, 1996, 2000) marquent le début des travaux de construction d'ontologies qui s'appuient presque exclusivement sur les textes. Aujourd'hui, le développement d'outils de TAL et de l'apprentissage, les évolutions de la linguistique de corpus et de la terminologie, l'évolution de la demande en ingénierie des connaissances conduisent à de nouvelles approches (Maedche & Staab 2000 ; Kang & Lee, 2001 ; Velardi et *al*., 2001) les plus automatisées possibles, pour construire des ontologies à partir de textes.

3.4.3 Les Corpus en ingénierie des connaissances

L'utilisation des corpus en ingénierie des connaissances se veut donc une réponse au problème de l'accès aux connaissances d'un domaine pour un objectif particulier, lié à une application informatique. Les enjeux associés sont multiples et traduisent les avantages attendus d'une analyse de ces corpus :

- gain de temps et réduction du coût par rapport à des entretiens d'experts ;
- fiabilité et stabilité de ce qui est modélisé, puisqu'il s'agit de connaissances fixées par l'écrit, a priori plus consensuelles car diffusées et partagées ;
- meilleure maintenance et lisibilité des modèles grâce au textes qui en constituent une sorte de documentation.

Bien sûr, les textes à eux seuls ne répondent pas toujours complètement au besoin, et le recours à des spécialistes du domaine s'avère souvent un complément indispensable :

- pour la validation des résultats obtenus en fonction de leur propres connaissances du domaine, voire même pour aider à les structurer ;
- pour accéder à des connaissances liées à des savoir-faire, des pratiques non encore explicitées dans des textes ;
- pour s'assurer d'une bonne adéquation aux besoins des utilisateurs.

Ces corpus sont constitués d'un ensemble de textes choisis en fonction de critères pragmatiques comme la couverture du domaine, l'application et le type de modèle visé, les caractéristiques des textes (contenu, genre textuel, date, auteurs, etc.) mais aussi leur disponibilité ou leur confidentialité. Les textes peuvent être tirés de documents techniques, de manuels mais aussi de documents moins bien rédigés ou moins structurés comme des traces d'incidents, des retranscriptions d'entretiens ou des informations échangées par messagerie ou sur des forums.

Un corpus peut être caractérisé selon des critères qui font référence aux conditions de son élaboration (construit ou imposé), aux contraintes liées à l'application (ouvert ou fermé) et à l'ensemble des documents le constituant (sa taille, sa représentativité, son homogénéité). En outre, le type des documents, leur genre textuel (structuration, style, etc.), leurs conditions de production sont repérés. Les critères pour comparer les types de corpus comprennent entre autres le support, la qualité, les utilisateurs visés, la taille, le domaine.

Cependant, l'ingénierie des connaissances parle plus de « textes » que de documents, ce qui révèle bien que l'attention est portée plus au contenu qu'à la forme ou au support. La plupart des outils d'analyse traitent le contenu des textes comme un ensemble de mots, sans tenir vraiment compte de la structure du document, de sa mise en forme ou même de sa nature. Finalement, la plupart des dimensions mentionnées ici guident l'analyste à constituer son corpus, lui permettent de savoir comment appliquer les logiciels d'analyse, de mieux en interpréter, relativiser ou valoriser les résultats. Mais leur influence ne va pas au-delà.

3.4.4 Méthodes proposées par l'ingénierie des connaissances

L'analyse des travaux, dans le domaine de l'acquisition des connaissances à partir de textes, permet de dégager pour la construction d'ontologie, un cadre méthodologique :

- *consensuel*, qui rend compte des principales propositions du domaine,
- *abstrait*, non détaillé, justement pour conserver son statut de synthèse,
- *théorique*, qui propose un cadre complet et donc idéal.

Les propositions concrètes du domaine accordent aux différentes parties du cadre des importances inégales et permettent, dans certains cas, d'exploiter des résultats intermédiaires, sans forcément aller jusqu'à la constitution d'une ontologie opérationnelle.

Ce cadre méthodologique est constitué de quatre étapes relativement indépendantes, qui s'accompagnent d'un double mouvement, du linguistique au conceptuel et de l'informel vers le formel : (a.) la constitution d'un corpus de documents ; (b.) le traitement du corpus ; (c.) la normalisation sémantique ; (d.) l'élaboration de l'ontologie opérationnelle.

a. Constitution d'un corpus de documents

La constitution du corpus est la première étape, après l'analyse des besoins. En général, les corpus sont construits en fonction de l'application et des documents disponibles. Quand ils sont construits, les corpus rassemblent un ensemble de documents attestés dans la pratique d'un domaine et pertinents pour l'application à développer. On s'attend à y trouver les expressions linguistiques des notions qu'il faut

modéliser. La nature des textes accessibles va conditionner la ressource obtenue (le contenu du modèle final) de même que le choix des outils adaptés à leur traitement.

b. Pré-traitement du corpus par des outils de TAL

Le traitement du corpus revient à systématiser et à rendre plus efficace la recherche des données conceptuelles dans les textes en utilisant des outils de traitement de la langue naturelle (TAL), par exemple des concordanciers, des extracteurs de termes candidats ou de relations candidates. Les résultats de l'analyse distributionnelle (Harris et al., 1989) permettent à certains outils de proposer des regroupements de concepts et aident à l'organisation des hiérarchies (Bourigault & Fabre, 2000). Les résultats fournis par ces outils sont en général retravaillés par le cogniticien pour en retirer les erreurs les plus évidentes. Ces traitements viennent donc faciliter l'accès au texte en fournissant des données plus élémentaires comme des candidats termes (syntagmes nominaux ou verbaux), des segments de phrases contenant éventuellement des relations lexicales, des classes de termes ayant des comportements analogues, etc. Les résultats obtenus peuvent conduire à modifier le corpus pour mieux répondre aux besoins. Il est de moins en moins suggéré de prévoir un filtrage en collaboration avec l'expert car ce type de validation, très long et coûteux, est repris lors de la phase suivante.

c. Normalisation sémantique

La normalisation sémantique consiste à organiser au sein d'un modèle conceptuel des connaissances à partir de l'interprétation des résultats précédents, de la compréhension du domaine et de l'application visée. Cela revient à associer aux termes une signification qui fasse abstraction des variations de sens liées aux différents contextes textuels dans lesquels ils apparaissent. Cette abstraction du contexte conduit à construire des concepts, considérés en tant que «signifiés non contextuels», normés au sens où ils sont décrits selon un certain point de vue (celui de la tâche, qui fixer un contexte de référence). À ce stade, on peut parler d'ontologie, qui peut être spécifiée de façon informelle (à l'aide de la langue naturelle non contrainte), semi-informelle (à l'aide de la langue naturelle contrôlée et structurée) ou semi-formelle (du formel non interprété par la machine pour réaliser des inférences mais, par exemple, pour échanger des ontologies sur Internet).

La normalisation revient à combiner des tâches portant sur l'analyse des résultats tirés des textes ou des textes eux-mêmes, et des tâches de structuration au sein du modèle. Les tâches effectuées se situent sur un axe texte/modèle et vont soit des textes vers le modèle (tâches de dépouillement), soit du modèle vers les textes (tâches de fouille ou de recherche ciblée). On peut identifier 3 types d'activités de structuration des concepts et relations au sein du modèle : des regroupements ou des généralisations, des spécialisations ou encore une analyse centrée autour d'un concept précis.

La normalisation vise à produire des résultats suffisamment systématiques pour assurer une bonne cohérence sémantique au modèle. Pour cela, elle s'appuie sur des principes empiriques et sur des critères de normalisation, par différenciation par exemple. Ces principes reviennent à justifier la place d'un concept dans l'ontologie par les relations qu'il entretient avec les autres concepts et à le différencier par ses relations ou propriétés.

d. Élaboration de l'ontologie opérationnelle

Cette dernière étape consiste à traduire l'ontologie obtenue à l'étape précédente en une ontologie opérationnelle spécifiée, soit dans un langage de programmation de bas niveau, soit dans un langage de représentation de haut niveau doté de capacité d'inférence. En fonction du langage considéré et de sa puissance d'expression, certains éléments de sens identifiés à l'étape précédente peuvent ne pas être représentés dans l'ontologie opérationnelle ou doivent être revus.

3.4.5 Les langages

Initialement, le problème de la construction des ontologies a été abordé dans l'optique de favoriser leur réutilisation, et, étudié par des informaticiens, il a été ramené à un problème d'interopérabilité, de langage et de format d'échange. C'est ainsi qu'Ontolingua, historiquement le premier outil dédié à la construction et à l'échange d'ontologies, est orienté réutilisation, par fusion et extension, d'ontologies existantes disponibles dans une bibliothèque, et autorise l'exportation d'ontologies dans différents formats. Il permet à un utilisateur, ou à un groupe d'utilisateurs, de visualiser des ontologies existantes et de construire manière coopérative de nouvelles ontologies. Il s'appuie sur un langage compatible avec le format d'échange KIF, ce qui est supposé assurer une facilité de réutilisation des ontologies, pour lequel il offre des interfaces de saisie et d'organisation des éléments de l'ontologie. (Duineveld*et al.*, 2000 ; Corcho *et al.*, 2000) ont réalisé un état de l'art exhaustif respectivement sur les langages de descriptions d'ontologie et sur les environnements de développement. Dans le cadre du projet Européen OntoWeb, un inventaire des outils et méthodes pour la construction d'ontologies est en cours.

Pour construire des ressources terminologiques, y compris des ontologies, à partir de textes, les outils de TAL et d'apprentissage sont de plus en plus utilisés. Un état de l'art assez complet des travaux dans le domaine est présenté dans le rapport de Gomez-Perez A. et Manzano-Macho D (2003). Ces outils relèvent à part entière d'une perspective TAL (cf. Ci-dessus). D'autres outils constituent de véritables plates-formes (Cerbah & Euzenat, 2000 ; Maedche & Staab 2000 ; Kang & Lee, 2001 ; Szulman et al., 2002) qui intègrent des outils de TAL et d'apprentissage pour construire des ontologies à partir de textes de la manière la plus automatique possible. Ces travaux ont en commun d'avoir recours à plusieurs logiciels et analyses complémentaires pour enrichir une ontologie d'un domaine. Ils s'appuient sur des données extraites des corpus à l'aide d'analyses linguistiques et parfois sur des ressources générales.

3.4.6 Applications et évaluation

Fondamentalement, les ontologies sont utilisées pour améliorer la communication entre des agents humains ou artificiels. Une même ontologie peut remplir différents services. Elle fournit, d'une part, des ressources conceptuelles et notionnelles pour formuler et expliciter un savoir. C'est ainsi qu'elle peut être mobilisée dans le cadre de l'ingénierie des connaissances (Charlet et *al*., 2000). Elle constitue également un cadre partagé que différents acteurs peuvent mobiliser et dans lequel ils peuvent se reconnaître. A ce titre, les ontologies ont un rôle de partage et de fédération des connaissances. Elles sont ainsi utilisées dans le domaine de la gestion des connaissances (Dieng et *al*., 2001). Elle peut aussi servir dans le cadre du Web sémantique ou pour la recherche d'informations dans la mesure où elle permet de représenter le sens de différents contenus échangés dans des systèmes d'information.

L'évaluation constitue le point faible de la discipline (*a priori*, on ne sait ni dire combien de temps il faut pour construire un modèle, ni garantir sa qualité). Des questions se posent sur le moment où intervient l'évaluation des ressources construites : des critères sont à définir autant pour les modèles intermédiaires que pour les ontologies. Cependant, on ne peut se contenter d'évaluer les résultats : nous devons pousser plus loin l'évaluation de nos outils et des méthodes proposées. Le développement d'applications industrielles, plutôt que des études sur des exemples jouets ou des cas d'école, constitue sans doute un contexte favorable au développement de méthodes d'évaluation. L'évaluation de l'utilisabilité des ressources produites serait en effet alors possible.

3.4.7 Pour conclure

Les questions posées par la construction d'ontologie, actuellement examinées en ingénierie des connaissances, sont relatives à la notion de corpus (constitution, détermination du genre textuel du corpus et la détermination des outils adaptés à ces genres, la dépendance des outils de traitement à adopter en

fonction de l'application traitée) ; à la définition de critères aidant à décider des unités linguistiques les plus représentatives d'un type de textes ; à la caractérisation des méthodes d'extraction en fonction de la nature des applications, de celle des corpus, etc. ; à la réutilisation d'ontologie et enfin à l'évaluation des ontologies produites ; aux rôles réservés aux experts. Les enjeux pour les années à venir portent sur la maintenance et l'intégration des ontologies construites et l'existence de méthodes permettant de répondre avec pertinence et rapidité aux demandes formulées par les entreprises.

3.5 Prise en compte des thèmes retenus en sciences de l'information et recherche d'information

3.5.1 Présentation

La recherche d'information (RI) et les sciences de l'information (SI) ont pour objectif de construire et d'analyser les processus permettant de répondre à des besoins et des usages de publics (utilisateurs) par rapport à des corpus (collections de documents). Ces processus ne se limitent pas à des actions de recherche d'informations ponctuelles de documents, mais englobent aussi la prise en compte d'attentes et besoins divers comme par exemple la recherche de granules d'information, la réponse à des questions précises, la détection et le suivi de sujets, des actions de veille, etc.
Les problématiques de la RI et des SI s'inscrivent dans celles étudiées par ASSTICCOT essentiellement sur les points suivants :

- le corpus est vu comme l'ensemble des documents à gérer. Cette notion peut varier en fonction de l'objet.
- La terminologie : la RI et les SI travaillent sur la médiation entre des auteurs des documents et des utilisateurs via la représentation des besoins et des contenus. La terminologie est d'abord un élément de ce dispositif de médiation.
- L'interdisciplinarité : la RI et les SI s'appuient sur des traitements statistiques et linguistiques des textes pour en obtenir une représentation adaptée aux traitements.

Depuis les années 1970, la RI s'est dotée de mesures et de moyens d'évaluation des mécanismes proposés. Il s'agit là d'un élément important du domaine.

3.5.2 Objectifs de la recherche d'information et des sciences d'information

Comme on va le voir, il existe un certain recouvrement entre les besoins entre les ressources terminologiques pour le TAL et pour la recherche d'information. Ce constat met en évidence, s'il en était encore besoin, la nécessité de développer des réflexions interdisciplinaires.

Recherche d'information

Il s'agit de répondre au besoin d'information d'un utilisateur en lui restituant les unités documentaires susceptibles de l'intéresser (Salton, 1971), (Van Rijsbergen, 1979), (Baeza et al., 1999). Ce principe se base sur la représentation des unités documentaires et des besoins sous forme de termes représentatifs pondérés. Les documents ayant une représentation suffisamment proche de la requête sont alors restitués. Les terminologies sont utilisées dans la phase de représentation d'information (indexation) et dans la phase d'interrogation (aide à la formulation et à la reformulation de requêtes). Les approches basées sur des traitements linguistiques et celles basées sur des traitements statistiques ont donné des résultats comparables dans les campagnes de test (SparckJones, 2003).

Recherche d'information multilingue

Les documents à retrouver ainsi que les requêtes peuvent être dans des langues différentes. Dans le cas particulier de la recherche d'information en langues croisées, la requête est dans une langue et les documents restitués dans une autre. Les mécanismes utilisés reposent sur la traduction de la requête ou/et des documents ou sur une représentation des informations dans un langage pivot. Des dictionnaires multilingues ou des ontologies peuvent alors être utilisés pour la traduction ou pour la désambiguïsation des termes (CLEF[4]).

Recherche de passages et recherche sur la structure des documents

Il s'agit de permettre à l'utilisateur d'accéder directement aux passages susceptibles d'être les plus pertinents. Dans le cadre de documents non structurés explicitement, ce sont de paragraphes ou des fenêtres de taille fixe qui sont restituées à l'utilisateur (Salton et al., 1994). Lorsque les documents sont structurés de façon explicite, des parties composantes des documents peuvent être restituées comme dans le cas de documents SGML [Corral 95] ou des documents XML. Le programme INeX s'intéresse aux documents XML et à des besoins d'information pouvant combiner des aspects contenus et des aspects structurels (INEX[5]). La recherche des phrases pertinentes est également étudié (TREC[6]).

Filtrage d'information

Il s'agit d'un processus dual à la recherche d'information. Le profil de l'utilisateur est mémorisé de sorte que lorsqu'un flux d'informations nouvelles (documents) est fourni au système, celui-ci puisse décider de transmettre ou pas chacune des informations à l'utilisateur. Les mécanismes utilisés reposent sur une fonction de décision et une adaptation du profil par apprentissage au fur et à mesure du traitement des documents.

Question/Réponse

Les systèmes question/réponse visent à répondre à des questions précises que se posent les utilisateurs. Les mécanismes les plus efficaces reposent sur des traitements linguistiques. Alternativement, les mécanismes de recherche d'information traditionnels permettent de restituer les passages – de taille variable - susceptibles de contenir la réponse.

Catégorisation

Les mécanismes de catégorisation permettent d'associer à chaque unité documentaire une ou plusieurs catégories, ces catégories pouvant être hiérarchisées. Ce principe est à rapprocher des mécanismes d'indexation à partir d'un langage. Les évaluations ont par exemple été réalisées sur des collections médicales (Medline avec la hiérarchie de termes médicaux MeSH ou sur des dépêches Reuters et l'ensemble des catégories associées).

Classification

Les mécanismes de classification de documents permettent de regrouper des documents qui ont un contenu sémantique proche. Le souci de classifier les documents est lié au fait que les documents qui sont pertinents pour une requête se ressemblent (Van Rijsbergen, 1979), c'est-à-dire ont des représentations

[4] http://clef.iei.pi.cnr.it:2002/
[5] www.is.informatik.uni-duisburg.de/projects/inex03/
[6] http:// trec.nist.gov/

proches. Les mécanismes de classification de documents reposent sur des principes de classification traditionnels basés sur les représentations pondérées des documents.

Recherche de la nouveauté

Il s'agit de proposer des mécanismes permettant d'éviter la redondance dans les éléments restitués à l'utilisateur en réponse à son besoin d'information et donc de détecter parmi les éléments supposés pertinents ceux qui sont nouveaux au point de vue de leur contenu. Le niveau de granularité correspond à des parties de documents (dans le cadre de TREC 2002 et 2003, il s'agit de la phrase).

Détection et suivi d'événements

La détection d'événement réfère à la recherche de thèmes nouvellement abordés dans les documents ainsi qu'à leur suivi. Le programme TDT[7] -topic Detection and Tracking- a permis d'évaluer les mécanismes proposés.

Veille stratégique scientifique et technologique

La veille stratégique scientifique et technologique a pour objet l'étude des éléments d'un corpus ciblé pour en extraire les éléments principaux et leurs corrélations : acteurs du domaine, technologie utilisée, sous domaines, évolutions, etc. .

Infométrie et scientométrie

La scientométrie s'intéresse à l'étude quantitative de la science et de la technologie en s'appuyant sur des méthodes et les techniques mathématiques, statistiques et de l'analyse des données appliquées à des corpus de données ciblés (i.e. d'un domaine).

3.5.3 Document

Un document, qu'il soit électronique ou papier est composé d'un ensemble d'unités documentaires, chacune desquelles respecte l'unicité de média (texte, son, image, vidéo). Un document multi-média est ainsi composé d'un ensemble hétérogène (au sens du média) d'unités documentaires. Cet ensemble d'unités documentaires est organisé selon une structure plus ou moins explicite (explicitée à l'aide du format XML par exemple). La notion d'unité documentaire réfère à différentes phases du cycle de vie des documents :

- lors de la création, qui réfère à l'acquisition, la création du contenu, sa mise en forme et sa matérialisation, l'auteur décide de la structuration du document en unités documentaires, éventuellement guidé par un cadre imposé (exemple : une DTD) ;
- lors du stockage qui a pour fonction de classer ou de représenter l'élément pour pouvoir y accéder, les unités documentaires peuvent être différentes de celles définies par l'auteur; en fonction de l'application et de ses objectifs le document pourra être décomposé de différentes façons ;
- lors de la production, résultant de la procédure de recherche ou d'analyse, le système décidera de la granularité la plus adaptée au besoin exprimé par l'utilisateur. Le résultat fourni par un système pourra alors correspondre à des unités documentaires stockées ou à une information élaborée à partir de plusieurs unités documentaire.

[7] www.nist.gov/speech/tests/tdt/

Pour de nombreuses applications, unités documentaires et documents se confondent. Dans ce cas, l'élément traité est le document pris dans sa globalité. Cela n'est pas sans poser des problèmes pour sa représentation et son traitement dans le cas de documents multimédias où les modèles d'indexation et de recherche et d'analyse doivent être combinés. Le cas des documents hypertextes ou hypermédias est plus complexe. Dans le cas de la toile (www), le document correspond à une unité d'information qui peut être identifiée par son URL (Uniform Resource Locator). Ainsi les hyperliens ou les contenus des éléments d'information liés à un document donné peuvent être pris en compte soit lors du stockage (indexation), soit lors de la production (recherche ou analyse).

3.5.4 Corpus et collection de test

Le terme "corpus" est peu utilisé en RI et SI : on y parle de fonds documentaires, de collections, de sources, le plus souvent au pluriel.

Corpus en recherche d'information

Pour la RI, un corpus réfère généralement à une collection de documents de test qui peut contenir des documents homogènes ou hétérogènes. L'hétérogénéité porte en particulier sur le format, sur la langue d'écriture ou sur le contenu thématique.

Le soucis de fournir des principes d'évaluation des mécanismes proposés dans le domaine a probablement conduit à cette notion. Les bases des principes d'évaluation ont été posées avec la collection de test Cranfield (Cleverdon, 1968). Selon cette approche, une collection de test est composée d'un ensemble de documents, un ensemble de requêtes et d'un ensemble de jugements de pertinence qui spécifient les documents qui auraient dû être restitués pour chaque requête. Cette collection doit répondre à un certain nombre de critères, en particulier en terme de nombre de requêtes.

Ce modèle d'évaluation a été repris et étendu dans les programmes actuels d'évaluation internationaux : TREC (Text Retrieval Conference) dont certaines collections de test contiennent plusieurs Giga octet de textes, Amaryllis (pour le français), CLEF (Cross Language Evaluation Forum), NTCIR (NII-NACSIS Test Collection for IR Systems) et INEX (Initiative for the Evaluation of XML Retrieval). Les extensions concernent les éléments pris en compte comme la langue dans laquelle les documents et/ou les requêtes sont écrites (différentes langues européennes pour CLEF, langues asiatiques pour NTCIR), les unités à retrouver (images, vidéos, parties de textes), la richesse de la description du besoin d'information (qui intègre des éléments pour permettre de décider de la pertinence d'un document et rend la requête plus proche d'un réel besoin d'information); le modèle d'évaluation lui-même reste en revanche identique. Alors que les premières collections de tests étaient de taille modeste (Cranfield contient 1400 documents et 225 requêtes), le facteur d'échelle a été pris en compte au début des années 1990 en particulier avec le programme TREC (les collections de TREC contiennent entre 500 000 et 1 million de documents et l'ensemble des documents utilisés pour une tâche varie et peut utiliser des sous parties de collections ou plusieurs collections).

Alternativement le corpus peut référer au résultat obtenu comme réponse à un besoin utilisateur. Il correspond alors à un ensemble d'unités documentaires qui ont une certaine homogénéité de contenu thématique. Il peut alors servir à des traitements tels que l'analyse pour la reformulation automatique du besoin permettant de rapprocher le point de vue des auteurs et celui de lecteurs. Une autre utilisation de ce type de corpus est l'élaboration d'information avancée (scientométrie, recherche de liens entre éléments d'information, détection de la nouveauté) (Mothe, 2000).

Corpus en sciences de l'information

En SI, l'exhaustivité, l'idée de détenir l'ensemble de la production éditoriale, reste encore (pour beaucoup) l'idéal à atteindre : cette posture est contraire à la notion de corpus, représentatif, clos. Corpus au singulier n'a que peu de réalité pour cette discipline (sauf dans les campagnes d'évaluation d'outils d'ingénierie linguistique appliqués au traitement de l'information).

Nous avons la plupart du temps à faire, comme point de départ, avec des fonds documentaires hétérogènes, ouverts, cumulatifs. Peut-on considérer que le dépôt légal constitue un "corpus"? Peut-on considérer que les listes d'acquisition dressées par les enseignants à l'intention des BU constituent des corpus ? La notion de corpus ne convient pas ni pour désigner ce que nous traitons comme matière première ni ce que nous produisons comme matière secondaire : notre activité consiste à structurer un ensemble quantitativement vaste et hétérogène en des "séries" de taille réduite, "informativement homogènes" : c'est le rôle de l'indexation que de créer des sous-ensembles de documents proches. Le terme "corpus" ne convient pas, non plus, pour désigner cette réalité-là.

Or cette réalité-là influe considérablement sur l'analyse du contenu des documents : l'information que l'on choisit de présenter n'est pas uniquement fonction du document lui-même, elle est aussi fonction des autres documents, le but étant de créer des séries de documents homogènes (en laissant provisoirement de côté la question de la recherche d'information "factuelle", qui ne pose pas du tout les mêmes questions d'indexation). Donc, la matière première à traiter changerait de nature. Reste que la nécessité de structurer cette matière de départ en entité "documentaire" de taille nécessairement réduite et de contenu "homogène" persiste. Comment désigner ces entités-là ? Le traitement documentaire doit résoudre le double problème du repérage des "bons" textes (dits textes pertinents) et de l'exploitation de ces textes (le niveau de l'"information", de l'invariant informationnel construit à travers plusieurs documents).

3.5.5 Représentation d'information : terminologie et connaissances

La RI et les SI travaillent sur la médiation entre des auteurs des documents et des utilisateurs via la représentation des besoins et des contenus des documents. Ces dispositifs de médiation s'appuient sur des représentations (représentation d'attentes ou des besoins des utilisateurs, représentation de documents, de parties de documents ou de sous-ensembles de documents) qu'il faut pouvoir mettre en relation. Les représentations mises en jeu sont construites à partir de langages, qui peuvent s'exprimer sous des formes diverses (listes, graphes, réseaux, ontologies…).

La RI et les SI ont beaucoup travaillé sur l'aspect «représentation des documents », et de manière plus générale, sur la construction de langages documentaires. Il existe deux manières d'aborder la conception de langages documentaires :

- comme structurant un domaine, et organisant les connaissances de ce domaine. Un tel langage permet de structurer le corpus en classes de documents, il permettra également un accès aux informations selon cette même structuration (langage normé),
- comme issus d'un corpus, et permettant une représentation de chaque unité sous la forme de termes représentatifs et discriminants. Les besoins d'information sont alors exprimés en langage libre.

Dans les deux cas, la démarche de conception tentera d'anticiper sur les usages et les attentes des utilisateurs.

a. Indexation des textes

- Représentation à partir d'un langage contrôlé

Lorsque le langage est contrôlé (c'est-à-dire défini), la représentation des documents ou leur indexation s'apparente à une catégorisation multi-libellée des unités documentaires (par exemple, basée sur la terminologie MeSH, sur les catégories Reuters ou Yahoo!). Cette catégorisation est effectuée manuellement par des documentalistes ou de façon automatique. Dans ce dernier cas, la catégorisation se base sur des termes discriminants des catégories (profils) et des mécanismes d'apprentissage basés sur des exemples positifs et négatifs de documents classés (Sebastiani, 2003). La mise à jour du langage d'indexation pose différents problèmes : sélection des termes à rajouter et positionnement dans le langage existant, prise en compte de cette évolution pour les documents déjà catégorisés. Il n'existe pas de mécanismes permettant une automatisation complète de tels processus mais des aides automatisées sont possibles.

- Représentation à partir du langage libre issu des documents

Le langage d'indexation est construit directement à partir de l'analyse des contenus des documents. Les étapes du processus automatique sont communément les suivantes : suppression des mots considérés comme vides pour la recherche ultérieure, radicalisation des termes (en utilisant des approches combinant linguistique et statistique), éventuellement extraction de groupes de mots et pondération des termes retenus. Cette dernière étape permet de quantifier le pouvoir représentatif et discriminant des termes retenus pour la représentation de chaque document. Cette représentation qualifiée de «bag of words » s'avère efficace lors des étapes de recherche de documents. Elle peut être complétée par des représentations multi-facettes permettant en particulier de prendre en compte les méta-données ou les différents points de vue qu'un utilisateur peut avoir sur un document. Ces derniers éléments sont indispensables dans des activités de veille.

b. Représentation du besoin d'information

Les ressources terminologiques disponibles, quelle que soit la forme qu'elles prennent (ensemble de mots liés, ontologies de domaine ou ontologie générique), peuvent s'avérer utiles pour l'utilisateur lors de l'interrogation des systèmes. Nous pouvons distinguer deux usages complémentaires :

- Reformulation automatique de requête : cette problématique vise à rapprocher la représentation du besoin exprimé par l'auteur des représentations induites par les choix des auteurs pour exprimer le contenu des documents. Typiquement, une terminologie est utilisée pour étendre la requête lorsque le nombre de mots de la requête initiale est trop faible (2-3 mots pour les recherches sur le web par des non spécialistes) ou pour modifier l'importance relative de chacun des termes de la requête. Les travaux initiaux dans ce domaine ont porté sur l'étude de la co-occurrence des termes. Des études plus récentes ont porté sur l'utilisation de ressources génériques de type WordNet.
- Prise en compte de la connaissance du domaine (par navigation en général) pour choisir les termes les plus adaptés pour exprimer le besoin d'information. Des terminologies comme MeSH ont fait par exemple l'objet de différentes études.

c. Information et connaissances

Du point de vue des SI, l'information est un objet à construire : elle n'existe pas, toute prête, dans des documents : on peut alors distinguer "donnée", "information", "connaissance". Les données comme point de départ de l'activité documentaire, l'information comme point d'arrivée de l'activité documentaire. La connaissance se situant à la fois en amont (du côté des auteurs) et en aval (du côté des utilisateurs).

L'enjeu des activités documentaires, c'est de permettre que des connaissances produites par un auteur engendrent des connaissances "nouvelles" c'est-à-dire différentes, pour les utilisateurs. Typiquement, les connaissances diffusées dans un brevet d'invention vont permettre de produire d'autres connaissances

pour les utilisateurs (connaissances se traduisant par un positionnement stratégique par exemple dans le domaine de la veille). L'information ne peut alors se réduire à la seule transmission de "connaissances" d'un point A-auteur à un point U-utilisateur.

4 PROSPECTIVE

Les réunions plénières ont permis des mises en commun à partir desquelles ont pu se dégager d'une part un socle d'éléments partagés et d'autre part, un ensemble de questionnements, dont la plupart viennent d'être rapportés dans la partie 3. Discutés, argumentés, élaborés, ces questionnements nous ont permis de dégager quatre axes de prospectives :

1. Développer et approfondir la notion de « genre textuel »
2. Prendre en compte les applications et usages pour comprendre la variabilité des méthodes, outils et types de ressources terminologiques
3. Définir des méthodes pour assurer la maintenance des ressources terminologiques
4. Se donner les moyens d'évaluer et de valider ces ressources mais aussi ces recherches.

Nous présentons chacun de ces axes dans les parties 4.1 à 4.4 en rapportant les questionnements que soulève leur étude.

4.1 DÉVELOPPER ET APPROFONDIR LA NOTION DE GENRE

La question des genres est apparue comme très fédérative lors des discussions. En effet, elle intervient dans au moins trois des disciplines concernées. La partie 4.1.1 insistera sur le point de vue de chacune des disciplines qui tient compte de cette notion et mettra en évidence les points de convergence. La partie 4.1.2 proposera des pistes de réflexion prospectives.

4.1.1 Constat à propos de la question des genres

Les points de vue et les perspectives de chacune des disciplines seront d'abord présentées puis seront dégagés les points de convergence en ce qui concerne les questionnements.

a. Linguistique de corpus

La problématique des genres textuels existe de longue date en linguistique. Comme le rappelle Branca-Rosoff, la notion de genre est utilisée depuis longtemps dans une perspective à la fois descriptive et prescriptive : si l'on voulait être reconnu comme un bon « écrivain », il fallait obéir aux règles de la bonne rédaction. Mais les siècles passant, la dimension prescriptive est débordée par la réalité des usages et il devient de plus en plus difficile de contrôler (ou décrire) l'évolution des genres :

> « La terminologie des genres, construite en vue de l'acquisition pratique des modèles a été descriptivement adéquate tant que l'institution scolaire a travaillé sur le corpus fermé des textes de la tradition. Mais une fois encore les classements se sont périmés à partir du XVIIè siècle parce que les pédagogues ont figé les catégories et ont exclu du champ littéraire les textes qui ne correspondaient pas à leur grille d'analyse alors que leur importance sociale allait croissant […]. A partir du XIXè siècle, la crise s'accroît car la modernité revendique la déstabilisation des genres … » (Branca-Rosoff,1999,18).

C'est surtout Bakhtine qui s'est intéressé à la question des genres.

> « Tout énoncé pris isolément est, bien entendu, individuel, mais chaque sphère d'utilisation de la langue élabore ses types relativement stables d'énoncés, et c'est ce que nous appelons les genres de discours » (Bakhtine,1984,265).

> « Les genres du discours, comparés aux formes de langue, sont beaucoup plus changeants, souples, mais, pour l'individu parlant, ils n'en ont pas moins une valeur normative : ils lui sont donnés, ce n'est pas lui qui les crée ».(ibid., 287).

Certains travaux ont été menés pour essayer de caractériser les genres textuels de manière consensuelle. Ainsi, deux projets ponctuels ont permis de faire le point, à deux époques différentes, et dans des domaines différents, sur des corpus existants et sur les types de recherches qui en étaient la source ou qui en sont nées. L'une s'est déroulée en 1975-1977, l'autre en 1994-1995. Un historique des recherches linguistiques sur corpus a été proposé dans le cadre d'une étude méthodologique réalisée pour le CNRS en 1975-1977. Il ne s'agissait pas prioritairement, alors, de recherches au moyen d'outils informatiques. Mais c'est bien de recherches linguistiques sur corpus qu'il était question.

Des travaux européens ont été conduits pour un corpus de référence, il s'agit des projets NERC (Network of European Reference Corpora) et PAROLE (Preparatory Action for linguistic Resources Organization for Language Engineering). D'après enquête menée dans le cadre du projet PAROLE sur les besoins des utilisateurs, le rapport EAGLES met l'accent sur un nécessaire équilibre entre critères internes et critères externes (intralinguistiques et extralinguistiques) dans la typologie textuelle. Il a été réalisé dans la proximité du projet NERC. Parmi les critères extra-linguistiques, il faut citer le sujet (topic), le genre littéraire, le support ("medium"), le mode (écrit, oral, électronique) et parmi les critères intralinguistqiues, le thème, le "style", les collocations.

b. TAL

Dans une orientation plus automatique, la question des genres a surtout été travaillée par la linguistique anglo-saxonne et plus particulièrement par D. Biber (Biber, 1988). Comme d'autres auteurs, Biber distingue le genre du texte, obtenu sur la base de critères propres à la situation de production du type du texte qui permet des regroupements sur la base de régularités linguistiques.

> « I use the term " genre " to refer to text categorizations made on the basis of external criteria relating to author/speaker purpose » (Biber, 1988, 68).
>
> « I use the term " text type " on the other hand, to refer to groupings of texts that are similar with respect to their linguistic form, irrespective of genre categories » (*ibid.*, 70).

L'objectif de Biber consiste à essayer de donner une assise linguistique à ce qui relève d'abord d'une classification intuitive. Après avoir constitué un corpus sur la base de situations de production qui semblent assez bien stabilisées, il prend en compte un grand nombre de phénomènes linguistiques et essaie de constituer des dimensions par rapport auxquelles vont s'organiser les textes du corpus. Il constate ainsi que, selon la dimension considérée, il y a plus de similarité de fonctionnements linguistiques entre X et Y qu'entre Y et Z alors que Y et Z paraissaient intuitivement proches (*i.e*, de même genre); par exemple, du point de vue de la dimension narrative, les dialogues spontanés sont plus proches des biographies que des conversations téléphoniques (Biber, 1988, 136). Comme le signale Beauvisage : « les travaux de D. Biber sont intéressants pour l'étude des genres en ce qu'ils introduisent l'idée fondamentale que l'énonciation de traits pertinents pour différencier des groupements de textes doit venir des textes eux-mêmes. Nous sommes ici dans une linguistique de corpus pour laquelle les textes doivent être le matériau des travaux d'ingénierie linguistique, l'objet, la source d'observation et non le moyen de vérifier des hypothèses » (Beauvisage 2001, 583).

Ces travaux relèvent d'une approche TAL dans la mesure où le repérage des critères se fait de manière automatique. Mais ce type d'approche est peu utilisé dans les applications du TAL au sens où très peu de travaux prennent en compte la variation pour élaborer des catégoriseurs, des analyseurs morphologiques par exemple. Quelques travaux s'inscrivent cependant dans cette optique, par exemple (Illouz *et al*., 1999) qui vise à organiser des flux de données de différentes natures (différents genres) sur la base de l'analyse automatique de certains critères linguistiques.

Dans le cas particulier du traitement automatique de la langue appliqué au domaine médical, P. Zweigenbaum a proposé des mots-clés visant à caractériser les genres textuels. Le projet global s'intéressait à un corpus échantillonné pour le français, des registres dans le domaine médical fournissant un inventaire des types d'énoncés en circulation dans cette sphère d'activité. Voici quelques uns des mots-clés repérés :

- Dossier patient (compte rendu, prescription, lettre de correspondant)
- Enseignement (livre, polycopiés, questions de cours…)
- Ressources (monographie, guide, notice de médicament, …)
- Publications
- Oral (cours, réunion d'équipe soignante, exposé…).

En pratique, le TAL prend acte aujourd'hui que les fonctionnements linguistiques varient d'un corpus à l'autre mais il est moins intéressé par le repérage ou l'identification du genre/type d'un texte que par l'adaptation des traitements à la langue particulière de ce texte. Un exemple bien connu s'inscrit dans cette perspective, c'est celui de l'étiqueteur de Brill qui se présente comme un étiqueteur par défaut à entraîner sur des données particulières (comme les systèmes de reconnaissance de la parole). C'est aussi l'une des raisons de l'intérêt du TAL pour les techniques d'apprentissage. Si le TAL n'a pas davantage avancé dans cette direction, ce n'est pas qu'il néglige ce problème de l'adaptation mais bien parce que cette adaptation est coûteuse et difficile à réaliser.

c. Sciences de l'information

Les sciences de l'information se sont intéressées à la notion de genre dans la perspective de constituer un corpus homogène ou plutôt d'organiser des collections en groupes de documents homogènes. Cette homogénéité peut être interrogée sous un double point de vue :

- les documents du corpus sont homogènes sur un plan thématique,
- ils relèvent du même genre textuel.

Du point de vue des sciences de l'information, la notion de genre devrait permettre de répondre à des questions comme : de quelle façon le texte traite-t-il du sujet ? sous quel angle ? en privilégiant quel aspect ? en s'adressant à quel type de lecteur ? dans quelle perspective ? avec quel mode d'expression ? Il s'agit en fait de prendre en compte des méta-informations de deux types : méta-informations simples et méta-informations correspondant à des jugements de valeur.

Méta-informations simples

Certaines d'entre elles sont objectives et factuelles. Il s'agit d'informations que l'on aimerait voir renseignées, et dont l'instanciation par des experts serait relativement aisée et probablement consensuelle. Ces méta-informations s'appliquent à l'ensemble du document et leur attribution est simple à effectuer (pour certaines d'entre elles une expertise devrait pouvoir être mémorisée et ré-utilisée). Nous en donnons des exemples à titre d'illustration :

- type d'environnement éditorial (presse grand public, presse fondamentale, presse professionnelle, mémoire universitaire…) ;
- type d'article (article primaire, article de synthèse, article de vulgarisation grand public, article de vulgarisation public averti, ouvrage de référence, texte didactique, texte de conférence, résumé de conférence, article technique du métier…) ;
- communauté de l'auteur (étudiant, professionnel, chercheur, enseignant, journaliste, …);
- domaine disciplinaire (informatique, physique, chimie, médecine, biologie …).

Méta-informations prenant en compte des jugements de valeur

Les jugements de valeur dont nous parlons ici ne sont pas des jugements au sens absolu (du type « ce document est bon » ou « ce document n'est pas bon ») mais ce sont des jugements de valeur relatifs à certains types d'attentes. Par exemple, si nous imaginons une grille évaluant le degré d'accessibilité à un profane, des valeurs maximales sur cette échelle correspondraient à des documents utilisables par un public n'ayant qu'un minimum de connaissances préalables sur le sujet, alors que des valeurs minimales correspondraient à des documents qui ne pourraient être utiles qu'à des lecteurs très avertis et spécialistes du même domaine que l'auteur du document.

Il faut noter que, sans que les questionnements soient profondément différents, la notion de genre apparaît à deux moments du processus de recherche d'information. D'une part, et c'est de cette étape dont il a été question majoritairement lors des discussions : au moment de la constitution du corpus qui va être remis à l'utilisateur ; corpus qui doit être homogène pour s'adapter au mieux à la question qui a été posée. D'autre part, dans une phase amont, au moment de la constitution du thesaurus. Ainsi, si l'on veut constituer un thesaurus, c'est-à-dire une norme pour faciliter la détermination des mots-clés, il est important, dans le cas où le thesaurus est constitué à partir d'un corpus, de s'interroger sur l'homogénéité de ce corpus. La difficulté est de savoir à quel type d'indexation s'adresse le thesaurus : grand public, experts, semi-experts ? La constitution du corpus devrait refléter les différences de publics visés, ce qui se traduit par la prise en compte des genres des textes mis en commun pour constituer le corpus.

4.1.2 Des besoins et des conclusions proches

Deux conclusions, apparemment contradictoires peuvent être tirées.

1- D'une part, l'intuition qui permet de penser que *les genres textuels sont corrélés avec des régularités linguistiques* semble valide comme en témoignent les trois exemples suivants :

- Une étude de commentaires de dégustation a permis de mettre en évidence une utilisation importante d'adjectifs mis en apposition. Ce modèle morpho-syntaxique reflète une pratique de la dégustation qui nécessite dans un temps restreint de décrire un ensemble de propriétés d'un produit. L'adjectif qualificatif permet de répondre à ce besoin de description, le protocole même de la dégustation structurant l'énoncé permettant de faire l'économie d'unités nominales (Normand, 2002).
- Une étude de discours de vulgarisation a montré l'importance du métalangage dans les discours destinés aux non-spécialistes. La plupart des analyses linguistiques montrent que les énoncés de vulgarisation sont marqués par le métalangage qui exhibe la « traduction » d'un vocabulaire spécialisé en vocabulaire commun. Le but de ce métalangage est d'élucider, d'interpréter le sens de mots supposés obscurs ou pouvant relever d'une autre acception, et donc nécessitant une explication, une précision, un complément d'information Ces formes peuvent être repérées par une analyse à entrée lexicale. Ce type d'analyse apparaît remarquablement fécond pour étudier les discours destinés à exprimer des savoirs. En effet, en sélectionnant les termes scientifiques pris comme « termes-pivots » et en examinant leurs cotextes, divers procédés de reformulation ont été dégagés. Les mots spécialisés servent d'une certaine façon de traceurs et permettent de repérer une activité de ce qui est en train de se faire. (Delavigne, 2003)
- Une étude en traitement automatique des langues sur le genre du roman policier (Beauvisage 2001) illustre la pertinence de travailler sur les variables morphosyntaxiques et la ponctuation pour donner une représentation de la spécificité des genres et ceci en se dédouanant des aspects lexicaux. Dans l'étude des discours, une distinction est posée entre ce qui relève des critères lexicaux pour la mise en évidence de thèmes liés à un discours et les critères morphosyntaxiques ainsi que la parataxe. Ces différents critères permettent alors de spécifier le discours de façon exploitable pour le TAL. Cependant, on peut s'interroger sur l'apport de la notion de genre au

TAL. Pour un traitement TAL, on a besoin de savoir entraîner les outils ou choisir des outils spécialisés (étiqueteurs, analyseurs), choisir et éventuellement adapter des ressources lexicales. Un jeu de critères (morphologie, lexique, syntaxe, voire sémantique) peut alors être suffisant. Ces critères peuvent d'ailleurs croiser des traits de genres et des traits de domaine. Ceci étant, disposer de catégories de genre (si elles sont stables), chacune étant associée à une certaine combinaison de critères, serait évidemment utile. Mais leur définition semble très coûteuse et on peut se demander si elle est réaliste.

Il est donc possible de corréler des situations de production de textes avec des régularités linguistiques, même si ces régularités varient en fonction du genre étudié.

2- D'autre part, *des difficultés sont apparues, assez récurrentes dans chaque discipline, remettant en cause la possibilité de définir des genres :*

- Critères de classifications variés.

Les critères de classification peuvent être de différentes natures comme le note par exemple Bronckart

> « Les genres de textes demeurent cependant des entités foncièrement vagues. Les multiples classements existants aujourd'hui restent divergents et partiels, et aucun d'entre eux ne peut prétendre constituer un modèle de référence stabilisé et cohérent […]. Cette difficulté de classement tient d'abord à la diversité des critères qui peuvent légitimement être utilisés pour définir un genre… » (Bronckart, 1996, 76).

Par ailleurs, un même corpus peut relever de différents genres. Il est courant qu'un même texte puisse comporter des parties qui relèvent d'un genre et d'autres d'un autre genre.

- Nature des critères linguistiques mis en œuvre.

Les éléments linguistiques qui sont observés peuvent être de différentes natures ; trois facteurs de variation sont apparus :

•Importance de cet élément au sein du genre lui-même : on l'a vu ci-dessus, les éléments linguistiques pertinents varient selon le genre textuel dont relèvent les textes ; chaque situation de production s'accompagne certes de régularités linguistiques mais elles sont propres à chaque situation ;

•importance de l'objectif de l'analyse : selon que l'on vise la constitution d'un corpus pour un ou l'autre objectif (vérification d'une hypothèse linguistique, constitution d'un thésaurus, description des caractéristiques linguistique d'une population…), les genres pertinents ne seront pas les mêmes.

•moyen que l'on met en œuvre pour examiner ces fonctionnements linguistiques. Ainsi, si l'on utilise une méthode automatique, il est clair que seuls les éléments ayant une forme identifiable par une machine seront examinés et pas ceux qui relèvent d'une interprétation sémantique (utilisation de tel mot avec tel sens par exemple) ou dont on ne peut circonscrire la forme a priori (phénomènes anaphoriques par exemple).

- Rôle du point de vue de la discipline.

Ce problème est lié à l'objectif d'analyse. Il suffit de revenir sur les propositions de méta-données faites dans les parties b et c, l'une relevant des sciences de l'information, l'autre du TAL, pour comprendre que le point de vue qui préside à la définition des genres est fortement lié à la discipline. Ces points de vue disciplinaires sont aussi en lien avec les objectifs et les méthodes de ces disciplines. Ils ont une influence très grande sur la façon de considérer les critères de classifications pertinents en lien avec des régularités linguistiques pertinentes. Ce double constat, à la fois de l'importance du genre textuel et de la difficulté à stabiliser les modes de définition de ces genres, nous a amenés à définir des pistes de réflexion pluridisciplinaires.

4.1.3 Pistes de réflexion pluridisciplinaire sur la notion de genre

La dimension pluridisciplinaire de la réflexion devrait donner un cadre de réflexion adapter pour mieux cerner la notion de genre et la possibilité de la systématiser. Quatre pistes semblent prometteuses.

a. Critères de classement pour les situations de production

Certains critères de classement semblent être intéressants pour toutes les disciplines. Une réflexion conjointe serait nécessaire pour définir ces critères généraux : par exemple, date et lieu de production, niveau de compétence du/des rédacteur/s, objectif initial, public visé mais aussi, mode de présentation… …

Certains critères au contraire n'ont de pertinence que pour une discipline, par exemple, mode d'édition pour les sciences de l'info, qualité de la rédaction pour le TAL…

Certains critères sans doute ne sont pertinents que de manière très ponctuelle : par exemple, nature des relations personnelles entre le producteur et le destinataire.

2- Critères de classement en termes de besoin

Comme on l'a vu, la nature de l'analyse qui va être menée peut avoir une influence majeure sur la façon dont on va considérer les textes et le genre dont ils relèvent et surtout, la manière dont on va les interpréter.

Anne Condamines propose la notion de genre interprétatif (Condamines, 2003a), c'est-à-dire une notion qui, parallèlement à la notion de genre textuel, permettrait d'identifier des modes d'interprétation récurrents, en lien avec des objectifs d'interprétation. Plusieurs pistes pourraient être suivies.

- Repérer des classes d'utilisation de ressources terminologiques et les caractéristiques des textes, dans toutes les disciplines représentées dans ASSTICCOT, un peu sur le mode de ce qu'ont fait la communauté TAL (cf.3.3.2) et les SI (3.5.2).
- Dans un second temps, il faudrait voir si ces classes d'utilisation sont au moins compatibles d'une discipline à l'autre et voir si on peut identifier de grandes classes de besoins, c'est-à-dire de grandes classes de genres interprétatifs.
- Il faudrait ensuite évaluer si ces objectifs interprétatifs peuvent être mis en relation avec des régularités linguistiques.
- Enfin, il faudrait évaluer comment le genre textuel et le « genre interprétatif » se combinent afin d'essayer de trouver un mode de balisage des textes qui prendrait en compte cette double influence. Cela permettrait aussi de caractériser les textes en fonction des besoins.

c. Définition ou adaptation de méthodes TAL pour aider à dégager des régularités linguistiques

Certains outils proposés par le TAL ou la recherche d'information devraient pouvoir être utilisés pour aider à la mise au jour de régularités linguistiques, dans une perspective proche de celle proposée par Biber. Certes, ces méthodes ont leurs limites (seuls des critères basés sur la forme peuvent être testés) mais dans un certain nombre de cas, ces approches permettraient de tester rapidement des hypothèses, ou d'identifier rapidement des proximités entre des textes, quitte à les affiner et/ou à les faire suivre d'une analyse plus manuelle.

4- Prise en compte par le TAL des genres textuels

La notion de genre est certainement pertinente pour le TAL afin d'adapter les outils. Reste à savoir comment elle peut être prise en compte. Tout un courant de recherche depuis une dizaine d'années a visé à l'intégration de méthodes d'apprentissage dans les outils de TAL. Une des raisons est que l'on souhaite avoir des outils relativement faciles à adapter pour de nouvelles applications. Des efforts ont porté sur l'étiquetage, l'analyse et les ressources lexicales. Ce travail a été réalisé pour permettre de passer d'un domaine à l'autre, d'une langue générale à une langue de spécialité et donc finalement d'un genre de texte à un autre. Il faut désormais envisager l'adaptabilité au genre comme un objectif à part entière.

4.2 PRENDRE EN COMPTE LES APPLICATIONS ET USAGES POUR COMPRENDRE LA VARIABILITÉ DES MÉTHODES ET OUTILS

4.2.1 Motivation

Comme nous l'avons rapporté dans la partie 3.1, un des cadres unificateurs de nos recherches est méthodologique. L'expérience des groupes de travail antérieurs a permis de faire ressortir les convergences des différentes démarches mises en œuvre pour associer des corpus et des ressources terminologiques, que ce soit pour les construire ou les utiliser. Finalement, parmi tous les types de ressources déjà cités, la différence ne vient pas seulement de leur structure ou de leur support. Leur diversité se justifie essentiellement par l'usage pour lequel la ressource est prévue. Or la nature de l'application, ou l'usage, a de multiples impacts, qui vont du choix des outils ou des techniques d'exploration de corpus à appliquer jusqu'à l'interprétation de leurs résultats ou encore la structure de la ressource construite. La pertinence des données pour l'utilisateur final préside au choix des occurrences pertinentes, à l'organisation des connaissances en un modèle, qui doit comporter toutes les connaissances nécessaires et uniquement celles-là.

La constitution à partir de textes de ces ressources requiert donc de définir un cadre méthodologique à la fois unificateur et différencié, situant l'usage d'outils et techniques de traitement de la langue. En nous appuyant sur diverses expériences, nous avons pu montrer comment la nature de l'application visée conditionne chacune des étapes de ce processus, depuis la constitution du corpus jusqu'à la structuration des connaissances. L'influence sur l'adaptation ou le choix de techniques et logiciels est plus complexe à définir. Elle nécessite de l'expérience dans la construction de modèles de types différents, et relève tout à fait du projet de recherche d'un groupe comme ASSTICCOT bien plus que de recherches individuelles.

Aussi, ce problème a été retenu comme sujet d'étude du groupe TIA en 2002 et 2003. La réflexion a été poussée plus loin au sein d'ASSTICCOT, où elle a été menée en impliquant plus de disciplines, en se référant à une gamme très large de ressources terminologiques et ontologiques. Nous reprenons ci dessous les différents temps de cette réflexion :

- confirmation de la variabilité des ressources selon les applications qui les utilisent;
- lien avec le débat sur la généricité des terminologies et des ontologies;
- repérage d'un premier ensemble d'indices associant type de modèle et méthode;
- identification plus fine de la nature des choix en fonction des applications ciblées pour une sélection d'applications particulières.

Ces analyses sont reportées dans deux articles, l'un présenté au nom du groupe TIA lors des assises du GRD I3 en déc. 2002 (Aussenac-Gilles *et al.*, 2002) et l'autre à paraître dans la revue RIA (Bourigault *et al*, 2004).

4.2.2 Variabilité des ressources en fonction de leurs applications

Nos analyses s'appuient sur plusieurs applications menées au sein des différentes équipes de l'AS, dont nous mentionnons quelques unes ci-dessous pour illustrer cette diversité:

- constitution d'index de documents sur papier ou électroniques (Ait el Meikki et al., 2002)
- constitution de thésaurus ;
- construction de terminologie structurée pour guider le pilotage de programme (Baudouin et al., 2003)) ;
- construction de BCT pour favoriser la communication et la collaboration entre équipes (Aussenac-Gilles et al., 2001) ;
- langage documentaire pour l'expression de requêtes ou l'indexation de documents dans des bases documentaires (Lainé, 2001) ;
- construction de modèles conceptuels et d'ontologies pour la veille technologique et la catégorisation de documents (Aussenac-Gilles et al., 2003a) ;
- construction d'ontologies pour l'analyse du langage naturel
- construction de modèles pour la recherche d'information dans des textes (Séguéla, 2001) ;
- construction de hiérarchies de termes pour définir des catégories de documents ou organiser la recherche au sein d'une base de documents (Mothe, 2002) ;
- extraction de mots clés pour le référencement de pages Web;
- utilisation d'une base de données lexicale pour l'expansion de requêtes (Baziz et al, 2003);
- comparaison diachronique de terminologie (Condamines *et al.*, 2003), (Aussenac-Gilles *et al.*, 2003c).

Comme nous l'avons souligné dans les enjeux technologiques, les corpus et les ressources terminologiques jouent un rôle charnière au sein de ces applications. En effet, elles supposent de construire ces ressources à partir de corpus puis de les utiliser ensuite pour accéder à des connaissances dans ces textes ou d'autres documents. L'adéquation ressource-corpus est donc déterminante pour garantir l'usage de l'application, et doit rester valide tout au long de son utilisation.

4.2.3 Variabilité versus généricité

La recherche de l'efficacité et de la rationalisation de la production incite à constituer des ressources terminologiques les plus génériques et réutilisables possibles. Elle tend également à imaginer des outils génériques pour les construire à partir de textes, des logiciels de TAL basés sur des principes universaux qui seraient applicables sur tout corpus. Enfin, les besoins opérationnels d'interaction entre systèmes ou entre un système et des utilisateurs pousse à produire des connaissances consensuelles, favorisant l'interopérabilité. Or la pratique, d'une part, et certaines bases théoriques linguistiques sur le lien entre le sens des mots et leur usage, d'autre part, montrent que cette généricité va à l'encontre de la bonne adéquation à un usage particulier.

Reconnaître cet état de fait traduit une position tranchée par rapport au courant fondateur de la terminologie autant qu'avec les raisons d'être des ontologies en intelligence artificielle, voire même avec la vocation très normative des langages documentaires. Pour illustrer le débat, nous partirons du domaine de la terminologie, car il a ensuite fortement influencé tout le courant anglo-saxon qui a conduit aux ontologies actuelles.

Rappelons que le terrain théorique, en terminologie, a longtemps été occupé par la *Théorie Générale de la Terminologie.* Cette théorie, fondée par Wüster à la fin des années trente, est née dans le courant positiviste de l'entre-deux guerres et dans la mouvance du Cercle de Vienne. Elle défend une vision

unificatrice de la connaissance : le monde de la connaissance est découpé en domaines stables, dont chacun est équivalent à un réseau fixe de concepts, les termes étant les représentants linguistiques de ces concepts. Or le rapprochement récent entre la terminologie et l'informatique vient bouleverser cette position. On s'est intéressé à la conception de bases de données terminologiques susceptibles d'aider les traducteurs professionnels dans les tâches de gestion et d'exploitation de lexiques multilingues. Les réflexions ont porté essentiellement sur le format de la fiche terminologique : à l'aide de quels champs décrire un terme dans une base de données qui sera utilisée par un traducteur humain? Mais elles ont vite glissé vers un problème fondamental : la description du terme et de son sens sont-elles universelles ? uniques et adaptées a priori à toutes sortes d'usages.

Depuis la fin des années 90, la terminologie classique voit donc les bases théoriques de sa doctrine ainsi que ses rapports avec l'informatique ébranlés par le renouvellement de la pratique terminologique que suscite le développement des nouvelles applications de la terminologie. La production de documents sous forme électronique s'accélère sans cesse, et à ce phénomène, s'ajoute une diversification des besoins. La gamme des produits à base terminologique nécessaires pour y répondre à ces besoins s'élargit, ce qui met à mal le principe théorique de l'unicité et de la fixité d'une terminologie pour un domaine donné, ainsi que celui de la base de donnée terminologique comme seul type de ressource informatique pour la terminologie. Le constat de la variabilité des terminologies s'impose: étant donné un domaine d'activité, il n'y a pas *une* terminologie, qui représenterait le savoir sur le domaine, mais autant de ressources terminologiques ou ontologiques que d'applications dans lesquelles ces ressources sont utilisées. Selon l'application, ces ressources peuvent différer sensiblement quant aux unités retenues et à leur description. L'ensemble de ces constats empiriques entraîne des changements en profondeur de la pratique terminologique, et appelle du même coup à un renouvellement théorique de la terminologie. Ce renouvellement est sollicité aussi sur des bases philosophiques et épistémologiques. Il s'est produit de façon concomitante et parallèle à l'émergence, dans le domaine de l'Intelligence Artificielle, de la notion d'ontologie.

4.2.4 Premiers paramètres liant type de ressources et méthode

Finalement, l'objectif d'usage de la ressource terminologique a de multiples conséquences sur l'ensemble du processus qui va des corpus d'origine à l'application en passant par la constitution d'un produit terminologique. Parmi les nombreux paramètres qui entrent en jeu dans sa construction, nous avons identifié les textes disponibles, les acteurs intervenant dans le projet (linguistes, experts du domaine, analystes ou cogniticiens), les utilisations et les utilisateurs visés. Ces paramètres sont déterminants dans le choix de logiciels adéquats pour l'analyse des textes autant que pour la structuration et la représentation des connaissances. Chaque facteur doit être pris en compte pour espérer construire des produits utilisables et utilisés. Or peu d'auteurs ont essayé d'expliciter la part de ces facteurs dans l'évaluation de leur approche ou de leurs résultats. Nous avons donc cherché à repérer, à partir d'expériences précises, des éléments pouvant guider, en fonction de l'application ciblée, le choix de ces paramètres.

Pour focaliser la réflexion, un premier travail a été conduit sur la base de trois expériences concrètes de construction de ressources terminologiques à partir de textes, réalisées par des équipes différentes mais ayant adopté un même cadre méthodologique et utilisé les mêmes logiciels. Le cadre méthodologique et la démarche suivie relèvent à des degrés divers des propositions promues par B. Bachimont (Bachimont, 2000) et le groupe TIA. Ce cadre est aussi celui des méthodes de construction d'ontologie associées aux logiciels DOE (Troncy et al., 2002) et Terminae (Szulman et al., 2002).

Le retour de ces expériences a permis d'enrichir les éléments théoriques, méthodologiques et logiciels pour la tâche de construction de ressources terminologiques ou ontologiques à partir de textes. Plus concrètement, nous avons montré comment la spécification de l'application cible dans laquelle doit être intégrée la ressource détermine les choix méthodologiques à tous les stades de la tâche de construction de cette ressource. Nous avons cherché en quoi l'utilisation ciblée des outils facilite le recueil de

connaissances et la modélisation de ressources adaptées. Nous avons ensuite identifié en parallèle les problèmes fondamentaux qui se posent du fait de la spécificité de l'usage qui sera fait de la ressource construite et, lorsqu'elles existent, les solutions, techniques ou théoriques, qui peuvent être envisagées.

L'impact de l'application cible sur le processus de modélisation a été examiné à travers les points suivants : profil du « constructeur », construction du corpus, choix et manière d'utiliser les outils de TAL, choix et utilisation des outils de modélisation. Sur chacune de ces dimensions, nous avons mis en évidence les points communs et les divergences entre les projets, les propositions et les limites des travaux actuels.

4.2.5 Impact des applications ciblées sur les choix méthodologiques et logiciels

L'impact de l'application cible sur le processus de modélisation a été examiné à travers les points suivants : profil du « constructeur », construction du corpus, utilisation des outils de TAL, utilisation des outils de modélisation.

a. Le profil de l'analyste

Dans l'idéal, la personne chargée de construire la ressource, que nous appellerons «analyste », devrait avoir à la fois des compétences métier, des compétences en modélisation des connaissances et en linguistique et des compétences en informatique. Un oiseau rare ? Dans la réalité, il faut mettre en place une collaboration entre acteurs de spécialités différentes et toute une gamme de situations peuvent être rencontrées. Pour les applications à forte dimension cognitive, l'expérience montre que l'efficacité maximale peut être atteinte quand la ressource est construite par un spécialiste métier, passionné par les problèmes de langue et de connaissance ou formé à ceux-ci. Il doit avoir la volonté d'acquérir les compétences minimales requises en modélisation et en analyse linguistique. Il doit bien comprendre les spécifications de l'application cible et être capable de dialoguer avec les informaticiens qui la développent. Ce type de situation a un coût d'autant plus élevé que le temps de cet expert est précieux. A l'opposé, certaines applications, de type documentaire, ne requièrent pas une implication forte des spécialistes. Ou encore, les spécialistes métier ne sont ni assez disponibles ni assez impliqués dans la réalisation de l'application pour y consacrer du temps. La ressource est alors réalisée par des personnes ayant le profil et l'expérience de documentaliste, de linguiste ou de terminologue. L'expert intervient alors uniquement pour des séances de validation. Une intervention extérieure sur la terminologie et les concepts d'un domaine est aussi un moyen d'y apporter un regard neuf et relativement neutre, celui sui ressort de l'usage de la langue, et qui peut différer du point de vue particulier d'experts. Dans tous les cas, l'intervention d'un analyste médiateur est nécessaire quand l'application exige la participation de plusieurs spécialistes.

b. Construction du corpus

Cette tâche de construction du corpus est à la fois primordiale et délicate. Son enjeux est souligné dans la partie 4.1 « genre textuel ». D'une part, le corpus est la source d'information essentielle et, d'autre part, le corpus restera, une fois le processus achevé, l'élément de documentation de la ressource construite. Il doit donc être composé avec un maximum de précautions méthodologiques. Dans ce domaine, il n'est hélas pas possible de définir *a priori* des instructions méthodologiques très précises. Au-delà des problèmes techniques ou politiques de disponibilité des textes, cette collecte doit se faire avec l'aide des spécialistes et en fonction de l'application cible visée. En effet, les spécialistes permettent de s'assurer que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure de la part d'utilisateurs. Ils sont également capable de juger de l'adéquation du niveau du vocabulaire présent dans les textes par rapport aux besoins : mots-clés, méta-discours sur le domaine, ou au contraire, vocabulaire du domaine, qu'il soit de vulgarisation ou spécialisé. Par ailleurs, une boucle de rétroaction

doit être prévue, au cours de laquelle une première version du corpus sera modifiée et enrichie en fonction d'une première d'analyse des résultats fournis sur cette version initiale.

Le critère de la taille est évidemment important, même s'il est impossible de donner un chiffre idéal pour les différents types d'applications. En règle générale, pour constituer une terminologie, on a le souci d'obtenir une couverture assez large par le corpus du domaine abordé pour ne pas découvrir, une fois l'analyse du corpus effectuée, que certains aspects importants ont été oubliés. Cette tendance vers une augmentation de la taille du corpus est encouragée quand on sait pouvoir disposer d'outils de TAL pour le dépouiller de façon efficace. Mais il convient de résister à cette tendance, en gardant à l'esprit que le corpus doit rester suffisamment petit ou redondant pour pouvoir être appréhendé de façon globale par l'analyste, même à l'aide de logiciels. Une fourchette entre 50 000 et 200 000 mots semble raisonnable pour une ontologie par exemple. Le problème de la taille est directement lié à celui de l'homogénéité. L'homogénéité est requise pour permettre de repérer des régularités d'usage de la langue et des concepts. Mais, dans la majorité des cas, le corpus est hétérogène dans le sens où il a été constitué en rassemblant des textes d'origine variée. Il est alors nécessaire de procéder à un balisage du corpus qui permettra aux outils d'analyse, ainsi qu'à l'analyste, de repérer les différents sous-corpus pour procéder éventuellement à des analyses contrastives.

c. Outils de TAL

Tous les outils et types d'outils présentés dans la partie 3.3.1., dits «d'aide à la construction de ressources terminologiques », ont été définis dans le cadre de collaborations entre des spécialistes du TAL et linguistes d'une part, et des terminologues ou des ingénieurs de la connaissance d'autre part. Leur mise au point suppose d'enchaîner plusieurs traitements élémentaires d'analyse et d'étiquetage des textes, et de développer des interfaces pertinentes pour en dépouiller les résultats efficacement.

La variabilité des applications est alors un obstacle à une large utilisation de ce type de logiciel pour construire tout type de ressource. Il est clair que chaque type de ressource va requérir de pousser plus ou moins loin certaines analyses pour parvenir à un modèle plus ou moins riche. Par exemple, il peut être indispensable de disposer de toutes les formes lexicales synonymes dans une terminologie, alors que cela pourrait être inutile dans certaines ontologies qui, au contraire, requièrent le repérage de relations binaires ou de règles associant plusieurs concepts. Or il est difficile de prévoir, au sein d'un même logiciel, une réponse sophistiquée et pertinente à toutes la gamme des besoins possibles.

Dans un souci de réutilisabilité, il ressort donc deux types de besoin portant sur deux types de logiciels très différents, et soulevant des problématiques complémentaires :

- répondre à des besoins ad hoc, sophistiqués, soulevant des enjeux de taille, comme le suivi terminologique sur des projets technologiques à très long terme et très coûteux, les systèmes de veille technologique associés à du classement documentaire propre à une entreprise, etc. Dans ce cas, on souhaite disposer de logiciels effectuant des traitements élémentaires pouvant être combinés facilement pour définir une nouvelle chaîne de traitements spécifiques. L'adaptation au nouveau besoin suppose la mise au point d'une nouvelle application TAL. Il faut prévoir des collaborations entre les demandeurs, spécialistes métier, des spécialistes de TAL et des spécialistes de l'ingénierie des connaissances. Ce type d'application sera d'autant plus simple à développer que l'on pourra disposer de modules élémentaires « génériques » et faciles à combiner pour le traitement automatique des textes, comme des étiqueteurs, des analyseurs syntaxiques, des outils qui facilitent la désambiguïsation, qui calculent des synonymies, etc. ainsi que de ressources, génériques ou par domaine, comme des stop-liste ou des dictionnaires de synonymes ou des ressources morpho-syntaxiques.

- anticiper la réponse à des classes de besoins qui deviennent de plus en plus partagés, comme la recherche d'information dans un domaine particulier, la construction de terminologie ou d'ontologie, la définition d'un langage documentaire dans un domaine donné, l'indexation de documents dans ce domaine, etc. Dans ce cas, les besoins commencent à être bien identifiés et les recherches en cours peuvent commencer à définir des principes communs de construction de ces ressources à partir de textes, ou encore l'association entre textes et ressources. Pour chacun de ces besoins, la collaboration pour la mise au point des outils peut avoir lieu en amont, entre spécialistes de l'ingénierie des connaissances et spécialistes du TAL, de manière à définir des plate-forme et des méthodes particulièrement adaptées à ces besoins. Dans ce cas, les logiciels de TAL proposés seront de haut niveau, plus complexes et produisant des résultats plus directement exploitables pour construire une ressource, comme c'est le cas par exemple de Syntex. L'utilisateur doit disposer d'une interface de consultation et de validation l'orientant vers le type de résultats qu'il recherche. C'est au niveau de la mise au point de ces environnements de haut niveau que peuvent être réutilisés des modules plus élémentaires ou même des approches sophistiquées, comme l'analyse distributionnelle, avec des guides pour en exploiter au mieux les résultats. Cette complexité peut rester transparente à l'utilisateur, ou alors la complémentarité des différents modules doit être explicitée. Ainsi, la plate-forme IndDoc propose une gamme d'outils et prévoir une méthode pour les utiliser de manière complémentaire pour la construction d'index (Aït el Mekki et al., 2002).

d. Outils de modélisation

Dans tout projet de modélisation des connaissances à partir de textes, il est particulièrement crucial de pouvoir utiliser conjointement des outils de TAL et un outil de modélisation, de pouvoir définir aisément des concepts à partir de termes jugés « à retenir » ou de pouvoir revenir à des résultats bruts ou aux textes à partir d'un modèle. Le travail de modélisation n'est ni linéaire ni direct. Il requiert de naviguer dans le réseau terminologique autant que de parcourir le modèle, et de pouvoir passer de l'un à l'autre en conservant les contextes en cours d'étude. C'est pourquoi il est nécessaire de travailler avec des interfaces qui permettent, d'un côté, de consulter, de façon la plus efficace possible, la masse des résultats fournis par les outils de TAL et, de l'autre côté, de procéder à la construction d'un modèle de connaissance plus ou moins formel. Cette double contrainte s'impose d'abord pour des raisons d'efficacité dans le travail d'analyse et de modélisation et, ensuite, pour maintenir une trace, au sens informatique, entre les textes, les extractions linguistiques et le modèle de connaissance construit. Des exemples de ce type d'interface sont GEDITERM (Aussenac-Gilles, 1999), TERMONTO et TERMINAE (Szulman *et al.*, 2002), qui reposent sur des bases techniques différentes. Chacune de ces interfaces a ses points forts : l'intérêt de GEDITERM est de permettre de construire des bases de connaissances terminologiques à l'aide d'une interface graphique et de gérer leur sauvegarde en conservant les composants lexicale, conceptuelle et textuelle ; en tant qu'interface de consultation des résultats des outils SYNTEX et UPERY, TERMONTO présente un ensemble de fonctionnalités puissantes pour la navigation dans ces résultats ; la force de TERMINAE réside dans les possibilités de modélisation des connaissances, à des degrés de formalisation divers, et dans la méthodologie associée (Aussenac-Gilles *et al.*, 2003b).

Pour un projet donné, le choix de l'interface dépend de plusieurs sortes de critères :

- ceux liés au type de la ressource à construire comme le degré de formalisation souhaité, la nature des éléments de connaissances qui la composent (fiches terminologiques, simple liste de termes, réseau conceptuel, etc.) ou encore les liens entre ces éléments de connaissances ;
- des besoins en logiciels d'analyse de textes : suivant qu'ils soient nécessaire ou non, efficaces sur le corpus ou non, qu'ils doivent être intégrés au sein de la plate-forme ou pas, qu'il soit possible de récupérer leurs résultats dans la ressource à construire ou pas, etc. ;
- ceux liés aux compétences du constructeur et au degré d'automatisation ou d'intervention humaine souhaités : certains outils de modélisation formelle d'ontologie requièrent des compétences en

logique et offrent des formalismes très expressifs, alors d'autres environnement sont plus destinés à être utilisés par des linguistiques ou des terminologues et d'autres enfin par des spécialistes du domaine à décrire.

4.2.6 Bilan

Les conclusions de ce travail concernent plusieurs aspects de la recherche:

- la méthode d'étude de ce problème ;
- les enjeux en matière de disponibilité et d'intégration des logiciels ;
- les évolutions nécessaires des structures de représentation des ressources terminologiques.

Au niveau de la méthode à retenir pour ce type d'étude, nous mesurons le chemin qu'il reste à parcourir pour passer de la mise en parallèle de quelques retours d'expérience à la définition d'une méthodologie générique en acquisition de connaissances à partir de textes et à la présentation de résultats illustrant de manière convaincante les retombées pratiques des recherches dans ce domaine. Nous constatons les limites d'une démarche expérimentale : beaucoup de recherches doivent encore être effectuées pour affiner des critères d'évaluation. Le travail considérable à mettre en œuvre pour élaborer un produit terminologique et le peu de réutilisations envisageables ne rendent pas facile la reproduction d'expériences. Néanmoins, nous pensons que les progrès ne pourront désormais venir que de la confrontation des problèmes rencontrés et des solutions choisies dans des expériences concrètes, face à des communautés d'utilisateurs. En effet, le terrain théorique est désormais relativement bien balisé, grâce, entre autres, aux travaux de B. Bachimont, même si des confrontations plus poussées avec les travaux philosophiques sur les ontologies formelles pourraient être envisagés. Ensuite, il existe maintenant bon nombre d'outils d'analyse de corpus qui permettent d'extraire des corpus tel ou tel type d'information. Il existe aussi plusieurs interfaces de modélisation qui chacune privilégie tel ou tel aspect, important, de la tâche de modélisation. Enfin, on sait que toute tâche de modélisation des connaissances doit coordonner une approche ascendante basée sur l'exploitation des résultats des outils d'analyse de corpus et une approche descendante guidée par l'application cible et l'organisation générale du domaine.

Le verrou se situe désormais au niveau de l'*intégration* de ces résultats dans des tâches réelles de construction de ressources terminologiques ou ontologiques à partir de textes, et les avancées sont conditionnées par la mise de ces résultats à l'épreuve de la pratique. Il faut alors sortir d'un cercle vicieux puisque des expérimentations en vraie grandeur ne peuvent être menées à bien que si les outils et interfaces sont arrivés à un stade de maturité assez élevé. L'une des principales difficultés est liée à la largeur du spectre des types de ressources terminologiques ou ontologiques envisageables. Une autre difficulté importante est le manque de modularité de certains logiciels. Il semble prometteur de prévoir des outils très simples, dont on sait clairement quelles sont les entrées et les sorties, quelle peut être leur localisation possible et leur apport dans une chaîne de traitements. Après avoir misé sur la possibilité de réaliser des outils génériques, il faut maintenant étudier comment la pertinence et le mode d'utilisation de ces outils sont conditionnés par la structure de la ressource à construire et donc *in fine* par l'usage prévu de cette ressource. Quel que soit le contexte, le problème le plus important à gérer restera celui de la masse (la quantité d'information que les outils d'analyse peuvent d'extraire des corpus est parfois énorme) qui génère un problème de temps. La construction d'une ressource exige du temps expert. Il faudra alors bien arriver à montrer que le coût de ce temps expert est largement compensé par une augmentation importante de la qualité et de l'efficacité des systèmes exploitant les ressources.

Le coût de développement de ressources terminologiques et ontologique pose également la question de leur évolution. La nature même de ces structures doit être revue pour mieux prendre en compte de nouveaux besoins, tels que la facilité de mise à jour et le lien étroit avec les textes. Il semble indispensable de se tourner à l'avenir vers des structures de données plus souples, dont la mise à jour serait plus dynamique, qui permettent de revenir aux sources de connaissances (les textes) et de

reconstruire le sens, voire les modèles, en fonction des usages. L'idée est de réduire les coûts en évitant de reconstruire une ressource pour chaque application dans un domaine donné. Au lieu de repartir uniquement de corpus pour faire une ressource complètement nouvelle, il s'agit de favoriser la réutilisation par l'adaptation de ressources existantes. Ces structures devraient donc laisser une part active à l'utilisateur et ne pas figer le sens définitivement. Ce point sera repris et approfondi dans la partie 4.3 de ce rapport.

4.3 Définir des méthodes pour assurer la maintenance des terminologies

4.3.1 Les ressources terminologiques : objets stables ou en évolution ?

Les ressources terminologiques et ontologiques ont jusqu'ici eu vocation à rendre compte de connaissances considérées comme suffisamment consensuelles et stables pour pouvoir être partagées au sein de différentes applications et par plusieurs utilisateurs, ou pour être mises à disposition au sein de communautés professionnelles, scientifiques ou techniques. Or les conditions même de leur utilisation, que ce soit la recherche d'information, le « Web Sémantique », la gestion des connaissances des entreprises ou la veille technologique pour ne citer que quelques applications, sont mouvantes, évoluent rapidement et avec elles, le vocabulaire et les conceptualisations en jeu. Les évolutions ont deux origines :

- elles correspondent à une adaptation à de nouveaux usages : les besoins des utilisateurs évoluent ou bien ils les expriment par un vocabulaire différent ;
- les domaines et objets traités par les applications en lien avec les ressources évoluent : par exemple, le système qui utilise la ressource doit accéder à de nouvelles sources à consulter ou de nouvelles données, comme les textes sur lesquels se fait une recherche d'information, les objets techniques et sociaux auxquels se réfère une terminologie, ou encore les collections de documents à indexer.

Les ressources terminologiques et ontologiques sont donc l'objet d'un paradoxe : prévues pour figer des connaissances et des usages (dont elles rendent compte en partie), elles doivent donner accès à des connaissances qui évoluent dans des contextes dynamiques. Supposées fournir une norme ou une référence stable, elle peuvent être remise en question faute de répondre aux besoins des utilisateurs. Le risque est qu'elles deviennent rapidement obsolètes ou inadaptées. Le problème se pose pour toutes les disciplines ayant participé à l'AS. Ainsi, on peut rapprocher le problème de la partie lexicale des ontologies de celui de la maintenance des thésaurus ou celui de l'actualisation des dictionnaires : les nouvelles variations introduites, avec leurs définitions, ne doivent pas perturber la cohérence de celles déjà consignées.

4.3.2 Enjeux et identification des évolutions au sein des ressources

Les enjeux sont de taille. Par exemple, mettre à disposition auprès d'un utilisateur un outil lexical daté décrédibilise la recherche d'informations au sein d'une collection. Au contraire, autoriser une indexation automatique ou encore une indexation libre par n'importe quels termes fournis par les utilisateurs développe une inflation de termes en entrée qui complique les recherches par la suite. Enfin, on n'est plus sûr de retrouver les bons documents avec d'anciens descripteurs. Les sciences de l'information sont sans cesse confrontées au problème récurrent de l'évolution des thésaurus. Le cas le plus commun est celui d'une variation lexicale au sein du thésaurus alors qu'au niveau conceptuel, on constate une certaine stabilité structurelle.

On peut espérer que leur maintenance régulière permette de conserver leur validité et leur pertinence. Mais se posent alors plusieurs types de problèmes :

- *Comment repérer des besoins nouveaux qui justifient des mises à jour ?* Les évolutions peuvent avoir trois origines finalement : évolution des usages (à travers les textes) ; évolution des ressources ; évolution des besoins. Les études diachroniques de corpus en linguistique, à l'aide de statistiques sur l'utilisation du lexique, d'analyses distributionnelles ou de marqueurs, devraient fournir des indicateurs des évolutions des usages dans le temps (Condamines et al., 2003). Une évolution souvent observée grâce à des analyses linguistiques est par exemple l'utilisation d'un terme désignant la fonction d'un objet pour désigner cet objet. De nombreuses recherches restent à mener sur les indices de la variation en diachronie. Ce problème pose la vraie question de la variation du sens. Un autre type d'évolution parfois nécessaire au sein d'un thésaurus est de prendre en compte un autre registre de langue, par exemple passer d'un registre technique ou scientifique à un registre « grand public ». Or ces termes « grand public » ne sont pas forcément dans les textes techniques retenus au départ, et doivent pouvoir être rattachés aux termes techniques et aux textes.
- *Sur quoi les mises à jour doivent-elles porter ?* Toute une gamme de phénomènes peuvent être observés dans l'évolution du lien terme-concept : introduction de nouveaux termes pour désigner des concepts anciens, glissement de sens de termes déjà identifiés, introduction de nouveaux concepts avec ou non de nouveaux termes associés, etc. Au-delà de l'ajout, de la suppression ou du renouvellement d'étiquettes lexicales, les évolutions de ressources ont des enjeux structurels. Ainsi, en sciences de l'information, un ensemble de documents forme une classe et l'organisation en classes et sous-classes suit une certaine logique, un point de vue reflété au niveau lexical. La mise à jour du langage pose des problèmes de cohérence avec l'organisation des collections.
- *Comment conserver une cohérence entre la ressource et ses différentes utilisations ?* Ainsi, en sciences de l'information, mettre à jour un thésaurus pose ensuite la question de la mise à jour des documents indexés à l'aide des termes de ce thésaurus. Pratiquement, on se trouve face à des doubles indexations (avec les 2 versions du thésaurus) ou encore à la nécessité de définir un protocole de ré-indexation (avec le risque d'avoir des indexations anachroniques par rapport aux documents). Dans certains cas, maintenir un langage documentaire revient autant à faire évoluer les étiquettes lexicales qui servent d'index qu'à réorganiser la collection selon de nouvelles catégories. De même, l'utilisation des ontologies dans le cadre du Web Sémantique ou de la gestion des connaissances d'entreprises, pour l'annotation de pages web ou de documents, suppose de maîtriser l'impact de la mise à jour des ontologies sur les pages annotées.

4.3.3 Anticiper les problèmes de maintenance

Aujourd'hui, la maintenance des ressources terminologiques est mal maîtrisée et soulève des problèmes rarement abordés. Plus encore, ce problème, lorsqu'il est abordé, est traité d'un point de vue disciplinaire, en fonction des contraintes propres aux applications de cette discipline. Or, étroitement liée au mode de conception, la maintenance se heurte aux mêmes obstacles : difficulté pour trouver les connaissances à modifier pour que la ressource réponde mieux à un besoin particulier, caractère laborieux de la réorganisation d'une structuration, difficulté à évaluer l'impact et la pertinence d'une modification sur la structure conceptuelle qui est complexe et, dans un second temps, sur les utilisations mêmes de la ressource. Cette maintenance peut être facilitée si la ressource a été documentée, et surtout si les décisions de définition, de sélection et d'organisation des concepts ou des termes sont explicitement notées dans la ressource. Dans le cas de ressources construites à partir de textes, on constate que la possibilité de revenir aux textes d'origine constitue une aide à l'interprétation et donc à la maintenance.

La réflexion de notre groupe a fait ressortir plusieurs perspectives pour aborder ce problème en partageant les expériences des disciplines concernées. Ces pistes sont au service d'une meilleure traçabilité et reproductibilité du processus de modélisation, pour parvenir à des ressources faciles à maintenir et à adapter :

- C*apitaliser les expériences d'utilisation d'outils (de TAL et de modélisation) et les démarches :* un premier travail d'inventaire, amorcé par le groupe TIA par exemple et poursuivi au sein de notre AS, commence à être mis en forme (deux articles, l'un dans les actes des assises du GDR I3 et un autre soumis à la revue RIA, traduisent l'état d'avancement de cette réflexion). Il reste encore à mieux évaluer la portée de certains outils, à rendre compte de démarches dans plus de disciplines, et surtout à être plus précis dans l'adéquation entre démarches-outils et types d'application. Il s'agit également de revoir les démarches et méthodes pour faire de la construction et de la maintenance des ressources terminologiques un processus cyclique et continu, favorisant l'apprentissage ou l'enrichissement incrémental selon les besoins. Définir une méthode de maintenance revient à définir ce qui doit rester fixé par rapport à ce qui peut évoluer, et ce dans le temps, de manière à gérer des parties de ressource les unes après les autres.
- Se donner les moyens de pouvoir *archiver conjointement corpus et ressources, et même les outils spécifiques ayant servi à les construire*, en privilégiant l'apprentissage : il ressort clairement de toutes les expériences disciplinaires que conserver les termes seuls ne suffit pas pour guider la maintenance d'une ressource. Des structures de données plus riches sont donc nécessaires. L'idée serait de pousser plus loin la proposition faite dans les BCT (conserver avec la ressource les termes mais aussi un réseau conceptuel ainsi que les données qui ont servi à tout structurer) en y associant les logiciels et les connaissances (marqueurs, patrons de recherche ou d'extraction, classes lexicales, etc.) qui serviront à la faire évoluer dans le temps. En quelque sorte, la ressource serait associée à des outils d'acquisition qui lui permettraient de s'adapter aux changements contextuels.
- Décider si la ressource doit ou non comporter une « *épaisseur diachronique* ». Ainsi, les ontologies actuelles n'ont pas la prétention de pouvoir s'adapter à plusieurs contextes chronologiques. Plusieurs ontologies seraient nécessaires pour accéder à des documents d'un même domaine mais d'époques différentes. Au contraire, certains thésaurus veulent rendre compte de ce type d'information et pouvoir fournir des informations lexicales à différentes époques. En sciences de l'information, il a été montré la nécessité de disposer d'une épaisseur historique qui rende compte de l'évolution des termes, comme le font certains dictionnaires.
- Outiller l'intégration des variations dans les ressources : il semble important de confronter deux approches : un *processus incrémental* qui s'appuie sur le repérage d'écart ; *un processus itératif* où l'on reconstruit à partir de textes mis à jour une nouvelle version de la ressource. Le fait de tout refaire automatiquement à partir de nouveaux textes pour rendre compte de nouveaux usages peut présenter un intérêt dans des contextes comme l'indexation automatique de sites web, ou toute autre application où la ressource est utilisée automatiquement. Par exemple, pour l'indexation de documents de presse, une application en cours a choisi une méthode qui consiste à refaire, tous les 6 mois, un apprentissage du nouveau référentiel à partir de textes, à l'aide d'outils de TAL et d'apprentissage. Ainsi, on obtient de nouveaux descripteurs, en faisant l'hypothèse que la structure ontologique est assez simple pour rester stable. Cela suppose de reprendre également toute l'indexation (faite automatiquement). L'inconvénient en est la perte des localisations fines des changements et des raisons de ces changements, l'impossibilité de repérer et qualifier les écarts. Or ceci est fondamental dans d'autres contextes d'applications en terminologie et sciences de l'information.

En repensant globalement la construction et la maintenance des ressources terminologiques, nous sommes amenés à revoir leur statut et à leur associer un caractère dynamique. Cette nécessité de rendre les ontologies plus souples, de trouver les outils qui permettraient de les adapter et de les faire évoluer, y compris à partir des mêmes textes source, a été envisagée de manière très innovante par B.Bachimont lors d'une conférence invitée d'un workshop de l'ECAI 2002. Avec la notion de méta-ontologie, il souligne la nécessité de conserver un contexte qui permette de reconstruire des sous-ensembles de l'ontologie en fonction de l'évolution des besoins liés aux usages ou des connaissances dans le domaine. D'autres travaux rappellent enfin que tout n'est pas mouvant simultanément, ce qui autorise à fixer des

états stables pouvant servir de référence, et qui ne rend pas vaine la notion de ressource terminologique et ontologique dans son ensemble.

Les recherches en apprentissage fournissent ici des repères et des formalisations mathématiques du phénomène. Alors que la reconstruction d'une ressource revient à optimiser une fonction globale, une démarche incrémentale n'assure pas d'être optimal à chaque étape, mais à des paliers que se fixe l'utilisateur. De plus, elle réduit considérablement les coûts. Cette nouvelle manière d'envisager l'acquisition de ressources est tout à fait prometteuse, ce que souligne l'intérêt croissant des collaborations entre TAL, apprentissage et ingénierie des ressources terminologiques et ontologiques, tel que cela est développé au sein du groupe A3CTE depuis 1998. La maintenance pose en fait plusieurs problèmes à l'informatique qui se déclinent de manière tout à fait originale suite à notre analyse :

- Outiller le diagnostic qui va permettre de localiser les besoins d'évolution et d'en identifier la nature ; décider de procéder selon un suivi continu ou ponctuel ;
- Outiller l'intégration des évolutions dans les ressources, et opter entre une démarche soit incrémentale soit ;
- Outiller l'archivage des informations liées à la maintenance et aux évolutions, qui vont des versions successives du lexique et du réseau conceptuels aux versions du corpus en passant par les outils d'analyse ;
- Adapter le processus par la définition d'outils spécifiques en fonction des ressources et des besoins, de la nature des évolutions dans le projet, etc.

4.4 Etudier les problèmes d'évaluation et de validation

4.4.1 De la validation de ressources et l'évaluation d'outils à l'évaluation des recherches

Nous terminerons par quelques réflexions sur les problèmes de la validation et de l'évaluation. Nous distinguons l'évaluation d'une ressource terminologique ou ontologique particulière construite dans un contexte particulier, de l'évaluation de tel ou tel outil de TAL d'aide à la construction de ressources d'un troisième type d'évaluation : celle des recherches. Dans les deux premiers cas, il faut adopter une démarche d'ingénierie, en retenant les principes de base du génie logiciel, ce qui exige, a minima, de prendre en compte autant que possible le contexte global d'utilisation de la ressource ou de l'outil. Ces deux premiers types d'évaluations contribuent globalement à l'évaluation des recherches sur la construction de ressources à partir de textes ou leur gestion en lien avec des textes. La capacité à évaluer les recherches conditionne autant le crédit qui peut être donné à des résultats obtenus selon les méthodes que nous avons identifiées que la reconnaissance des fondements scientifiques sur lesquels ils s'appuient. En effet, c'est en montrant que nos disciplines sont à même de valider leurs méthodes et d'évaluer leur propositions et hypothèses que nous pourrons en souligner les apports et l'intérêt.

4.4.2 Validation et évaluation de ressources

En ce qui concerne les ressources, distinguons validation et évaluation.

Dans le processus de construction d'une ressource, il y a plusieurs moments de *validation*, c'est-à-dire de moment où l'analyste présente la ressource à l'experts (ou à des experts), et lui (leur) demande de valider ou d'invalider certains choix de modélisation effectués. Ces moments de validation sont d'autant moins nombreux que les experts sont peu disponibles. Ils font partie intégrante du processus de modélisation, et ceci encore plus lorsque c'est un spécialiste du domaine qui prend en charge la construction de la ressource. Ce sont donc des étapes très importantes dans le processus. L'enjeu est de s'assurer avec les experts que la conceptualisation représentée dans la RTO n'est pas en contradiction sur tel ou tel point avec les connaissances expertes et avec le rôle que la ressource va jouer auprès de ces

utilisateurs. Le problème ne se pose pas tant en terme de vérité, qu'en terme de non violation des connaissances de l'expert et de justification par le besoin de l'application. En effet, pour construire la modélisation, l'analyste a adopté un point de vue, celui de l'application cible dans laquelle sera intégrée la ressource, qui n'est pas nécessairement exactement celui de l'expert dans son activité. La tâche n'est pas simple. L'analyste doit aider l'expert, qui peut être déstabilisé par les résultats proposés, à prendre le recul nécessaire pour déceler la présence d'erreurs, voire d'absences, flagrantes.

Une fois la RTO construite, s'engage un processus *d'évaluation.* Comme nous l'avons déjà évoqué, l'évaluation doit être réalisée selon les procédures de base du génie logiciel. Il s'agit de vérifier si la RTO satisfait bien le cahier des charges et répond aux attentes spécifiées au début du projet. La difficulté, habituelle, est que l'ontologie n'est qu'un élément de l'application cible, qui est le dispositif à valider. Il faut donc concevoir des expériences et des bancs d'essais qui permettent de cibler l'évaluation sur la seule ressource. Une fois ces généralités affirmées, nous pouvons difficilement aller au-delà, parce que nous manquons encore de retour d'expérience. Chaque cas étant particulier, il sera difficile de définir des procédures à la fois précises et relativement génériques, et cela dépasse quelque peu le cadre de la recherche.

4.4.3 Validation des ressources : quels critères et quels acteurs ?

La question de la validation est évidemment cruciale dans l'élaboration de ressources terminologiques. Encore faut-il s'entendre sur ce que l'on entend par validation. En effet, différents types de validations sont envisageables, selon que l'on se place par rapport à un besoin précis, à une possibilité de réutilisation ou encore par rapport à une perspective scientifique.

La validation d'une ressource spécialisée peut se faire selon plusieurs points de vue :

- Selon son adéquation avec le corpus dont elle est issue ; même si la ressource spécialisée est construite avec un objectif précis, le corpus reste un élément de référence important. En principe donc, la représentation élaborée à partir de ce corpus doit être très proche de la connaissance consensuelle qu'il est censé véhiculer. Ce mode de validation se justifie principalement lorsque l'étude menée se situe dans une problématique plus strictement linguistique.
- Selon sa pertinence par rapport à un besoin ; on se rapproche là du concept d'utilisabilité en ergonomie. Cette validation n'est pas plus évidente à mettre en place car il est difficile de traduire comment les besoins exprimés permettent de décider des termes, concepts et propriétés à retenir. Un autre biais vient compliquer cette validation : la réutilisabilité. En effet, l'organisation de la ressource correspond à une représentation dont on peut souhaiter qu'elle serve aussi de base à un autre projet.
- Selon sa capacité à être réutilisée (généricité, clarté de la méthodologie) ; l'adéquation trop parfaite d'une ressource à un besoin peut être un handicap si cette ressource doit servir à plusieurs applications ; en effet, la terminologie constituée peut alors se révéler inutilisable pour un autre type de besoin. Dans une perspective de réutilisation, il faut a minima développer des méthodes de constitution aussi documentées que possible afin que les utilisateurs puissent avoir accès non seulement à la ressource construite mais aussi à la façon dont elle a été construite.

Au-delà de la validation qui concerne l'utilisabilité, la question du rôle de chacun des intervenants dans l'élaboration des ressources doit être posée. Plusieurs types d'acteurs interviennent en effet dans ce type de processus : le client (celui dont on a identifié le besoin), le demandeur (celui qui sert d'intermédiaire entre le client et le constructeur pour spécifier le besoin), le constructeur (terminologue, ingénieur de la connaissance ou documentaliste), l'expert. Il se peut que plusieurs rôles soient tenus par la même personne. Il convient de s'interroger sur les avantages et les inconvénients qu'il peut y avoir à diminuer le nombre d'intervenants. Dans l'objectif qui consiste à proposer des méthodes de constructions, on peut se demander par exemple, s'il y aurait un sens à ce que ces méthodes soient si précises en termes de

connaissances et d'outils à mettre en œuvre qu'elles permettraient que le « constructeur » disparaisse et que l'expert ou le demandeur, voire le client construise directement ses ressources.

4.4.4 Evaluation de logiciels de construction de ressources

L'évaluation des outils de construction de ressources est un autre problème qui nous concerne ici. C'est un problème lui aussi difficile. La source des difficultés est double : d'abord il s'agit d'outils d'aide, ensuite chaque outil est rarement utilisé seul. Quand il s'agit d'évaluer un outil automatique, du type « boîte noire », il est possible d'évaluer ses performances en comparant les résultats qu'il fournit à des résultats attendus (« gold standard »). En revanche, la situation est plus complexe dans le cas des outils d'aide qui nous intéressent ici. Les résultats fournis par les outils sont interprétés par l'analyste, et le résultat de cette interprétation est variable : il peut déboucher sur une modification, un enrichissement de la ressource à un ou plusieurs points du réseau, voire, dans certains cas, l'absence d'action immédiate, sans que cela signifie nécessairement que les résultats en question soient faux, ni même non pertinents. De plus, chaque interprétation s'appuie normalement sur une confirmation par retour aux textes. Or il n'y a pas systématiquement de trace directe entre un résultat (ou un ensemble de résultats) de l'outil et telle ou telle portion de la ressource. Si on rajoute à cela qu'une portion de RTO n'a de sens que dans la globalité de la ressource, et la ressource elle-même ne peut être évaluée qu'en contexte, on saisit l'ampleur de la tâche. Il y a un tel parcours interprétatif entre les résultats de l'outil et la ressource construite que le mode d'évaluation par comparaison entre les résultats de l'outil et une ressource de référence ne peut apporter que des résultats limités, même si cela peut donner des indications très intéressantes pour faire évoluer l'outil (Nazarenko *et al.,* 2001). Là encore, nous n'avons de solution miracle à proposer. L'idéal serait par exemple de comparer en termes de temps de réalisation et de qualité deux ressources ontologiques, l'une construite avec tel outil, et l'autre sans. Quand on connaît le temps de développement d'une ontologie, on imagine la lourdeur, et la difficulté de mise en œuvre d'une telle méthodologie. Le problème reste ouvert. Pour mesurer, ne serait-ce que d'un point de vue qualitatif, l'intérêt des outils, considérons pour le moment qu'il est primordial de les tester dans des contextes nombreux et variés et aussi réels que possible pour faire avancer la recherche.

5 SYNTHÈSE

L'action spécifique « Corpus et Terminologie » a démarré début 2002 et a déroulé son activité jusqu'en juillet 2003. Elle a rassemblé une trentaine de chercheurs d'horizons disciplinaires variés: en informatique, recherche d'information, traitement automatique des langues (TAL), apprentissage et ingénierie des connaissances (IC); en sciences du langage, terminologie et linguistique de corpus; et enfin en sciences de l'information.

Le contenu de cette action spécifique s'est élaboré sur un double constat. D'une part, le constat de l'existence de projets collaboratifs autour de la constitution de terminologies à partir de corpus faisant intervenir le plus souvent deux approches (par exemple, TAL et linguistique de corpus ou IC et TAL) mais aussi de groupes de recherche fonctionnant sur des thématiques proches (par exemple, A3CTE[8] ou TIA[9]). D'autre part, une demande sociétale très importante en matière de ressources terminologiques et des réponses qui se font souvent au coup par coup, en fonction des opportunités mais sans que les compétences propres aux disciplines soient clairement établies. Il a semblé urgent de dresser un état des lieux des compétences des disciplines concernées et de leurs complémentarités afin de donner une assise aux recherches et de définir les lignes forces de ce qui pourrait constituer la recherche sur ce thème dans les prochaines années.

Etant donnée la situation (différentes approches en présence, nécessité de définir des convergences), nous avons choisi un fonctionnement par réunions de travail (au total, 10 réunions ont eu lieu en 18 mois). Nous avons également eu le souci d'ouvrir le débat et de diffuser notre réflexion par l'organisation de deux ateliers au cours de deux conférences (CFD[10] et Plate-forme AFIA[11]) et par la gestion d'un site web[12] qui donne accès à nos documents de travail. Nous avons alterné réunions plénières, et travaux en sous-groupes. En effet, il est apparu primordial que les différents points de vue disciplinaires s'accordent dans un premier temps sur leurs objectifs et leurs méthodes. La réflexion en quatre sous-groupes a donc été encouragée : traitement automatique des langues, linguistique de corpus et terminologie, ingénierie des connaissances, sciences de l'information et recherche d'information. Afin de baliser cette réflexion, nous avons proposé une grille de questions à débattre qui s'est organisée autour de quatre thèmes: besoins ciblés, place et nature des corpus, type de ressource utilisées ou produites, méthodes et outils. Les réunions plénières ont permis des mises en commun à partir desquelles ont pu se dégager d'une part un socle d'éléments partagés et d'autre part, un ensemble de questionnements. Discutés, argumentés, élaborés, ces questionnements nous ont permis de dégager quatre axes de prospectives.

1 - Développer et approfondir la notion de « genre textuel »

Presque toutes les disciplines en présence dans l'action spécifique ont repéré une dépendance entre régularités linguistiques et situations de production de langage (écrit dans le cas qui nous intéresse). Cette co-variance est appelée genre depuis l'antiquité. Il s'agit d'une notion séduisante mais difficile à mettre en place. En effet, différents points de vue (en lien avec l'identification des critères pertinents dans la situation de production mais aussi avec l'objectif de l'analyse de corpus) peuvent intervenir pour définir les genres. Un travail important serait nécessaire pour essayer de définir des critères pertinents, soit généraux, soit à relier à une situation particulière. Les méthodes de TAL devraient permettre de tester rapidement des hypothèses. Une approche interdisciplinaire sur cette thématique devrait être encouragée.

[8] http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/index.html
[9] Groupe « Terminologie et IA » rattaché à la S.A 6.1 du GDR-I3 . http://www.biomath.jussieu.fr/TIA/
[10] Conférence Fédérative sur le Document, Hammamet (Tunisie), Octobre 2002.
[11] Plate-forme de conférences de l'Association Française d'Intelligence Artificielle, Laval (F), juillet 2003
[12] http://www.irit.fr/ASSTICCOT/

2 - Prendre en compte les applications pour comprendre la variabilité des méthodes et des ressources terminologiques

La nature de l'application a de multiples impacts : choix des outils, choix des techniques d'exploration, mode d'interprétation des résultats. De nombreuses expériences témoignent de l'importance de la prise en compte de l'application et ce, très en amont du processus, au moment même de la constitution du corpus. La réflexion théorique commence à se mettre en place sur ce sujet. Mais la prise en compte de l'application ne se fait pas encore de manière systématique, loin s'en faut; cette nécessité se heurte à deux éléments. D'une part, les outils ne sont pas adaptés. Conçus dans l'optique d'un fonctionnement « générique », ils ne prennent que difficilement en compte une variation quelle que soit sa nature. D'autre part, les résultats ne sont pas encore nombreux et il est difficile d'en tirer des généralisations qui permettraient d'améliorer les outils et les modes d'interprétation.

3 - Définir des méthodes pour assurer la maintenance des terminologies

Les ressources terminologiques sont l'objet d'un paradoxe : elles doivent à la fois « normaliser » des connaissances, c'est-à-dire les figer à un moment donné et être utilisées pour accéder à des connaissances qui évoluent dans des contextes dynamiques. Des propositions ont été faites pour conserver des connaissances sur le contexte dans lequel a été élaboré la ressource (corpus, connaissances linguistiques pertinentes, voire logiciels…). Mais, comme pour le point précédent, réflexion théorique (sur l'évolution de la connaissance) et adaptation des outils (de diagnostic, d'intégration des évolutions, d'archivage…) et des méthodes (rôle des méthodes d'apprentissage par exemple) devraient se développer conjointement dans le cadre de projets interdisciplinaires.

4 - Etudier les problèmes d'évaluation et de validation

Une des difficultés majeures, devant la prolifération des projets de construction de terminologies à partir de corpus, est celle de l'évaluation et de la validation des résultats. Si la validation concerne majoritairement l'expert auquel on soumet la ressource construite, l'évaluation est, elle, à recadrer dans l'ensemble d'un projet voire, dans l'ensemble des besoins d'une entreprise par exemple (compatibilité de la ressource avec un projet mais aussi avec d'autres projets). La question fondamentale est donc celle du choix entre d'un côté spécificité/adaptation à un besoin/faible réutilisation et, d'un autre côté généricité/faible adaptation à un besoin particulier/réutilisabilité. En réalité, les trois termes de ces possibilités doivent être examinés avec précaution et objectivité afin de définir au mieux les outils et méthodes à mettre en place mais aussi à calculer au plus juste l'investissement nécessaire en temps et en argent.

A l'issue de cette Action Spécifique, nos propositions, en termes de prospectives vont dans deux sens:

- *Proposer et soutenir le financement de projets d'envergure qui prennent en compte la variation dans les ressources terminologiques.*

Il s'avère que les ressources terminologiques doivent pouvoir s'adapter au changement, qu'il soit lié à la nature des corpus, à l'application ou à la diachronie. Il est nécessaire de mettre sur pied des projets d'envergure qui permettent d'une part le développement d'outils modulaires et paramétrables, aptes à prendre en compte cette variation et, d'autre part, l'évaluation et l'analyse théorique des résultats fournis par les outils afin de leur donner un sens et d'envisager des modes de généralisation. Pour les deux axes, très complémentaires, l'objectif est de diminuer les coûts de construction, coûts qui restent la pierre d'achoppement de la construction de terminologies à partir de corpus. Pour mener à bien ces projets à moyen terme, les financements ne peuvent pas venir uniquement des entreprises.

- *Donner les moyens humains, par des recrutements de jeunes chercheurs compétents formes à cette interdisciplinarité, et les moyens financiers, par des programmes pluridisciplinaires portant sur ces thématiques.*

- *Continuer à soutenir des groupes de réflexion interdisciplinaire.*

Les réunions d'ASSTICCOT, qui ont pu paraître parfois décousues et peu finalisées, ont permis à chacun d'appréhender objectifs, problématiques et méthodes des autres disciplines. Il s'est dégagé de ces échanges un fond de problématiques similaires, qui ne remet pas en question chaque approche disciplinaire mais qui permet de repérer un axe commun, composé de similitudes à la fois dans les demandes et les réponses apportées, sans doute soutenu par l'état de la technique et de la connaissance à un moment précis. Il serait donc dommage que ce type de groupe, organisé autour d'un besoin particulier tout en s'appuyant sur une recherche de théorisation, ne puisse pas trouver un cadre pour mener ce genre d'activité autour des axes de prospective dégagés.

SIGLES

A3CTE : Groupe de travail : Applications, Apprentissage, Acquisition de Connaissances à partir de Textes Electroniques (http://www-lipn.univ-paris13.fr/groupes-de-travail/A3CTE/)

AFIA : Association Française d'Intelligence Artificielle

BCT : Base de Connaissances Terminologiques

CFD : Conférence Fédérative sur le Document, Hammamet, Tunisie, octobre 2002.

ECAI : European Conference on Artificial Intelligence

ENSSIB : Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques

IC : Ingénierie des Connaissances

ISDN : Institut Supérieur du Document Numérique

RI : Recherche d'Information

RTO : Ressource terminologique ou ontologique

SI : Sciences de l'Information

TAL : Traitement Automatique de la Langue

TIA : Groupe de travail Terminologie et Intelligence Artificielle (http://www.biomath.jussieu.fr/TIA/)

BIBLIOGRAPHIE

AIT EL MEKKI T., NAZARENKO A., (2002), Comment aider un auteur à construire l'index d'un ouvrage ? L'architecture du système IndDoc , *actes du Colloque International sur la Fouille de Texte CIFT'2002*, Y. Toussaint et C. Nedellec Eds., oct. 2002, p. 141-158.

AUROUX S., (1998) *La raison, le langage et les normes*. Paris : PUF.

AUSSENAC-GILLES N. GEDITERM : un logiciel pour gérer des bases de connaissances terminologiques *Terminologies Nouvelles*, 19, Nov 1999. pp 111-123.

AUSSENAC N., SEGUELA P., (2000) Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, Numéro spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : ERSS. Pp 175-198.

AUSSENAC-GILLES N., CONDAMINES A. (2001), Entre textes et ontologies formelles : les Bases de Connaissances Terminologiques. In M. Zacklad et M. Grundstein (éds) *: Ingénierie et capitalisation des connaissances*. Paris : Hermès. 2001, pp.153-176.

AUSSENAC-GILLES N., CONDAMINES A., SZULMAN S. (2002), Prise en compte de l'application dans la constitution de produits terminologiques. *Actes des 2e Assises Nationales du GDR I3*, Nancy (F), Déc. 2002. Toulouse : Cépaduès Editions. Pp 289-302.

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S., (2003a) D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. *5e rencontres Terminologie et IA, TIA 2003*. Ed. F. Rousselot. Strasbourg (F), ENSSAIS, Avril 2003. pp 41-53.

AUSSENAC-GILLES N., BIÉBOW B., SZULMAN S. (2003b), Modélisation du domaine par une méthode fondée sur l'analyse de corpus. *In Ingénierie des Connaissances*. R. Teulier, P. Tchounikine et J. Charlet Eds. Paris : Eyrolles. A paraître en 2003.

AUSSENAC-GILLES N., BOURIGAULT D., TEULIER R. (2003c), Analyse comparative de corpus : cas de l'ingénierie des connaissances. *Actes de IC2003 (14° journées Francophones d'Ingénierie des Connaissances)*. Présidente : R. Dieng-Kuntz. Laval (F), 1-3 Juillet 2003. Presses Universitaires de Grenoble. pp 67-84.

BACHIMONT B. (1996), *Herméneutique matérielle et Artéfacture : des machines qui pensent aux machines qui donnent à penser ; Critique du formalisme en intelligence artificielle*. Thèse de doctorat d'épistémologie, École Polytechnique, 1996.

BACHIMONT B., (2000) Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In *Ingénierie des connaissances : évolutions récentes et nouveaux défis*. Eds. : J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT. Paris : Eyrolles, 2000. pp 305-324.

BAEZA-YATES, R., RIBEIRO-NETO, B., (1999) *Modern Information Retrieval*, Addison-Wesley Ed., ISBN 0-201-39829-X, 1999.

BAKHTINE M., (1984) : *Esthétique de la création verbale*. Paris : Gallimard.

BAUDOUIN N., HOLZEM M., SAIDAILI Y., LABICHE J., Acquisition itérative de connaissances en traitement d'image : rôle d'un collège d'experts. *In Actes des 14° journées Francophones d'Ingénierie des Connaissances*. Pdte : R. Dieng-Kuntz. Presse Universitaires de Grenoble. 2003. pp 101-116

BAZIZ M., AUSSENAC-GILLES N., BOUGHANEM M., (2003), Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. *ISI*. Paris : Hermès. à paraître fin 2003.

BEAUVISAGE T., (2001), Morphosyntaxe et genre textuel, *TALN*, n°2-2001, Paris : Hermès. 579-608.

BESSIÈRES P., NAZARENKO A. et NÉDELLEC C. (2001), Apport de l'apprentissage à l'extraction d'information : le problème de l'identification d'interactions géniques, *Actes de CIDE'01*, Toulouse, octobre 2001.

BIBER D., (1988), *Variation Across Speech and* Writing . Cambridge University Press.

BLASCHKE C., ANDRADE M. A., OUZOUNIS C. and VALENCIA A. (1999), Automatic Extraction of biological information from scientific text: protein-protein interactions, in Proceedings of *International Symposium on Molecular Biology, (ISMB'99)*, 1999.

BOUGHANEM M., TEBRI H. AND M. TMAR (2002), IRIT at TREC'2002: Filtering Track, *TREC 2002*.

BOURIGAULT D. (1994), *Lexter, un Logiciel d'Extraction de TERminologie. Application à l'extraction des connaissances à partir de textes.* Thèse en Mathématiques, Informatique appliquée aux sciences de l'homme. EHESS, Paris.

BOURIGAULT D., CONDAMINES A., (1995) Réflexions sur le concept de base de connaissances terminologiques. *Journée du PRC IA*, 1-2 février 1995, Nancy : Teknea, pp.425-445.

BOURIGAULT D., FABRE C., (2000) Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaires*, n° 25, 2000, Université Toulouse - Le Mirail, pp. 131-151.

BOURIGAULT D., JACQUEMIN C. (2000) :« Construction de ressources terminologiques ». J.-M. Pierrel (ed) : *Ingénierie des langues*, Traité I2C, Paris : Hermes. pp. 215-233.

BOURIGAULT D., (2002) Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2002), Nancy, 2002, pp. 75-84.

BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J., (2004) Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA*), Numéro spécial « Terminologie ». M. Slodzian (Ed.). Hermès : Paris. à paraître en 2004.

BOUVERET M. (1997) Le terme, une dénomination au sens réglé, *Actes de TIA 1997.*

BRANCA-ROSOFF S. (1999) Types, modes et genres : entre langue et discours. S.Branca-Rosoff (ed.) : *Langage et Société* n°87, *Types, modes et genres de discours*. pp. 5-24.

BRONCKART J.-P., (1996) *Activités langagières, textes et discours*. Lausanne : Delachaux et Niestlé.

CERBAH F., EUZENAT J. (2000) Integrating Textual Knowledge and Formal Knowledge for Improving Traceability. *Proc. of the 12th European Workshop on Knowledge Engineering and Knowledge Management (EKAW 2000)* Juan-les-Pins (F), Springer Verlag. 296-303.

CHARLET J., ZACKLAD M., KASSEL G., BOURIGAULT D. (2000), *Ingénierie des connaissances : évolutions récentes et nouveaux défis*. Eds. . Paris : Eyrolles, 2000.

CHARLET J. (2002), *L'ingénierie des connaissances : résultats, développements et perspectives pour la gestion des connaissances médicales*. Mémoire d'habilitation à diriger des recherches, Paris VI, déc. 2002.

CLEVERDON C.W. (1968), Effects of variations in relevance assessments, Cranfield library report 3.

CONDAMINES A., (2003a) : Vers la définition de genres interprétatifs. *Actes de TIA'2003*, Université Marc Bloch, Strasbourg, 31 mars-1è avril 2003, pp.69-79.

CONDAMINES A. (2003b) : *Sémantique et corpus spécialisé : Constitution de bases de connaissances terminologiques*. Mémoire d'Habilitation à Diriger les Recherches, Juin 2003, Université Toulouse Le Mirail; ERSS : Carnets de grammaire.

CONDAMINES A., REBEYROLLE J, (2000) Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault, (eds).*: Ingénierie des Connaissances, évolutions récentes et nouveaux défis*. Paris : Eyrolles. 2000, pp.225-242.

CONDAMINES A., GALARRETA D., PERRUSSEL L., REBEYROLE J., ROTHENBURGER B., VIGUIER-PLA S., (2003) Tools and methods for knowledge evolution measure in space project. Proceedings of *54th International Astronautical Congress*, 29 septembre, 3 Octobre, 2003, Bremen, Germany, to appear.

CORCHO O., FERNÁNDEZ-LÓPEZ M., GÓMEZ-PÉREZ A., VICENTE O. (2002) WebODE: An Integrated Workbench for Ontology Representation, Reasoning, and Exchange. Proceedings of EKAW 2002. Springer Verlag. 138-153

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

DAOUST F. (1992), *SATO (Système d'Analyse de Textes par Ordinateur)* version 3.6, Manuel de référence, Centre ATO Université du Québec à Montréal.

DAVID S. & PLANTE P. (1990), *Termino version 1.0*, Rapport du Centre d'Analyse de Textes par Ordinateur. Université du Québec à Montréal.

DELAVIGNE V. (2003), Quand le terme entre en vulgarisation, *Actes de TIA 2003*, Université Marc Bloch, Strasbourg, 31 mars-1 avril 2003. 80-91

DESPRÉS, S. (2001) Une comparaison raisonnée des apports de la terminologie et de l'intelligence artificielle pour servir et améliorer la construction d'ontologies, *TIA-2001*, Inist, Nancy, 3 et 4 mai 2001.

DIENG-KUNTZ, R., CORBY, O., GANDON, F., GIBOIN, A., GOLEBIOWSKA, J., MATTA, N. and RIBIÈRE, M. (2001). *Méthodes et outils pour la gestion des connaissances ; une approche pluridisciplinaire du Knowledge Management*. 2e édition, Dunod, Paris.

DOMINGUE, J., MOTTA, E. (2000). Planet-Onto: From News Publishing to Integrated Knowledge Management Support. *IEEE Intelligent Systems*, 15(3), p. 26-32.

DUINEVELD A.J., STOTER R., WEIDEN, M.R., KENEPA B., BENJAMINS R., Wondertools ? A comparative study of Ontological engineering tools. *International Journal of Human-Computer Studies*. 51 : 1111-1133.

ENGUEHARD C., PANTÉRA L., (1995), Automatic natural acquisition of terminology. *Journal of Quantitative Linguistics*, vol.2, n°1. pp. 27-32

FAURE D., NEDELLEC C, (1999), Knowledge Acquisition of Predicate Argument Structures from Technical Texts Using Machine Learning: The System ASIUM Proc. of the *11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW'99)*, Juan-les-Pins, France, 329-334.

FERNANDEZ-LOPEZ M., GOMEZ-PEREZ A., PAZOS J., PAZOS A., (1999), Building a Chemical Ontology using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems and their applications*. # 4 (1) : 37-45.

FORTIER, J.-Y., KASSEL, G. (2003). Modeling the information contained in an organizational memory to facilitate its access. In Proceedings of the 10th International Conference on Human-Computer Interaction: HCI'2003, June 22-27, Crete (Greece).

GANDON F., DIENG-KUNTZ R., Ontologie pour un système multi-agents dédié à une mémoire d'entreprise, in actes des *journées Ingénierie des Connaissances IC'2001,* Grenoble (F), Juin 2001, pp 1-20

GANGEMI A., GUARINO N., MASOLO C., OLTRAMARI A., SCHNEIDER L., (2002), Sweetening Ontologies with DOLCE. Knowledge Engineering and Knowledge Management, Proc. Of EKAW2002, A. Gomez-Perez and R. Benjamins Eds., LNAI 2473. Springer Verlag. 2002 :166-181.

GOLEBIOWSKA J., DIENG-KUNTZ R., CORBY O., MOUSSEAU D, (2001). Exploitation des ontologies pour la mémoire d'un projet-véhicule, *Actes des 4èmes rencontres "Terminologies et Intelligence Artificielle*" (TIA 2001) Nancy, pp 170-179.

GOMEZ-PEREZ A., MANZANO MACHO D., (2003). A survey of ontology learning methods and techniques. Deliverable 1.5. IST Project IST-2000-29243 OntoWeb. May 2003. http://www.ontoweb.org/

GUARINO N., WELTY C.-A. (2000) : Identity, Unity, and Individuality: Towards a Formal Toolkit for Ontological Analysis. ECAI 2000: 219-223

HAMON T.et NAZARENKO A. (2001), Detection of synonymy links between terms: experiment and results, *Recent Advances in Computational Terminology.* John Benjamins, 2001.

HAMON T., NAZARENKO A., (EDS), (2002) : *Structuration de terminologie.* TAL volume 43 - n°1/2002.

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

HUBERT G., MOTHE J, BENAMMAR A., DKAK T.. DOUSSET B., KAROUACH S., (2001), *Textual document mining using a graphical interface,* 9th Int. Conf. Human-Computer Interface, pp 918-922, Nouvelle Orléans, Août 2001.

ILLOUZ G, HABERT B., FLEURY S., FOLCH H., HEIDEN S., LAFON P., (1999), « Maîtriser les déluges de données hétérogènes ». A.Condamines, C.Fabre, M.-P.Péry Woodley (eds) : Actes de l'Atelier «*Corpus et TAL, Pour une réflexion méthodologique ».* TALN'99. pp. 37-46.

JOUIS C., (1993), *Contribution à la conceptualisation et à la modélisation des connaissances à partir d'une analyse de textes. Réalisation d'un prototype : le système Seek.* Thèse d'informatique, EHESS, Paris.

KANG S.-J. & LEE J.-H. (2001), Semi-automatic Practical Ontology Construction by Using a thesaurus, Computational Dictionaries and Large Corpora. *In Proceedings of the ACL workshop on Human Language technologies and Knowledge Management.* Toulouse, July 6-7, 2001. 29-36.

KASSEL G. (2002) OntoSpec : une méthode de spécification semi-informelle d'ontologies. *Actes de la 13e conférence d'ingénierie des connaissances (IC2002).* Rouen (F). mai 2002.

Le MOIGNO S., CHARLET J., BOURIGAULT D., & JAULENT M.-C. (2002), Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. *Actes des 13èmes journées francophones d'ingénierie des connaissances* (IC 2002), Rouen, 2002, pp. 229-238.

LAINÉ-CRUZEL S. (2001), Vers un nouveau positionnement des professionnels de l'information. *3ème colloque du Chapitre français de l'ISKO (International Society for Knowledge Organisation) : Filtrage et résumé automatique de l'information sur les réseaux*. Paris, 5-6 juillet 2001.

LAINÉ-CRUZEL S. (2001), *Conception de systèmes de recherche d'informations : accès aux documents numériques scientifiques*. Habilitation à diriger des Recherches, soutenue le 22 juin 2001, Université Claude Bernard Lyon 1. http://www.recodoc.univ-lyon1.fr/hdr_SLC.pdf

MAEDCHE A. & STAAB S. (2000), Mining Ontologies from Text. In *Knowledge Engineering and Knowledge management: methods, models and tools, proceedings of EKAW2000*. R. Dieng and O. Corby (Eds). Bonn : Springer Verlag. LNAI 1937.

MAINGUENEAU D. (1987), *Nouvelles tendances en analyse du discours*, Paris : Hachette, 1987.

MASOLO C. (2001) Ontology driven Information retrieval : Stato dell'arte. Report of the IKF (Information and Knowledge Fusion) E!2235. LADSEB-Cnr, Padova (I).

MEYER I., BOWKER L., ECK K., (1992), Cogniterm: An Experiment in Building a Terminological Knowledge Base. *Proceedings 5th EURALEX International Congress on Lexicography*, Tampere, Finland

MORIN E. (1999), Acquisition de patrons lexico-syntaxiques caractéristiques d'une relation sémantique, *TAL* (*Traitement Automatique des Langues*), vol.40, n°1, Paris : Université Paris VII, 143-166.

MOTHE J. (2000), *Recherche et exploration d'informations – Découvertes de connaissances pour l'accès à l'information*, Habilitation à diriger des recherches, Université Paul Sabatier, Décembre 2000. http://www.irit.fr/ACTIVITES/EQ_SIG/personnes/mothe/pub/HDR.ps.gz

MOTHE J., CHRISMENT C., DKAKI D, DOUSSET B., EGRET D., 2001, *Information mining: use of the document dimensions to analyse interactively a document set*, pp 66-77, European Colloquium on IR Research: ECIR, Avril 2001.

NAZARENKO A., ZWEIGENBAUM P., HABERT B. ET BOUAUD J. (2001) Corpus-based Extension of a Terminological Semantic Lexicon, *Recent Advances in Computational Terminology*. John Benjamins, 2001.

NAZARENKO A., HAMON T., (2002) Structuration de terminologie. *Traitement Automatique des Langues*. Vol 43 n°1, Hermès 2002.

NECHES R., FIKES R., FININ T., GRUBER T., PATIL R., SENATIR T., and W. SWARTOUT. (1991) Enabling technology for knowledge sharing. *AI Magazine*, 12(3):36--56, 1991.

NÉDELLEC C. (2002), Bibliographical Information Extraction in Genomics, in *IEEE Intelligent Systems: Trends & Controversies - Mining Information for Functional Genomics*, N. Shadbolt (éd.), p. 76-78, mai-juin, 2002.

NÉDELLEC C., OULD A., VETAH M., et BESSIÈRES P., 2001 :Sentence Filtering for Information Extraction in Genomics, a Classification Problem". In Proceedings of the Conference on Practical Knowledge Discovery in Databases, PKDD'2001, p. 326-338, Freiburg, sept. 2001.

NORMAND S., (2002), *Les mots de la dégustation. Analyses sémantiques d'un discours professionnel*.CNRS Editions,

O'LEARY, D.E. (1998). Using AI in Knowledge Management: Knowledge Bases and Ontologies. *IEEE Intelligent Systems*, 34-39.

PERY-WOODLEY M.-P., (1995), Quels corpus pour quels traitements automatiques. *TAL* (Traitement Automatique des Langues), n°36 (1-2). pp. 213-232.

PILLET V. (2000), *Méthodologie d'extraction automatique d'information à partir de la littérature scientifique en vue d'alimenter un nouveau système d'information*, thèse de l'Université de droit, d'économie et des sciences d'Aix-Marseille, 2000.

POIBEAU T. (2001), Extraction d'information dans les bases de données textuelles en génomique au moyen de transducteurs à nombre fini d'états, In *Actes de la Conférence Française de Traitement Automatique de la Langue, (TALN'2001)*, 2001.

RASTIER F., (1995), Le terme : Entre ontologie et Linguistique. *La Banque des Mots* n°7, Numéro spécial. pp. 35-64.

RASTIER F., (2001) *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques.

REBEYROLLE J., TANGUY L. « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires ». *Cahiers de Grammaire*, n°25, Université Toulouse - Le Mirail, 2000. p 153- 174.

REYNAUD C, M.C. ROUSSET, B. SAFA (2002). Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL. Revue I3. N°1. Vol. 1 Cépaduès-Editions.

VAN RIJSBERGEN K. (1979), *Information Retrieval*, Butterworths, London, 2 Edition. http://www.dcs.gla.ac.uk/Keith/Preface.html

RILOFF E. (1993), Automatically constructing a Dictionary for Information Extraction Tasks. In Proceedings of the *Eleventh National Conference on Artificial Intelligence (AAAI-93)*, p. 811-816, AAAI Press / The MIT Press.

ROUSSELOT F., FRATH P., OUESLATI R., (1996), Extracting Concepts and relations from corpora. *Proceedings ECAI'96, 12th European Conference on Artificial Intelligence*

ROUSSEY C., CALABRETTO, S., PINON, J.-M. (2001). SyDoM: A multilingual Information Retrieval System for Digital Libraries. In *Proceedings of the 5th International ICCC/IFIP Conference on Electronic Publishing: ELPUB'2001*, Canterbury (UK), p. 150-160.

SALTON G. (1971), *The SMART Retrieval System – Experiments in automatic document processing*, Prentice Hall Inc., Englewood Cliffs, NL, 1971.

SALTON G., ALLAN J., BUCKLEY C. (1994), Automatic structuring and retrieval of large text files, communication de l'ACM, 37(2), pp 97-108.

SEBASTIANI F. (2003), Text Categorization, In A. ZANASI (ed.), *Text Mining and its Applications*, WIT Press, Southampton, UK, Forthcoming 2003.

SÉROUSSI, B., BOUAUD J., ANTOINE E.-C., ZELEK L., SPIELMANN M. (2000): Using ONCODOC as a Computer-Based Eligibility Screening System to Improve Accrual onto Breast Cancer Clinical Trials. AIME 2001: 421-430

SÉGUÉLA P., AUSSENAC-GILLES N., (1999) Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. *Actes de la conférence IC'99 - Plate-forme AFIA*. Palaiseau (F), 14-18 Juin 1999. pp 79-88.

SÉGUÉLA P., (2001) *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse d'informatique, Université Paul Sabatier, Toulouse.

SLODZIAN M., (1995) La doctrine terminologique, nouvelle théorie du signe au carrefour de l'universalisme et du logicisme. *ALFA (Actes de Langue Française et de Linguistique : Terminologie et langues de spécialité*, 7/8, Dalhousiana, Halifax, Nova Scotia, Canada. 1994-1995. pp. 121-136.

SODERLAND S., (1999) : Learning Information Extraction Rules for Semi-Structured and Free Text, in *Machine Learning Journal*, vol 34, 1999.

SPARCK JONES K. (2003), Document Retrieval: Shallow Data, Deep Theories; Historical Reflections, Potential Directions, *ECIR*, pp 1-11.

STAAB, S., MAEDCHE, A. (2001). Knowledge Portals, Ontologies at Work. *AI Magazine*, Summer 2001, p. 63-75.

STAAB S., MÄDCHE A., NÉDELLEC C. and WIEMER-HASTINGS P., (2000) ECAI Workshop Notes of the *Ontology Learning*, workshop of the 14th European Conference on Artificial Intelligence (ECAI), Berlin, Août 2000.

STAPLEY B. J. and BENOIT G. (2000), Bibliometrics: Information Retrieval and Visualization from co-occurrence of gene names in MedLine abstracts. In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*.

SZULMAN S., BIÉBOW B.& AUSSENAC-GILLES N. (2002), Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE, *TAL*, Paris : Hermès. Vol 43, N°1. 2002. pp 103-128.

TRONCY R., ISAAC A., (2002) DOE: une mise en œuvre d'une méthode de structuration différentielle pour les ontologies, *Actes des 13ièmes journées francophones d'Ingénierie des Connaissances IC 2002*, Rouen (F), mai 2002. pp 63-74.

THOMAS, J., MILWARD, D., OUZOUNIS C., PULMAN S. AND CAROLL M. (2000) Automatic Extraction of Protein Interactions from Scientific Abstracts". In Proceedings of the *Pacific Symposium on biocomputing (PSB'2000)*, vol.5, p. 502-513, Honolulu.

VELARDI P., MISSIKOFF M., BASILI R. (2001) Identification of relevant terms to support the construction of domain ontologies. *In Proceedings of the ACL workshop on Human Language technologies and Knowledge Management*. Toulouse, July 6-7, 2001. 18-28.

WUSTER E., (1981) L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et les sciences des choses. G.Rondeau et H.Felber (eds) : *Textes choisis de terminologie*, GIRSTERM, Université de Laval, Québec. pp. 55-108.

Annexes

1 MODE DE FONCTIONNEMENT

Le mode de fonctionnement retenu a consisté à solliciter des collègues prêts à s'investir dans un projet pluridisciplinaire, et cela dans chacune des disciplines concernées par la thématique «terminologies et corpus » (ingénierie des connaissances, traitement automatique du langage naturel, sciences de l'information, recherche d'information et sciences du langage). Ces chercheurs, au nombre d'une trentaine, sont issus de domaines disciplinaires différents : linguistique, et particulièrement linguistique de corpus, terminologie, sociolinguistique, informatique, et particulièrement traitement automatique des langues, ingénierie des connaissances, apprentissage automatique et recherche d'information, et enfin sciences de l'information. Les chercheurs concernés sont listés dans l'annexe 2 de ce document. Le groupe a fonctionné par des réunions fermées, plénières ou en sous-groupes, auxquelles se sont ajoutées des actions plus ponctuelles, n'impliquant qu'une partie des chercheurs, comme l'organisation de journées scientifiques, la rédaction d'articles ou l'élaboration d'une déclaration d'intérêt pour un réseau d'excellence européen dans le cadre du 6e PCRD.

Nous avons jugé utile de mentionner comme compléments à ce rapport d'activité, qui se veut avant tout la synthèse de la réflexion scientifique menée dans le groupe, les différents documents et les informations accessibles sur le site web (http://www;irit.fr/ASSTICCOT/). Ce site regroupe des documents de travail produits par le groupe ainsi que des productions du groupe qui nous semblent importantes pour se rendre compte de la nature de ses activités et pour mieux connaître ses productions. La structure du site correspond aux différentes activités de l'AS.

Lors d'une première réunion de présentation, nous avons travaillé à la définition d'un ensemble de questions qui devaient servir de trame à la réflexion par groupe disciplinaire par la suite. Cet ensemble de questions, soumis à discussion lors de la première réunion, est commenté dans la partie 3, et leur liste est reportée intégralement dans l'annexe 2 de ce document. C'est cette trame commune et les discussions autour d'exposés venant y apporter des éléments de réponse qui articuleront le travail inter-disciplinaire.

Pour assurer une visibilité claire des points de vue disciplinaire et faciliter une meilleure connaissance réciproque, le groupe a été organisé en sous-groupes disciplinaires. Ces sous-groupes ont été animés chacun par une personne plus impliquée dans le projet. Nous avons fait en sorte que les groupes travaillent de manière indépendante avant de produire des synthèses. Leurs réflexions et synthèses ont exposés lors de réunions plénières à l'ensemble des membres du projet avec, pour objectif, d'identifier les similitudes, les différences et les complémentarités des approches.

Les participants à l'AS se sont réunis 10 fois, soit environ toutes les 6 semaines au sein de réunions fermées et de journées scientifiques, aux dates suivantes :

14 janvier 2002 ; 12 mars 2002, 7 mai 2002, 11 juin 2002, 10 septembre 2002, 20 octobre 2002 (CFD), 14 novembre 2002, 25 mars 2003, 30 juin 2003, 4 juillet 2003 (plate-forme AFIA)

Les comptes rendus de ces réunions constituent la rubrique « pages réservées aux membres du groupe », item « réunions plénières » du site, et les transparents correspondant à des présentations de chercheurs sont accessibles depuis les pages de chacune des réunions. Des textes de travail ont aussi été rédigés par les sous-groupes disciplinaires : ils sont regroupés dans la rubrique « pages réservées aux membres du groupe », « travaux des sous-groupes » sur le site.

Le groupe a également choisi de diffuser sa réflexion au sein d'ateliers associés à des conférences:

- un atelier au cours de la conférence CFD à Hammamet (Tunisie) en octobre 2002;
- un atelier au cours de la plate-forme AFIA à Laval, organisé conjointement avec le groupe A3CTE, le 4 juillet 2003.

Les programmes de ces ateliers ainsi qu'une partie des communications exposées constituent la rubrique « productions » du site.

Un autre des axes de travail d'ASSTICCOT a été de monter une réponse (déclaration d'intention, EoI) à l'appel européen dans le cadre du 6° PCRD, de type réseau d'excellence (NoE). Ce réseau aurait eu pour thématique « SemTech : Acquiring specific semantic knowledge for an access to textual content». Ce projet a insufflé une double dynamique au groupe : il a permis d'établir ou de renforcer des contacts au niveau européen (plus de 80 équipes contactées), et surtout de constater au passage combien était positif l'écho que recevait notre projet scientifique. Malheureusement, cette déclaration d'intention n'a pas été donné lieu à une proposition de réseau d'excellence car elle avait peu de chances d'aboutir. Le texte de cette déclaration d'intention est visible à la rubrique «pages réservées aux membres du groupe », « projet européen » du site.

Enfin, au cours des 18 mois de son fonctionnement, notre AS a rendu compte de ses projets et de ses avancées auprès du département STIC du CNRS, directement lors de la journée consacrée aux AS le 17 juin 2002, et indirectement auprès du réseau auquel elle est rattachée, le RTP 33 «document numérique » RTP-DOC. Les présentations faites lors de ces journées forment la rubrique «Assticcot et le RTP-DPC», « présentations d'ASSTICCOT au RTP-DOC » sur le site.

Le financement reçu, de l'ordre de 70 k€, a servi essentiellement à payer les missions des personnes participant aux réunions et aux ateliers, à la gestion du site web et à des investissements pour certains laboratoires. La justification des dépenses ainsi que l'attribution des crédits aux laboratoires sont détaillées dans l'annexe 4.

6 Première formulation du champ d'étude et des questions retenues pour confronter différents travaux de recherche en cours ou à envisager.

Ce projet visait à formuler et à explorer toutes les problématiques abordées dès que l'on cherche à rendre compte du contenu de textes et à organiser les connaissances qui peuvent en être tirées dans des structures plus ou moins formelles, que nous pouvons appeler « ressources terminologiques » (terminologies, thesaurus, bases terminologiques, ontologies, etc.)

La constitution du corpus correspond ainsi à la première étape de l'analyse. Dans un second temps, c'est l'objectif lui-même qui guide l'étude et l'interprétation ; dans un tel processus, l'application n'est pas une simple utilisation de données, elle devient première et joue un rôle tout au long du processus d'analyse. La vision envisagée était donc clairement ascendante puisqu'elle partait d'un matériau textuel pour élaborer des modèles mais elle prenait en compte très tôt l'objectif de l'étude. Dans une telle perspective, les questions qui se posaient ont été organisées en six thématiques, théoriques et appliquées, qui ont constitué un premier maillage de la réflexion.

1. Problèmes théoriques

Relation entre sens et information, sens et connaissance.

Rôle des corpus dans l'histoire de la discipline.
Position de la discipline par rapport à l'utilisation des corpus.
Lien entre vision ascendante et vision descendante.

2. **Définition du besoin**

Quelle est sa nature : appliqué vs théorique ?

Sur l'axe corpus / processus de dépouillement de corpus / ressource terminologique cible, quel est l'objet d'étude ? qu'est-ce qui est considéré comme un moyen pour l'étudier ?

3. Modèles

Quels modèles sont adéquats lorsqu'on prend les textes comme sources de connaissances, ou lorsqu'on vise des applications terminologiques ou ontologiques ?

Font-ils appel aux notions de concept ? de terme ? de relation ? en leur donnant quel sens ?

Quel est le statut des données recueillies dans le modèle : générique ? spécifique ?

Est-ce que le fait de partir de corpus a un impact fort sur la structuration des données, ou est-ce la nature de l'application visée qui est plus forte ?

4 - **Méthode**

Décrire l'approche retenue, en quoi elle consiste, à quels principes elle fait appel, ce qu'elle exploite du texte et ce qu'elle laisse de côté.

Méthode manuelle vs automatique (rôle des outils) ?

Prise en compte de l'application : A quel moment intervient l'application ? Sur quoi intervient-elle : choix du corpus, choix des outils, techniques, méthodes utilisés, nature du modèle ?

Quels sont les avantages et les limites connus de cette approche ?

D'autres approches complémentaires sont-elles utilisées ? ont-elles été comparées à cette approche ? quel en est le bilan ?

Quel est le spectre d'application de cette méthode (pour quel type de corpus et d'application cible est-

elle valide)?

5-**Evaluation des résultats**

Comment est-elle faite ?
Rôle des experts ? Rôle de l'application
Critères retenus : coût, qualité des modèles, performances de l'application, etc.
Quels sont les résultats théoriques et techniques acquis, considérés comme mûrs ?
Quelles sont les réalisations majeures, démonstratives (s'il y en a) ?
Vers quelles communautés ces résultats sont-ils communiqués / transférés ?
Sur quoi portent les projets de recherche en cours en France ?
Quels sont les points difficiles non résolus ? pourquoi ?
En quoi ces résultats sont-ils facilement maintenus et mis à jour ?
Sont-ils réutilisables ? spécifiques ?

6- **Corpus**

Réflexion sur la constitution du corpus.
Comment est constitué le corpus par rapport à l'application ?
Comment est constitué le corpus par rapport à la méthode mise en œuvre ?
Typologie des corpus, genre textuel.
Nature des connaissances qui peuvent/ne peuvent pas être tirées des textes.

7 Liste des participants

Nom-Prénom	Discipline	Laboratoire	Courriel
AMAR Muriel	Sciences de l'Information	IUT Paris V	muriel.amar@iut.univ-paris5.fr Tel:01.44.14.45.75 Fax:01.44.14.45.73
AUSSENAC-GILLES Nathalie	Informatique IC Section O7	IRIT, UMR 5505 CNRS	aussenac@irit.fr Tel: 05 61 55 82 93 Fax: 05 61 55 62 58
BIEBOW Brigitte	Informatique IC Section O7	LIPN UMR 7030 CNRS	brigitte.biebow@lipn.univ-paris13.fr Tel:01 49 40 36 09 Fax:01 48 26 07 12 http://www-lipn.univ-paris13.fr/~biebow
BISSON Gilles	Informatique Apprentissage Section 07	LEIBNIZ	gilles.bisson@imag.fr Tel:04 76 57 46 03 Fax:04 76 57 46 02 http://www-leibniz.imag.fr
BOURIGAULT Didier	Linguistique, TAL	ERSS, UMR 5610 CNRS	didier.bourigault@univ-tlse2.fr Tel:05 61 50 36 93 Fax:05 61 50 46 77
BOUVERET Myriam	Linguistique corpus	LIUM, EA1015 CNRS	Myriam.Bouveret@univ-lemans.fr myriam_bouveret@yahoo.com Tel:02 43 83 31 Fax:02 43 83 37 75
CANDEL Danielle	Linguistique	UMR 7597	candel415@aol.com dcandel@linguist.jussieu.fr
CHARLET Jean	Informatique IC section 07	SIM, AP-HP	jc@biomath.jussieu.fr Tel:+33 1 45 83 67 28 Fax:+33 1 45 86 56 85
CHAUDIRON Stéphane	Sciences de l'Information	CRIS	stephane.chaudiron@u-paris10.fr stephane.chaudiron@technologie.gouv.fr Tel : +33 1 55 55 80 37 Fax : +33 1 55 55 83 58 http://www.u-paris10.fr
CONDAMINES Anne	Linguistique section 34	ERSS, UMR 5610 CNRS	acondami@univ-tlse2.fr Tel:05 61 50 36 08 Fax:05 61 50 46 77
DAVID Sophie	Linguistique Section 34	CNRS UMR 7114 "Modèles Dynamiques, Corpus"	sophie.david@u-paris10.fr Tel:01 40 97 47 33
DELAVIGNE Valérie	Linguistique	DYALANG UMR 6065	valerie.delavigne@normandnet.fr Tel : 02 35 14 60 56 Fax : 02 35 14 69 40
DESPRES Sylvie	Informatique IC	EHEI	sd@math-info.univ-paris5.fr
DIENG-KUNTZ Rose	Informatique	INRIA Sophia Antipolis	rose.dieng@sophia.inria.fr Tel : 04 92 38 78 10 Fax : 04 92 38 77 83
HOLZEM Maryvonne	Sciences de l'Information	DYALANG UMR 6065	maryvonne.holzem@univ-rouen.fr Tel : 02 35 14 60 56 Fax : 02 35 14 69 40
KASSEL Gilles	Informatique IC	LaRIA, EA 2083 CNRS	kassel@laria.u-picardie.fr Tel : 06 80 42 38 87
KAYSER Daniel	Informatique IA	LIPN UMR 7030	daniel.kayser@lipn.univ-paris13.fr
LAINE-CRUZEL Sylvie	Sciences de l'Information	RECODOC	slaine@univ-lyon1.fr Tel : 04 72 43 13 91 http://dist.univ-lyon1.fr
LALLICH-BOIDIN Geneviève	Sciences de l'Information	RECODOC	genevieve.lallich-boidin@univ-lyon1.fr Tel : +33 4 72 44 58 34 http://dist.univ-lyon1.fr
MINEL Jean-Luc	Informatique TAL	CAMS - ISHA	minel@msh-paris.fr Fax : 01 44 39 89 51 http://www.lalic.paris4.sorbonne.fr/~minel

MOTHE Josiane	Informatique Recherche d'Information	IRIT UMR 5505	josiane.mothe@irit.fr Tel:+33 5 61 55 63 22 Fax:05 61 55 62 58 http://www.irit.fr/~josiane.mothe
NAZARENKO Adeline	Informatique TAL	LIPN UMR 7030	adeline.nazarenko@lipn.univ-paris13.fr
NEDELLEC Claire	Informatique Apprentissage	MIG	nedellec@versailles.inra.fr Tel:+33 1 30 83 33 53 Fax: +33 1 30 83 33 59
NORMAND Sylvie	Linguistique de Corpus	DYALANG IUFM Créteil	sylvie.normand@creteil.iufm.fr
SZULMAN Sylvie	Informatique IC	LIPN UMR 7030 CNRS	ss@lipn.univ-paris13.fr Tel:+33 1 49 40 36 09 Fax:+33 1 48 26 07 12 http://www-lipn.univ-paris13.fr/~szulman
TANGUY Ludovic	Linguistique TAL	ERSS, UMR 5610 CNRS	ludovic.tanguy@univ-tlse2.fr
TOUSSAINT Yannick	Informatique TAL	LORIA UMR 7503 CNRS	yannick.toussaint@loria.fr http://www.loria.fr/~yannick Tel [33] 03 83 59 20 91 Fax [33] 03 83 41 30 79
ZWEIGENBAUM Pierre	Informatique TAL	DIAM/SIM	pz@biomath.jussieu.fr Tel : 01 45 83 67 28 Fax : 01 45 86 56 85 http://www.biomath.jussieu.fr/~pz

8 BILAN FINANCIER DÉTAILLÉ

Récapitulatif :

Montants en €	ERSS	IRIT	DYALANG	LIPN	TOTAL
Dotation	10671,43	10 671,43	4573,47	4573,47	30 489,8
Part consommée	Missions 5307,00 CFD 1395,00 Gestion 1220,00	Missions 5558,78 CFD 456,06 Site Web 2142,00 Repro. 400,00 Gestion 1220,00	Missions 662 CFD 1771,00 Invest. 1545,00 Gestion 594,38	Missions 814,36 Gestion 1371,00	Missions 12342,14 CFD 3622,06 Site Web 2142,00 Repro 400,00 Invest 1545,00 Gestion 4405,00
Crédit engagés pour fin 2003	Missions 2749,43	Missions 895,00	0	Missions 738,11 Invest. 1650,00	Missions 4372,54 Invest. 1650,00 Reliquat 9,10

Justification :

Le financement reçu, de l'ordre de 70 k€, a servi essentiellement à payer les missions des personnes participant aux réunions et aux ateliers, à la gestion du site web et à des investissements pour certains laboratoires, justifiés par la nécessité de machines pour faciliter les échanges et la production de documents. Les crédits engagés pour la fin 2003 correspondent à des missions de participation à des groupes de travail ainsi qu'à des conférences ou workshops internationaux pour promouvoir les travaux de l'AS et donner une suite aux contacts pris pour la déclaration d'intention de réseau européen.