Construction d'ontologies à partir de textes

Didier Bourigault (1) et Nathalie Aussenac-Gilles

(1) ERSS – CNRS & Université Toulouse le Mirail
5, allées Antonio Machado
31 058 Toulouse Cedex 1
didier.bourigault@univ-tlse2.fr
(2) IRIT – Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse Cedex 4
aussenac@irit.fr

Résumé – Abstract

Cet article constitue le support d'un cours présenté lors de la conférence TALN 2003. Il défend la place du Traitement Automatique des Langues comme discipline clé pour le développement de ressources termino-ontologiques à partir de textes. Les contraintes et enjeux de ce processus sont identifiés, en soulignant l'importance de considérer cette tâche comme un processus supervisé par un analyste. Sont présentés un certain nombre d'outils logiciels et méthodologiques venant de plusieurs disciplines comme le TAL et l'ingénierie des connaissances qui peuvent aider l'analyste dans sa tâche. Divers retours d'expérience sont présentés.

This paper gathers the notes of a tutorial. We advocate in favour of the role of Natural Language Processing as a key discipline for the development of terminological and ontological resources from texts. The constraints and challenges of this process are identified, and lead to underline this task as a supervised processes carried out by an analyst. We present several software and methodological tools from NLP and knowledge engineering that can be use for to assist the analyst. Our suggestion rely on various experience feed-back.

Keywords – Mots Clés

Extraction de termes, extraction de relations, Terminologie, Ontologies, Ingénierie des connaissances, méthode, modélisation de connaissances, interdisciplinarité.
Term extraction, relation extraction, Terminology, ontologies, Knowledge Engineering, method, knowledge modelling, crossdisciplinarity.

1 Introduction

Dans cet article, nous développons les grandes lignes du cours présenté lors de la dixième conférence *Traitement Automatique des Langues* le 14 juin 2003 à Batz-sur-Mer. Ce cours fait suite aux tutoriels donnés en juin 2000 lors de la conférence *Ingénierie des connaissances* (IC 2000, Toulouse) et en janvier 2002 lors de la conférence *Reconnaissance des Formes et Intelligence Artificielle* (RFIA 2002, Angers), au cours desquels nous avons eu l'occasion de présenter les outils développés en Traitement Automatique des Langues (TAL) aux membres de la communauté d'Ingénierie des Connaissances (IC). L'objectif du présent cours est, symétriquement, de présenter aux chercheurs de la communauté Traitement Automatique des Langues les enjeux, pratiques et théoriques, l'utilisation de certains outils de TAL dans une perspective d'ingénierie des connaissances, ceci pour encourager les travaux interdisciplinaires autour de la problématique de construction de ressources termino-ontologiques (RTO)[1], à partir de textes. Cette problématique constitue en effet un nouvel enjeu important aussi bien pour le Traitement Automatique des Langues que pour l'Ingénierie des Connaissances. Les systèmes de traitement de l'information qui doivent fonctionner dans des domaines de connaissances spécialisés ne peuvent être efficaces que s'ils s'appuient sur des ressources termino-ontologiques, construites pour le domaine et l'application concernés. Les recherches et réalisations en TAL et en IC doivent être menées de façon pluridisciplinaire, pour, d'une part, développer les outils de TAL pertinents pour la tâche de construction de RTO à partir de textes, et, d'autre part, élaborer des méthodes d'acquisition des connaissances à partir de textes qui spécifient comment utiliser les outils de TAL et les environnements de modélisation des connaissances dans le contexte de la construction de RTO. Au delà, mais cet aspect sera moins développé dans ce cours, il s'agit de s'interroger sur le statut de la langue écrite comme révélateur de connaissances, dès lors que l'on veut y accéder au moyens d'outils informatiques.

2 Des ressources à construire variées, des outils génériques

2.1 RTO et systèmes de traitements de l'information

A la suite à l'utilisation généralisée des outils de bureautique, à l'internationalisation des échanges et au développement d'Internet, la production de documents sous forme électronique s'accélère sans cesse. Or pour produire, diffuser, rechercher, exploiter et traduire ces documents, les systèmes de gestion de l'information ont besoin de ressources termino-ontologiques, qui 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 :

- bases de données terminologiques multilingues classiques pour l'aide à la traduction,

[1] Dans ce cours, nous nous efforcerons d'utiliser systématiquement cette expression plutôt que le terme très en vogue d'ontologie, adopté dans le titre du cours pour des raisons de concision. Ce choix terminologique sera justifié plus loin dans cet article.

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

Du côté de la recherche, chacune de ces ressources est prise en charge par une discipline différente. La terminologie focalise ses recherches, depuis l'avènement des outils de bureautique, sur les bases de données terminologiques destinée aux traducteurs humains. Les sciences de l'information et de la documentation concentrent leurs réflexions sur les thesaurus et langages de classification ou langages documentaires, exploités par les documentalistes pour indexer et classer les éléments de fonds documentaire. En informatique, le domaine de la recherche d'information (RI) s'intéresse à des thesaurus d'un type différent, conçus pour limiter le bruit et augmenter le rappel des outils informatiques de recherche d'information. L'intelligence Artificielle et l'Ingénierie des Connaissances travaillent sur les ontologies formelles qui constituent le cœur des systèmes à base de connaissances. Ces différentes disciplines développent de façon relativement autonome et cloisonnée des recherches spécifiques sur ces différents types de ressources. Or, sous la pression des besoins et des applications, elles sont amenées à considérer que, pour des raisons de pertinence et d'efficacité, les ressources lexicales et/ou conceptuelles qu'elles doivent construire et exploiter peuvent ou doivent être construites à partir de sources textuelles. Elles sont donc naturellement amenées à procéder à un rapprochement interdisciplinaire, dont le Traitement Automatique des Langues peut être le catalyseur, en tant que pourvoyeur de méthodes et outils de construction de RTO à partir de textes.

En terminologie, au cours des années 80, un rapprochement avec l'informatique s'est opéré avec le développement de la microinformatique. 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 ? Depuis la fin des années 90, la terminologie classique voit 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 multiplication des types de ressources terminologiques 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. Depuis le milieu des années 90, un courant de recherche se développe autour de la terminologie textuelle, qui préconise la

construction de terminologies à partir de textes, et qui sollicite le TAL pour des méthodes et outils d'analyse de corpus (Slodzian, 2000). En Intelligence Artificielle, une évolution importante du domaine s'est produite de façon concomitante et parallèle à ce renouvellement théorique et méthodologique en terminologie. L'échec relatif des réalisations en IA a conduit à remettre en cause l'hypothèse qui était à la base du développement des systèmes experts, selon laquelle l'expert d'un domaine serait le dépositaire d'un système conceptuel qu'il suffirait de mettre au jour, en interrogeant l'expert ou en l'observant au travail. L'Ingénierie des Connaissances (IC) s'est alors imposée comme une direction de recherche en IA, avec pour ambition de résoudre les difficultés soulevées par la construction des systèmes experts, et de proposer des concepts, méthodes et techniques permettant d'acquérir et de modéliser les connaissances dans des domaines se formalisant peu ou pas. L'IC s'intéresse en particulier au processus de construction d'ontologies formelles pour les systèmes à base de connaissances ou pour l'interopérabilité entre systèmes dans le Web sémantique. Elle préconise elle aussi que, dans certains contextes, ce processus s'appuie sur l'analyse de corpus de textes. Elle sollicite le TAL pour des outils rendant possible et efficace la tâche de construction d'ontologies à partir de textes.

Des sollicitations analogues émanent aussi d'autres disciplines, comme les sciences de l'information et de la documentation. Au sein même du domaine du Traitement Automatique des Langues, certaines applications, comme la traduction automatique, la recherche d'information ou l'extraction d'information, ont besoin de ressources termino-ontologiques, Le TAL est donc ainsi doublement concerné par la problématique de la construction de RTO à partir de textes, en tant que consommateur de ressources et en tant que pourvoyeur d'outils pour les construire. Le TAL se trouve donc à la convergence de demandes émanant de disciplines diverses et concernant la mise à disposition d'outils et de méthodes d'analyse de textes pour la construction de ressources termino-ontologiques. Il peut adopter ainsi une position décalée par rapport à chacune de ces disciplines et saisir, grâce à cet angle de vue privilégié, les proximités et les différences entre des différents types de ressources, avec une objectivité et un recul, que ne peuvent avoir ces disciplines seules. En ce sens, le TAL peut favoriser le décloisonnement de ces disciplines et encourager le rapprochement pluridisciplinaire, autour d'une réflexion sur la notion de ressource termino-ontologique. Cette réflexion doit permettre de mettre en évidence is les ressemblances et les particularités de ces différents types de ressources, de façon à spécifier les types d'outils d'analyse relativement génériques et utilisables pour une large gamme de ressources et de contextes d'exploitation.

2.2 Ontologie, terminologie, thesaurus, …

Le TAL se trouve donc face à des disciplines chacune préoccupée par le problème de la construction de ressources termino-ontologiques de types différents, puisque destinées à des usages différents. Dans ce contexte de sollicitations diversifiées, il est non pertinent pour le TAL de se lancer dans une réflexion théorique visant à caractériser formellement et de façon générique ce qu'est une ressource termino-ontologique. L'approche consiste plutôt à mettre en perspective les différentes définitions travaillées par ces disciplines. L'objectif est de saisir en quoi les caractéristiques spécifiques de ces différents types de ressources dépendent des contextes applicatifs, pour finalement identifier ce qui différencie et, surtout, ce qui rapproche ces différents types de ressources. Il est alors possible de spécifier les différents types d'outils génériques de TAL dont il convient de promouvoir le développement.

Les réflexions sur les ontologies se sont d'abord développées en informatique (intelligence artificielle, sciences de la gestion), dans le cadre de travaux qui avaient comme objectif final la spécification de systèmes informatiques, avec plus particulièrement à l'origine la volonté de pouvoir réutiliser des composants génériques d'une application à une autre, ou encore de favoriser la communication entre différentes applications. C'est le cas encore des travaux menés en Ingénierie des Connaissances ou en représentation des connaissances autour des Systèmes à Base de Connaissances et du Web sémantique. Dans ce contexte, une ontologie est une conceptualisation des objets du domaine selon un certain point de vue, imposé par l'application. Elle est conçue comme un ensemble de concepts, organisés à l'aide de relations structurantes, dont la principale, celle avec laquelle est construite l'ossature de l'ontologie, est la relation *is-a*. Cette conceptualisation est écrite dans un langage de représentation des connaissances, qui propose des « services inférentiels » (classification de concept, capacité de construire des concepts définis à partir de concepts primitifs, etc.). A l'opposé, pour les thesaurus, un haut degré de formalisation et des services d'inférence ne sont pas nécessaires. Les thesaurus sont organisés avec les classiques relations d'hyperonymie et de synonymie, auxquelles s'ajoute la relation *voir aussi*. Néanmoins, il faut bien distinguer les thesaurus selon qu'ils sont exploités par des indexeurs et documentalistes humains, ou par des systèmes informatiques. Au cours d'une tâche d'indexation, pour choisir les meilleurs descripteurs, les agents humains procèdent à des interprétations et des inférences, qui s'appuient sur leur connaissance du domaine et des utilisateurs, connaissances implicites qui ne sont pas consignées dans le thesaurus. Les systèmes d'indexation automatique ne peuvent approcher de tels comportements intelligents qu'à condition que ces connaissances soient autant que possible explicitées et représentées dans les thesaurus, qui tendent ainsi à se rapprocher des ontologies de l'Ingénierie des Connaissances.

Le principal critère de discrimination entre RTO est le type de données d'entrée du système de traitement de l'information qui exploite la RTO. Selon que ces systèmes traitent de l'information de nature textuelle ou non, les caractéristiques des RTO vont être relativement différentes. Si le système analyse des entrée en langue naturelle, la première exigence est qu'il soit capable de reconnaître sous des formes linguistiques différentes des occurrences de la même unité et, inversement, de reconnaître des unités différentes sous une même forme. Il doit pouvoir gérer, aussi bien que l'application l'exige, les phénomènes de synonymie, de paraphrase, de variabilité linguistique aux niveaux morphologique ou syntaxique ou lexical, présents en masse dans les textes en langues naturelles (Zweigenbaum, 1999). Ceci n'est possible que si des règles de correspondance sont répertoriées dans la RTO que va exploiter le système. Une des tâches de l'analyste qui construit la RTO est donc de décrire des liens entre des motifs textuels et des unités de traitement, unités qui seront ensuite exploitées pour effectuer les traitements assignés aux système (classification de document, expansion de requête, extraction d'information, etc.). Quand les motifs textuels ont la structure de noms ou syntagmes nominaux, ils sont naturellement désignés sous le nom de termes. Les unités de traitement sont les concepts. C'est la raison pour laquelle nous parlons de *ressources termino-ontologiques*. De ce point vue, le concept peut être vu comme une classe d'équivalence de termes, ou plus généralement de motifs textuels, modulo les contraintes de l'application cible : deux motifs sont jugés équivalents, ou synonymes, en fonction de traitement que doit effectué par le système. Le concept est un mode de regroupement de termes. Ceci n'est pas incompatible avec sa fonction de regroupement d'objets (informatiques) du domaine qui lui est assignée dans les ontologies de l'Ingénierie des Connaissances. Le système de traitement de l'information dispose donc pour traiter de la synonymie de règles d'appariement qui

exploitent les liens termes/concepts présents dans la RTO. Il dispose de règles analogues pour le traitement de la polysémie, de l'homographie.

Si l'application cible n'est pas une application textuelle, l'analyse des textes n'en est pas moins fondamentale. Même s'il s'agit de construire une ontologie pour un système informatique, dont les données d'entrée ne seront pas textuelles, mais numériques, par exemple des résultats de mesures de capteur, l'analyse de textes et la description du vocabulaire sont néanmoins primordiales pour la construction de l'ontologie. En effet, l'analyse des textes sert d'indicateur à l'organisation d'un système conceptuel et donc à la mise en relation de concepts, et, par ailleurs, le choix des étiquettes de concepts doit être judicieux pour assurer l'interprétabilité et l'intelligibilité du système, ainsi que la maintenance de l'ontologie (Bachimont, 2000).

Cette position constructiviste et fonctionnelle des notions de terme et de concept s'éloigne quelque peu des positions référentialistes et fixistes - le terme comme étiquette de concept -, qui sont classiquement adoptées dans les domaines de l'Intelligence Artificielle, de la terminologie ou du Traitement Automatique des Langues, disciplines qui ont longtemps été largement influencées par une sémiotique du signe fondée sur la triade terme/concept/référent (Rastier, 1991). La conception classique pose que le terme existe en tant que représentant linguistique d'un concept faisant partie d'un système conceptuel unique et stable caractérisant a priori le domaine. Mais 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 termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées. L'ensemble de ces constats empiriques appelle à un renouvellement théorique de la terminologie (Rastier, 1995) (Slodzian, 2000). A rebours de la conception fixiste et apriorique, on peut voir le terme et le concept comme le *résultat* d'un processus d'analyse termino-conceptuelle. Un mot ou une unité complexe n'acquiert le statut de terme que par décision. Dans le cas qui nous concerne ici, cette décision est prise par l'analyste en charge de l'élaboration d'une RTO pour une application bien identifiée. Celui-ci définit son propre référentiel de décision. Il procède à un travail de *construction* d'*une* ressource termino-ontologique pour une application dans le domaine, et non de *découverte* de *la* terminologie du domaine. Ce travail est guidé par une double contrainte de pertinence :

- pertinence vis-à-vis du corpus. Il s'agit de retenir et de décrire des structures lexicales qui présentent des caractéristiques à la fois spécifiques au domaine et stables dans le corpus ;

- pertinence vis-à-vis de l'application visée. Les unités finalement retenues doivent l'être en fonction de leur utilité dans l'application visée, qui s'exprime en termes d'économie, de cohérence interne et d'efficacité.

3 Eléments méthodologiques

3.1 Des outils d'aide

Devant la masse des données à analyser et étant donnés les délais de réalisation imposés, les disciplines concernées par la construction de RTO se tournent vers le TAL pour des outils informatiques d'analyse de corpus.

Les travaux de conception d'outils de TAL pour la construction de RTO doivent développer une réflexion méthodologique sur l'activité de construction elle-même. Doit s'imposer d'emblée le postulat que cette activité est avant tout une activité humaine, intellectuelle, menée par un individu que nous nommerons ici « analyste ». Dans un projet de construction de RTO, les contraintes sont multiples et multiformes, les choix à effectuer nombreux et de types divers, et ces choix comme ces contraintes, de type heuristique, sont difficilement explicitables. ont jusqu'ici été peu explicités. Par conséquent cette tâche ne peut en rien se limiter à l'élaboration automatique d'un réseau de termes et de concepts par quelque outil que ce soit. Nous défendons que la contribution du TAL doit être la fourniture d'outils d'aide pour l'analyste. Les recherches doivent se développer dans le paradigme de la coopération, et non celui de l'automatisation, même partielle, et il faut assumer, dans une perspective ingénierique, le rôle central de l'analyste. Autant les outils de TAL consommateurs de ressources termino-ontologiques doivent et peuvent approcher l'automaticité, autant les outils de TAL d'aide à la construction de RTO exigent l'intervention d'un agent humain.

Au-delà des difficultés techniques traditionnellement liées au développement d'outils en TAL, il existe une tension particulière propre au développement d'outils d'aide à la construction de RTO : il s'agit de concilier le caractère ad hoc des ressources à construire avec les outils, avec les contraintes de généricité, transportabilité, reproductibilité, qu'impose le développement de la recherche. Autrement dit, il faut chercher à développer des outils de TAL relativement génériques quant au domaine et au type d'application, pour des utilisations elles très ciblées quant à ces deux points.

Rapidement, on peut classer les types d'outils à construire selon deux axes. Du point de vue fonctionnel, on peut distinguer les outils d'aide à l'acquisition de termes et les outils d'aide à la structuration de termes et au regroupement conceptuel (section 4). Du point de vue du mode d'utilisation, on peut distinguer les outils qui fonctionnent « en batch » (ils traitent l'ensemble du corpus, puis fournissent les résultats à l'analyste), et les outils interactifs. Par ailleurs, puisque les décisions prises par l'expert s'appuient *in fine* sur l'analyse de contextes dans le corpus, à côté des outils de traitement massif de corpus, il faut fournir à l'analyste des moyens d'*accès au texte* (concordanciers, outils de navigation hypertextuelle, etc.).

3.2 Rôle de l'analyste

Dans l'idéal, la personne chargée de construire la RTO, l'analyste, devrait avoir à la fois des compétences métier, des compétences en modélisation des connaissances et en linguistique et des compétences en informatique. Ce profil fait-il de l'analyste un oiseau rare ? Dans la réalité, il faut mettre en place une collaboration entre acteurs de spécialités différentes. Plusieurs sortes de situations peuvent être rencontrées. Pour les applications à forte dimension

cognitive, l'expérience montre que l'efficacité maximale peut être atteinte quand la construction de la RTO est assurée par un spécialiste métier, passionné par les problèmes de langue et de connaissance, ou formé à ceux-ci, qui comprend bien les spécifications de l'application cible et qui est capable de dialoguer avec les informaticiens qui la développent. A l'opposé, certaines applications, de type documentaire, ne requièrent pas une implication forte des spécialistes et la construction de la RTO peut être réalisée par des personnes ayant le profil et l'expérience de documentaliste ou de terminologue. Dans tous les cas, l'intervention d'un analyste médiateur est nécessaire quand l'application exige la participation de plusieurs spécialistes.

3.3 Place du corpus

Dans un projet de construction de RTO à partir de textes, la tâche de construction du corpus est à la fois primordiale et délicate. Puisque, d'une part, le corpus est la source d'information essentielle pour tout le processus de construction de la RTO et que, d'autre part, il restera, une fois le processus achevé, l'élément de documentation de la ressource construite, il doit être composé avec un maximum de précautions méthodologiques. Dans ce domaine, il n'est hélas pas encore possible de définir a priori des instructions méthodologiques très précises pour encadrer la tâche de sélection des sources textuelles qui viendront constituer le corpus. 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. Il convient en effet de s'assurer auprès des spécialistes que les textes choisis ont un statut suffisamment consensuel pour éviter toute remise en cause ultérieure de la part d'utilisateurs ou de leur part. Par ailleurs, il convient de prévoir d'emblée une boucle de rétroaction au cours de laquelle une première version du corpus sera modifiée et enrichie en fonction d'une première phase d'analyse des résultats fournis par les outils de TAL sur cette version initiale. Le critère de la taille est évidemment important, même s'il est impossible de donner un chiffre idéal. Le choix est ici encore un compromis. Le corpus doit être suffisamment « gros » pour justifier que des outils de traitement de la langue soient nécessaires pour le dépouiller de façon efficace. Mais il doit être suffisamment petit et/ou redondant pour pouvoir être appréhendé de façon globale par l'analyste, même à l'aide d'outils de TAL. Une fourchette entre 50 000 et 200 000 mots semble raisonnable. Les projets prenant le Web comme source de textes font rapidement exploser ces chiffres, posant par la même des problèmes spécifiques, comme celui de la définition d'un « échantillon » pertinent pour l'étude. Enfin, dans la majorité des cas, le corpus sera hétérogène dans le sens où il aura été constitué en rassemblant des textes d'origine variée. Il est alors absolument nécessaire de procéder à un balisage du corpus qui permettra aux outils d'analyse, et ainsi qu'à l'analyste, de repérer les différents sous-corpus pour procéder éventuellement à des analyses contrastives.

3.4 Utilisation de ressources existantes

On l'aura compris, nous ne nous intéressons ici ni aux ontologies génériques (« à la Cyc ») censées représenter un ensemble maximal de connaissances, de sens commun, ni aux ontologies formelles (au sens de Guarino) qui constitueraient un cadre référentiel universel et formellement valide, mais bien à des ressources termino-ontologiques exploitées par un système particulier de traitement de l'information dans un domaine particulier. C'est l'usage prévu de la ressource qui contraint et encadre sa construction. Pour autant, nous ne souhaitons

pas participer à une polémique sur l'opposition ontologies générales vs. ontologies spécialisées. Notre position est la suivante : il est primordial que les outils d'aide à la construction de RTO puissent recycler des données existantes afin de tirer le meilleur parti du patrimoine terminologique possédé par les entreprises et les institutions (Jacquemin, 1997). Pour une tâche de construction de RTO, il faut faire feu de tout bois, et chercher à exploiter autant que faire se peut toutes les ressources disponibles, et pas uniquement les textes. Sur le plan de la politique de la recherche, nous pensons qu'il est utile de promouvoir des travaux montrant l'utilité de ressources lexicales existantes (générales, comme la base WordNet ou des fichiers électroniques de synonymes, ou spécialisées, comme les grands thesaurus de la médecine comme UMLS) dans la perspective d'améliorer le rendement du couple analyste/outils de TAL. Nous sommes plus réservés sur la nécessité de dégager des financements lourds pour la réalisation de nouvelles ressources sémantico-conceptuelles de taille gigantesque, élaborées hors de toute spécification d'application cible. Il nous semble plus pertinent que soient encouragés des expériences d'évaluation, nécessairement très lourdes, proposant des protocoles expérimentaux capables de mettre en évidence à grande échelle les gains en temps et en qualité apportés par l'introduction d'une ontologie dans tel ou tel système de traitement de l'information par rapport au coût de la construction de cette ontologie (cf. section 6).

3.5 De la nécessité d'interface intégratrices

La tâche de construction d'une RTO est incrémentale et comporte de nombreux enchaînements d'essais/erreurs. Il faut des *interfaces* ergonomiques permettant une utilisation coordonnée et optimale des différents outils de traitement et de consultation du corpus de référence, par l'analyste qui construit une RTO, à l'instar de (Ait El Mekki et Nazarenko, 2002) pour la construction d'index d'ouvrages, de la plate-forme de modélisation TERMINAE pour la construction de terminologies et d'ontologies (Szulman et al., 2002). De façon plus générale, l'utilisation de ces différents outils doit être encadrée par une méthodologie précisant à quel stade du processus et selon quelles modalités il convient de les utiliser. En effet, la solution au problème de l'acquisition de ressources termino-ontologiques à partir de corpus ne réside pas uniquement en la fourniture d'un ou de plusieurs outils de traitement automatique des langues. La mise à disposition de tels outils doit s'accompagner d'une réflexion méthodologique poussée, conduisant à la réalisation de guides méthodologiques et de plates-formes logicielles intégratrices permettant la mise en œuvre efficace des outils proposés. Cette nécessité appelle une coopération entre TAL et IC. Cette réflexion sur l'utilisation combinée de différents types d'outils d'analyse de textes en ingénierie terminologique est aussi très présente dans un certain nombre de travaux en ingénierie des connaissances (Charlet et al., 2000).

3.6 Une proposition méthodologique

A titre d'exemple, nous évoquons une proposition méthodologique intégrant l'utilisation de plusieurs outils de TAL et qui se veut une réponse possible aux différents problèmes évoqués : la méthode TERMINAE (Szulman et al., 2002). Cette méthode s'appuie sur des travaux représentatifs du courant français de travaux à la convergence entre terminologie,

linguistique, ingénierie des connaissances et intelligence artificielle[2]. Elle s'appuie sur les principes suivants :

- Partir de textes du domaine comme sources de connaissances : ils constituent un support tangible, rassemblant des connaissances stabilisées qui servent de référence et améliorent la qualité du modèle final ;
- Enrichir le modèle conceptuel d'une composante linguistique : l'accès aux termes et aux textes qui justifient la définition des concepts garantit une meilleure compréhension du modèle ;
- Utiliser des techniques et outils de TAL basés sur des travaux linguistiques : ils permettent l'exploitation systématique des textes et leurs résultats facilitent la modélisation ;
- Construire des ontologies « régionales », c'est-à-dire consensuelles dans un domaine et adaptées à une application, mais non universelles ;
- Appliquer des principes de modélisation systématiques pour assurer une bonne structuration des données et faciliter la maintenance de l'ontologie.

TERMINAE vise essentiellement la constitution de terminologies, réseaux conceptuels et ontologies. La méthode comprend quatre étapes, les trois dernières étant mises en oeuvre de manière cyclique. L'importance de chacune dépend du produit terminologique visé et des objectifs d'utilisation de ce dernier.

- La Constitution d'un corpus vise à choisir documents techniques, comptes rendus, livres de cours, etc. à partir d'une analyse des besoins de l'application.

- L'étude linguistique consiste à identifier des termes et des relations lexicales, en utilisant des outils de traitement de la langue naturelle (SYNTEX comme extracteur de termes, UPPERY comme outil d'analyse distributionnelle, Caméléon pour l'aide au repérage de relations par des patrons linguistiques, YAKWA comme concordancier).

- La normalisation sémantique conduit à définir dans un langage formel des concepts et des relations sémantiques que nous appelons terminologiques car provenant des termes et relations précédemment étudiés (Biébow & Szulman, 1999). Leur structuration en réseau s'appuie sur les résultats du dépouillement des textes tout en tenant compte de l'objectif d'utilisation de l'ontologie. Elle nécessite l'ajout de nouveaux concepts et relations dits de structuration.

- La formalisation permet de préciser, compléter et valider le modèle construit lors de la nor-malisation. L'analyste indique si les concepts sont primitifs ou définis, vérifie que les relations sont à la bonne place pour favoriser un héritage maximum, etc.

Le logiciel TERMINAE associé à la méthode fournit des aides pour toutes les étapes de l'analyse des textes à la formalisation. Il offre un support méthodologique qui permet d'évoluer progressivement et en conservant des liens des textes vers les niveaux linguistique

[2] Ce courant, animé au sein du GDR-I3 et de l'AFIA par le groupe TIA (http://www.biomath.jussieu.fr/TIA/) dont les auteurs font partie.

et conceptuel. Le logiciel assure donc une continuité entre les différentes formes de l'ontologie. Celle-ci passe d'un état proche d'une taxinomie de termes à un réseau conceptuel enrichi de relations et de concepts de structuration pour aboutir à une ontologie formelle. Elle est décrite dans un langage formel masqué à l'analyste qui permet de vérifier des contraintes de validité minimale.

4 Outils de TAL pour la construction de RTO

4.1 Une typologie fonctionnelle

Dans cette section[3], nous passons en revue un certain nombre de travaux de recherche sur le développement d'outils d'aide à la construction de RTO à partir de textes. Nous avons choisi de les présenter selon une typologie de fonctionnelle.

- *Acquisition de termes.* Une première classe regroupe les outils dont la visée est l'extraction à partir du corpus analysé de *candidats termes*, c'est-à-dire de mots ou groupes de mots susceptibles d'être retenus comme termes par un analyste, et de fournir des étiquettes de concepts. Ces outils diffèrent principalement quant au type de techniques mises en œuvre (syntaxique, statistique, autres).

- *Structuration de termes et regroupement conceptuel.* Les ressources termino-ontologiques se présentent rarement sous la forme d'une liste à plat. Des outils d'aide à la structuration d'ensembles de termes sont donc nécessaires. Dans cette classe, nous évoquerons, d'une part, des outils de classification automatique de termes, et, d'autre part, des outils de repérage de relation. Signalons que beaucoup d'outils d'extraction proposent déjà une structuration des candidats termes extraits.

4.2 Acquisition de termes

L'outil TERMINO est une application pionnière de l'acquisition automatique de termes (David et Plante, 1990). Construit sur la base de l'atelier FX, un formalisme pour l'expression de grammaires du langage naturel et un analyseur associé, TERMINO se focalise sur le repérage des syntagmes nominaux qui sont les seules structures supposées produire des termes. Les candidats termes extraits par TERMINO sont appelés " synapsies " d'après les travaux de Benveniste. La chaîne de traitement de TERMINO se compose d'une phase d'analyse morphosyntaxique suivie d'une phase de génération des synapsies à partir des dépendances entre tête et compléments rencontrés dans la structure de syntagme nominal retournée par l'analyseur. ANA est un outil d'acquisition terminologique qui extrait des candidats termes sans effectuer d'analyse linguistique (Enguehard et Pantera, 1995). Les termes sont reconnus au moyen d'égalités approximatives entre mots et d'une observation de répétitions de patrons. ACABIT extrait des candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé (Daille, 1994). ACABIT mêle des traitements linguistiques et des filtres

[3] Cette partie est extraite et adaptée de (Bourigault et Jacquemin, 2000) parue dans l'ouvrage *Industrie des langues* (Hermès) coordonné par J.-M. Pierrel . Une bibliographie mise à jour sera fournie lors du cours.

statistiques. L'acquisition terminologique dans ACABIT se déroule en deux étapes : (1) analyse linguistique et regroupement de variantes, au cours de laquelle n ensemble de transducteurs analyse le corpus étiqueté pour extraire des séquences nominales et les ramener à des candidats termes binaires ; (2) filtrage statistique, au cours duquel les candidats termes binaires produits à l'étape précédente sont triés au moyen de mesures statistiques. A l'instar d'ACABIT, LEXTER extrait des candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé (Bourigault, 1994). Il effectue une analyse syntaxique de surface pour repérer les syntagmes nominaux maximaux, puis une analyse syntaxique profonde pour analyser et décomposer ces syntagmes. Il est doté de procédures d'apprentissage endogène pour acquérir des informations de sous-catégorisation des noms et adjectifs propres aux corpus. Il organise l'ensemble des candidats termes extraits sous la forme d'un réseau. FASTR est un analyseur syntaxique robuste dédié à la reconnaissance en corpus de termes appartenant à une liste contrôlée fournie au système (Jacquemin, 1997). Les termes n'ayant pas toujours, en corpus, la même forme linguistique, le principal enjeu est de pouvoir identifier leurs variantes. FASTR est doté d'un ensemble élaboré de métarègles, qui lui permettent de repérer différents types de variation : les variantes syntaxiques, morpho-syntaxiques et sémantico-syntaxiques. L'environnement SYMONTOS (Velardi et al., 2001) propose des outils pour repérer des termes simples et complexes dans des textes et des critères pour décider de définir des concepts à partir de ces termes.

4.3 Structuration de termes et regroupement conceptuel

La gamme des outils d'aide à la structuration de terminologie est très large. Sont susceptibles d'émarger à cette catégorie un certain nombre de types d'outils qui n'étaient pas initialement conçus spécifiquement pour cette tâche, mais qui ont été développés pour des applications d'informatique documentaire ou d'extraction d'informations, par exemple. Nous balayons rapidement un spectre assez large, couvrant les outils de classification de termes sur la base de cooccurrences dans des textes ou dans des fenêtres, les outils de classification de termes sur la base de distributions syntaxiques et les outils de repérage de relations. Les outils de cooccurrence développés dans le domaine de la recherche d'information rapprochent des termes qui apparaissent fréquemment dans les mêmes (portions de) documents, et qui possèdent donc sans doute une certaine proximité sémantique. La technique de recherche de cooccurrents est déjà ancienne (à l'échelle de l'histoire de l'informatique) puisqu'elle a été promue très tôt en informatique documentaire pour permettre l'expansion de requêtes (Sparck Jones, 1971). Parmi les applications dans le domaine de l'acquisition terminologique, on peut citer le projet ILIAD (Toussaint et al., 1998), et les travaux de G. Lame (2002). Toujours dans le domaine de l'informatique documentaire, les travaux dans le domaine de la construction automatique de thesaurus peuvent être réinvestis dans des applications terminologiques. Par exemple, la chaîne de traitement développée par G. Greffenstette construit automatiquement des classes comportant des noms qui se retrouvent régulièrement comme arguments des mêmes verbes (Greffenstette, 1994). Ce repérage de la position argumentale des noms se fait grâce à l'exploitation d'un analyseur syntaxique de surface à large couverture. Ces techniques inspirées de la linguistique harrissienne, qui visent à rapprocher les termes qui ont des distributions syntaxiques analogues, sont à la base de nombreux travaux depuis plusieurs années (Assadi, 1998) (Habert et al, 1996) (Faure, 2000).

Les outils que nous venons d'évoquer visent à rapprocher des termes à partir d'une analyse globale de l'ensemble de leurs occurrences. Ils ne touchent que les termes fréquents, et donc

le plus souvent des noms simples, et proposent une simple relation d'équivalence (appartenance à une classe). A côté de ces outils qui travaillent sur les types comme regroupement des occurrences, on trouve les outils de repérage de relations, qui travaillent au niveau des occurrences elles-mêmes. Ces outils détectent en corpus des mots ou contextes syntaxiques répertoriés comme susceptibles de " marquer " telle ou telle relation entre deux éléments. Les travaux de M. Hearst, sur l'extraction automatique des liens d'hyperonymie, font figure de référence (Hearst, 1992). Les recherches sur ce thème se déclinent de multiples façons. L'un des enjeux principaux concerne la généralité des relations, et celles des marqueurs de relations. D'un côté, il existe probablement des relations que l'on jugera toujours pertinentes pour décrire un domaine de connaissance, par exemple les relations de type hiérarchique ou partitive, et des marqueurs pour ces relations eux aussi généraux (Garcia, 1998). A l'opposé, il est indéniable que chaque domaine est structuré par des relations qui lui sont spécifiques, et qu'il convient nécessairement de prendre en compte pour décrire le domaine. De plus même dans le cas de relations considérées comme générales, il est possible que les marqueurs susceptibles de conduire à les identifier diffèrent d'un corpus à l'autre. Se pose alors le problème de l'apprentissage inductif de ces marqueurs de relation. Un certain nombre de travaux en TAL et en IC sont consacrés à ce problème. Ils partent tous du même principe d'une recherche itérative alternée dans le corpus à la fois des marqueurs d'une relation donnée et des couples de termes qui entrent dans cette relation (Rousselot et al., 1996) (Séguéla et Aussenac-Gilles, 1999) (Morin, 1999) (Condamines et Rebeyrolles, 2000) (Maedche et Staab, 2000).

5 Trois retours d'expérience

5.1 Contextes

Les exemples, démonstrations et expérimentations proposés pendant le cours sont issus principalement de 3 expériences réelles de construction de RTO à partir de textes[4]. Ces trois expériences couvrent un spectre large de types de domaines et de types d'applications : la première expérience a été menée dans le domaine technique de la fabrication du verre (projet VERRE), avec comme application cible la classification de documents ; la deuxième expérience a été menée dans un domaine médical de la réanimation chirurgicale (projet REA), avec comme application cible le codage d'actes médicaux ; la troisième expérience a été menée dans le domaine juridique du Droit français codifié (projet DROIT), avec comme application cible l'aide à la reformulation de requêtes. Il s'agit à chaque fois de projets de Recherche et Développement, dans lesquels l'application cible n'est pas strictement spécifiée au départ du projet, comme cela devrait l'être dans un « vrai » projet industriel. On doit donc être prudent au moment de tirer des conclusions générales. Néanmoins, chacun de ces projets est allé à son terme, en ce sens qu'il n'a pas conduit à des RTO « jouets », mais à des ressources complètes qui sont ou seraient exploitables. Par ailleurs, chaque projet a permis de tester certaines hypothèses méthodologiques faisant ainsi progresser les recherches dans le domaine de l'acquisition des connaissances à partir de textes. C'est en multipliant ce type

[4] Cette partie est extraite et adaptée d'un article à paraître dans un numéro spécial de la Revue d'Intelligence Artificielle, coordonné par M. Slodzian et J.-M. Pierrel (Aussenac-Gilles & al, 2003)

d'expériences que l'on avancera sur la définition d'un cadre méthodologique relativement précis qui aille au-delà d'un simple recueil de bonnes pratiques et qui puisse satisfaire les exigences d'un transfert vers les applications industrielles.

5.1.1 Le projet VERRE : une ontologie dans le domaine de la fabrication et d'utilisation de la fibre de verre

Le premier projet vient répondre à une demande du centre de recherche du groupe Saint-Gobain. Au sein des différentes filiales du groupe, l'avance technologique et industrielle est primordiale pour conserver une place compétitive par rapport aux entreprises concurrentes. Les activités de veille documentaire et technologique jouent alors un rôle crucial, et font l'objet d'un outillage informatique de plus en plus performant. Parmi ces activités, une demande récurrente des documentalistes porte sur la définition d'un outil d'aide au repérage de nouveaux documents pertinents sur le Web (comme des brevets, des dépêches de presse, etc.) et à leur classement en fonction des domaines d'intérêt des ingénieurs qui les consultent. Or la plupart des outils de routage de documents s'appuient sur un réseau conceptuel d'autant plus performant qu'il est enrichi des connaissances et de la terminologie du domaine de l'entreprise. L'objectif du projet était donc de tester la faisabilité du développement d'une ontologie dans l'objectif de l'utiliser pour guider le classement de documents en fonction des profils des utilisateurs. Dans ce projet, les aspects méthodologiques étaient tout aussi importants que l'ontologie elle-même. L'étude a été menée par deux chercheurs de l'IRIT, A. Busnel pour l'analyse terminologique et ontologique, et N. Aussenac-Gilles sur les aspects méthodologiques. Un début d'ontologie (50 concepts, 20 relations) a été mis en forme à l'aide du logiciel de modélisation TERMINAE, à partir de l'analyse d'un corpus de langue anglaise composé de différents types de documents sur le domaine. Les logiciels de Traitement Automatique des Langues SYNTEX, UPERY et YAKWA ont été utilisés pour le dépouillement de ces corpus. Une proposition méthodologique utilisable dans le contexte de cette entreprise et pour ce type d'application a été mise en forme (Aussenac-Gilles & Busnel, 2002).

5.1.2 Le projet REA : une ontologie dans le domaine de la traumatologie en réanimation chirurgicale

Le deuxième projet a été encadré par M.-C. Jaulent et J. Charlet et a été mené à bien au sein de l'UFR Broussais-Hotel-Dieu. Le contexte est celui du codage des actes médicaux par les médecins. Pour leur activité de codage obligatoire, les praticiens s'aident d'un thésaurus de spécialité qui a été élaboré de façon à ce que les séjours de réanimation soient le mieux possible valorisés. Il est aujourd'hui reconnu que l'ambiguïté du thésaurus est une source d'erreurs et de disparités de codage. Dans un domaine particulier tel que la réanimation chirurgicale, on ne peut envisager de réaliser des outils informatiques d'aide au codage qu'après avoir préalablement organisé des objets du domaine, en fonction de la tâche à résoudre, par le biais d'une ontologie. L'objectif de ce deuxième projet était donc de construire une ontologie du domaine de la réanimation chirurgicale. Les outils de Traitement Automatique des Langues SYNTEX et UPERY ont été utilisés pour traiter un corpus de comptes rendus d'hospitalisation. Le travail a été réalisé par S. Le Moigno, médecin spécialiste, dans le cadre d'un stage de DEA en informatique médicale. L'ontologie comprend environ 2 000 concepts et 200 liens (Le Moigno et al., 2002).

5.1.3 Le projet DROIT : une ressource ontologique dans le domaine du Droit

Le troisième projet a été mené par G. Lame, au cours de sa thèse au Centre de Recherche en Informatique de l'Ecole des Mines de Paris (Lame, 2002). Ce centre de recherche a créé et héberge le site juridique droit.org, qui diffuse l'édition *Lois et décrets* du Journal Officiel de la République française, ce qui représente 95 000 documents (lois, décrets, arrêtés), ainsi que les codes du droit français (Code civil, Code pénal, etc.) et des textes européens (directives, règlements). L'objectif du travail était de tester l'intérêt et la faisabilité d'une approche consistant à intégrer une ontologie du Droit susceptible de faciliter l'accès au site par les utilisateurs. Le résultat est une ressource ontologique de très large couverture, couvrant tous les domaines du Droit, constituée d'environ 130 000 termes et 200 000 liens. Cette ressource est utilisée comme support pour un système d'expansion de requêtes : à un mot posé par l'utilisateur, le système propose tous les termes reliés à ce mot dans la ressource et laisse l'utilisateur choisir ceux qu'ils souhaitent retenir pour modifier sa requête. Cette ressource a été construite en utilisant les résultats bruts, sans aucun filtrage manuel, de différents outils ou techniques de Traitement Automatique des Langues (SYNTEX, cooccurrence statistique, UPERY), obtenus par analyse d'un corpus constitué de l'ensemble des Codes de la législation française.

5.2 Trois outils de TAL pour la construction de RTO à partir de textes

5.2.1 Extraction de termes : SYNTEX

Dans les trois projets, les résultats de l'outil SYNTEX ont été utilisés. SYNTEX (Bourigault et Fabre, 2000) est un analyseur syntaxique de corpus. Il existe actuellement une version pour le français, qui a été utilisée dans les projets REA et DROIT, et une version pour l'anglais, qui a été utilisée dans le projet VERRE. Après l'analyse syntaxique en dépendance de chacune des phrases du corpus, SYNTEX construite un réseau de mots et de syntagmes (verbaux, nominaux, adjectivaux), dit « réseau terminologique », dans lequel chaque syntagme est relié d'une part à sa tête et d'autre part à ses expansions. Les éléments du réseau (mots et syntagmes) sont appelés « candidats termes ».

A chaque candidat terme sont associées un certain nombre d'informations numériques, sur lesquelles l'utilisateur peut se baser pour organiser son dépouillement :

- *fréquence* : c'est le nombre d'occurrences du candidat terme détectées par le logiciel dans le corpus. L'interface d'analyse des résultats permet à l'analyste d'accéder à l'ensemble des contextes d'apparition du candidat terme dans le corpus. Cet accès au texte est d'autant plus crucial que l'utilisateur n'est pas un spécialiste du domaine.

- *productivité en Tête (resp. Expansion)* : c'est le nombre de « descendants en Tête » (resp. « descendants en Expansion ») du candidat terme, c'est-à-dire le nombre de candidats termes plus complexes qui ont le candidat terme en position tête (resp. expansion). A partir de ces informations, l'analyste peut visualiser des listes paradigmatiques de candidats termes partageant la même tête ou la même expansion (cf. figure 1), ce qui le guide vers la constitution de taxinomies locales.

La difficulté essentielle pour l'utilisateur vient de la masse des résultats fournis par l'extraction. Même s'il existe de nombreux travaux fort intéressants sur le filtrage statistique de candidats termes extraits automatiquement de corpus, l'expérience montre qu'aucune mesure statistique ne peut suppléer l'expertise de l'analyste, en particulier parce qu'il y a toujours des candidats termes de fréquence 1 dont l'analyse est intéressante. De façon générale, sachant qu'il ne pourra analyser tous les candidats termes extraits du corpus, l'analyste doit adopter une stratégie optimale qui, étant donné le temps qu'il a choisi de consacrer à la tâche d'analyse textuelle et en fonction du type de la ressource à construire, lui garantit que, parmi les candidats qui auront échappé à son analyse, la proportion de ceux qui auraient pu être pertinents est faible.

5.2.2 Analyse distributionnelle : UPERY

Dans les trois projets, les résultats de l'outil UPERY ont été utilisés. UPERY (Bourigault, 2002) est outil d'analyse distributionnelle. Il exploite l'ensemble des données présentes dans le réseau de mots et syntagmes construits par SYNTEX pour effectuer un calcul des proximités distributionnelles entre ces unités. Ce calcul s'effectue sur la base des contextes syntaxiques partagés. Il s'agit d'une mise en œuvre du principe de l'analyse distributionnelle du linguiste américain Z. S. Harris, réalisée dans la lignée des travaux de H. Assadi (Assadi & Bourigault, 1996). L'analyse distributionnelle rapproche d'abord deux à deux des candidats termes qui partagent un grand nombre de contextes syntaxiques. Par exemple, dans le corpus REA, les candidats termes *insuffisance rénale* et *détresse respiratoire* sont rapprochés car on les trouve dans les contextes syntaxiques suivants : complément de *prise en charge*, de *apparition*, de *installation*, de *admettre en réanimation chirurgicale pour*.

Trois mesures permettent d'appréhender la proximité entre deux candidats termes. Le coefficient *a* est égal au nombre de contextes syntaxiques partagés par les deux termes. Cette mesure donne une première indication de la proximité entre deux termes. Mais cette mesure reflète de façon insatisfaisante la proximité. Il faut tenir compte de la productivité en Tête des contextes partagés : plus un contexte partagé par deux termes est productif, moins sa contribution au rapprochement des deux candidats termes doit être importante. Cette intuition est prise en compte par le coefficient *prox* qui pondère chaque contexte partagé par l'inverse de sa productivité. Enfin, pour évaluer la proximité entre deux unités, il est important de tenir compte non seulement de ce qu'elles partagent, mais aussi de ce qu'elles ont en propre. On caractérise la proximité entre deux candidats termes à l'aide de deux indices supplémentaires : pour chacun des deux candidats termes, le rapport entre le nombre de contextes partagés et le nombre total de contextes dans lesquels apparaît le candidat terme.

Le module d'analyse distributionnelle UPERY calcule chacun de ces coefficients pour chaque couple de candidats termes, et ne sont présentés à l'utilisateur que les couples dont les coefficients dépassent certains seuils. Ceux-ci sont définis de façon empirique et varient en fonction d'une part de l'homogénéité et de la redondance du corpus et d'autre part du contexte dans lequel doivent être exploités les résultats de l'analyse distributionnelle. L'analyse distributionnelle implémentée dans UPERY est symétrique : on calcule aussi la proximité entre contextes syntaxiques. Deux contextes syntaxiques sont proches si on y trouve les mêmes termes. Par exemple, dans le corpus REA, les verbes *montrer* et *mettre en évidence* sont proches car ils partagent en position sujet les termes *échographie*, *bilan infectieux*, *tomodensitométrie*, *artériographie*, *auscultation pulmonaire*, etc.

Il s'avère que les rapprochements effectués par UPERY sont extrêmement utiles et pertinents pour la construction de classes conceptuelles. Le nombre de rapprochements effectués dépend de la redondance du corpus. Par exemple, les corpus REA et le corpus du Code civil, l'un des corpus exploité dans le projet DROIT, sont deux corpus différents quant à ce paramètre de la redondance. Le corpus REA est constitué dans un ensemble de comptes rendus médicaux qui décrivent tous les mêmes types d'événement et donc dans lesquels les mêmes structures syntaxiques reviennent régulièrement. A l'opposé, dans le Code civil, les redondances, répétitions, reformulations sont évitées. Cela se répercute de façon assez sensible sur la richesse des résultats fournis par UPERY sur chacun des 2 corpus, puisque dans le corpus REA 30% des syntagmes nominaux et 47% des noms, de fréquence supérieure ou égale à 5, sont rapprochés d'au moins un autre mot, alors qu'ils ne sont que respectivement 20% et 43% pour le corpus du Code civil. Le phénomène est encore plus accentué dans le corpus LIVRE du projet VERRE. La taille du corpus est relativement réduite (100 000 mots, contre 400 000 pour le corpus REA et 150 000 pour le Code civil) et les redondances sont très faibles (chaque chapitre traite d'un sujet spécifique, et l'auteur s'efforce de varier son style). De ce fait, seuls 3% des SN et 18% des noms, de fréquence supérieure ou égale à 5, ont des voisins,.

5.2.3 Extraction des relations : YAKWA et CAMELEON

Développé à l'ERSS par L. Tanguy, YAKWA est un concordancier pour corpus étiquetés (Rebeyrolle et Tanguy, 2000). Il permet de rechercher des phrases et/ou des paragraphes contenant une séquence définie par des marqueurs. Ce marqueurs s'appuient sur les informations notées par l'étiqueteur dans les corpus, comme les catégories grammaticales des mots. Leur contenu peut être formé de formes lexicales (tronquées, exactes, etc.), de formes canoniques des unités lexicales du texte, de catégories morpho-syntaxiques et de leur combinaisons (disjonctions, conjonction de marqueur lexical et de marqueur morpho-syntaxique), de la négation d'un des types de marqueurs précédents ou de jokers (mots non comptabilisés). YAKWA peut s'adapter à tout type d'étiqueteur, par exemple CORDIAL université pour le français ou TREETAGGER pour l'anglais. Son interface guide la construction de marqueurs et permet d'en visualiser la projection sur un corpus.

CAMELEON est un logiciel de recherche de relations lexicales à partir de marqueurs linguistiques (Séguéla, 1999). Il est associé à un module de modélisation qui permet de valider (ou de rejeter) ces relations lexicales pour les intégrer sous forme de relations sémantiques dans un modèle conceptuel. Les marqueurs utilisés dans CAMELEON peuvent être des marqueurs génériques prédéfinis ou leur adaptation ou encore des marqueurs spécifiques définis par l'utilisateur. L'idée est de rechercher des relations avec des moyens adaptés au corpus étudié. Les relations sont donc génériques (comme EST-UN) ou spécifiques au corpus (comme « used-in » dans le projet VERRE), et les marqueurs associés à toutes les relations sont revus et adaptés à chaque corpus. Le langage d'expression des marqueurs est moins riche que celui de YAKWA car CAMELEON fonctionne sur un corpus brut non étiqueté. En revanche, Caméléon présente deux points forts pour la construction de RTO : il propose une base générique de relations et de marqueurs associés ; il s'appuie sur les résultats de SYNTEX pour suggérer les concepts qui pourraient être en relation à partir de la forme lexicale trouvée.

6 Le problème de l'évaluation

Nous terminerons par quelques réflexions sur le problème de l'évaluation. Il faut distinguer l'évaluation d'une RTO particulière construite dans un contexte particulier, de l'évaluation de tel outil ou tel outil de TAL d'aide à la construction de RTO. Dans les deux cas, il faut adopter une approche ingénierique, en adoptant 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 RTO ou de l'outil.

En ce qui concerne les RTO, il faut distinguer *validation* et *évaluation*. Dans le processus de construction d'une RTO, il y a plusieurs moment de *validation* de la RTO, 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. 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. Le problème ne se pose pas tant en terme de vérité, qu'en terme de non violation des connaissances de l'expert. 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 ne reconnaît pas nécessairement à première vue ses petits, 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, et parce que chaque cas étant particulier il sera de toutes façons difficile de définir des procédures à la fois précises et relativement génériques, et que cela dépasse quelque peu le cadre de la recherche.

L'évaluation des outils de construction de RTO est le problème qui nous concerne ici. C'est un problème lui aussi difficile. La source des difficulté est double : d'abord il s'agit d'outils d'aide, ensuite chaque outil est rarement utilisé seul. Quand il s'agit d'évaluer d'un outil automatique, du type « boite noire », il est possible d'évaluer les performances de l'outil 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 : une modification, un enrichissement de la ressource à un ou plusieurs points du réseau, voire dans certain cas l'absence d'action immédiate, sans que cela signifie nécessairement que les résultats en question soient faux ni même on pertinents. De plus, chaque cette interprétation s'appuie normalement sur une confirmation par retour aux textes. Il n'y a pas systématiquement de trace directe entre un résultats (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 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 entre 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.

Référence

Ait El Mekki T., Nazarenko A (2002), Comment aider un auteur à construire l'index d'un ouvrage ?*,* Actes du *Colloque International sur la Fouille de Texte CIFT'2002*, Y. Toussaint et C. Nedellec Eds., oct. 2002, pp. 141-158

Assadi H. (1998), *Construction d'ontologies à partir de textes techniques. Application aux systèmes documentaires*, thèse de l'Université Paris 6

Aussenac-Gilles N. (1999), GEDITERM, un logiciel de gestion de bases de connaissances terminologiques*, in* Actes des Journées Terminologie et Intelligence Artificielle (TIA'99), Nantes, *Terminologies Nouvelles* n°19, 111-123.

Aussenac N., Séguéla P. (2000), Les relations sémantiques : du linguistique au formel. *Cahiers de grammaire*, N° spécial sur la linguistique de corpus. A. Condamines (Ed.) Vol 25. Déc. 2000. Toulouse : Presse de l'UTM. Pp 175-198.

Aussenac-Gilles N., Biébow B., Szulman N. (2000), Revisiting Ontology Design: a method based on corpus analysis. *Knowledge engineering and knowledge management: methods, models and tools, Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management*. Juan-Les-Pins (F). Oct 2000. R Dieng and O. Corby (Eds). Lecture Notes in Artificial Intelligence Vol 1937. Berlin: Springer Verlag. pp. 172-188.

Bachimont, B. (2000), Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances ». In J. Charlet *et al*. (eds), *Ingénierie des Connaissances ; Evolutions récentes et nouveaux défis*, Eyrolles, pp. 305-323

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., 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, in J.-M. Pierrel (éd.), *Industrie des langues*, Hermès, Paris, pp. 215-233

Charlet J., Zacklad M., Kassel G. & Bourigault D. (eds) (2000), *Ingénierie des connaissances : évolutions récentes et nouveaux défis,* Eyrolles : Paris - Collection technique et scientifique des télécommunications

Charlet J. (2002), *L'ingénierie des connaissances : résultats, développements et perspectives pour la gestion des connaissances médicales*. Mémoire d'habilitation à diriger des recherches, université Pierre et Marie Curie

Chaumier J. (1988), *Travail et méthodes du/de la documentaliste : Connaissance du problème, Applications pratiques.* 3e éd. mise à jour et complétée. Paris : ESF, 1988

Condamines A. et Rebeyrolles J (2000), Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In Charlet J, Zacklad M., Kassel G. & Bourigault D. éds. *Ingénierie des connaissances. Tendances actuelles et nouveaux défis*. Editions Eyrolles/France Telecom, Paris

Daille B. (1994), *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse en Informatique Fondamentale, Université de Paris 7, Paris

David S. et Plante P. (1990), De la nécessité d'une approche morpho-syntaxique dans l'analyse de textes. *Intelligence Artificielle et Sciences Cognitives au Québec*, 3(3):140-154

Enguehard C. et Pantera L. (1995), Automatic natural acquisition of a terminology. *Journal of Quantitative Linguistics*, 2(1):27-32

Faure D. (2000), *Conception de méthode d'apprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*, thèse de Doctorat Université de Paris Sud

Garcia D. (1998), *Analyse automatique de textes pour l'organisation causale des actions. Réalisation du système informatique COATIS.* Thèse en informatique. Université Paris IV

Grefenstette G. (1994), *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publisher, Boston, MA

Habert B., Naulleau E. et Nazarenko A . (1996), Symbolic word clustering for medium-size corpora. *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, Copenhagen, pp 490-495

Jacquemin C. (1997), *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes

Lame G. (2002), *Construction d'ontologie à partir de textes. Une ontologie du Droit français dédiée à la recherche d'information sur le Web*, thèse de l'Ecole des Mines de Paris

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.

Maynard D. et Ananiadou S. (2001), Term extraction using a similarity-based approach, in Bourigault D., Jacquemin C. & L'Homme M.-C., *Recent advances in computational terminology*, John Benjamins Publishing, Amsterdam, pp 261-278

Morin E. (1999), Des patrons lexico-syntaxiques pour aider au dépouillement terminologiques, *Traitement Automatique des Langues*, volume 40, Numéro 1, pp. 143-166

Nazarenko A., Zweigenbaum P., Habert B. & Bouaud J. (2001), Corpus-based extension of a terminological semantic lexicon, in Bourigault D., Jacquemin C. & L'Homme M.-C., *Recent advances in computational terminology*, John Benjamins Publishing, Amsterdam, pp 327-352

Rastier F. (1991), Sémantique et recherches cognitives, Presses Universitaires de France, Paris, 1991

Rastier F. (1995), Le terme : entre ontologie et linguistique, Actes *des 1ères Journées " Terminologie et Intelligence Artificielle"*, Villetaneuse, avril 1995*, La banque des mots*, Numéro spécial 7-1995, pp. 35-65

Rousselot F., Frath P. et Oueslati R. (1996), Extracting concepts and relations from corpora, *Procedings of the 12th European Conference on Artificial Intelligence* (ECAI'96), workshop on Corpus-Oriented Semantic Analysis, Budapest

Séguéla P. et Aussenac-Gilles N. (1999), Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine, *Actes de la conférence Ingénierie des Connaissances (IC'99)*, Paris, pp 79-88

Slodzian M. (2000), L'émergence d'une terminologie textuelle et le retour du sens, in *Le sens en terminologie*, publication du Centre de Recherche en Terminologie et Traduction de l'Université Lyon 2

Sparck Jones K. (1971), *Automatic Keyword Classification for Information Retrieval*. Butterworth, London

Szulman S., Biébow B. & Aussenac-Gilles N. (2002), Structuration de Terminologies à l'aide d'outils d'analyse de textes avec TERMINAE, *Traitement Automatique de la Langue (TAL)*. Numéro spécial sur le Structuration de Terminologie. Eds A. Nazarenko, T. Hammon. Vol43, N°1; pp 103-128. 2002.

Toussaint Y., Namer F., Daille B., Jacquemin C., Royauté J. et Hathout N. (1998), Une approche linguistique et statistique pour l'analyse de l'information en corpus. Actes *de la 5ème conférence annuelle sur le Traitement Automatiques des Langues Naturelles (TALN'98)*, Paris, pp. 182-191

Velardi P., Missikoff M. & Basili R. (2001) Identification of relevant terms to support the construction of domain ontologies. *In ACL WS on Human Language Technologies and Knowledge Management*. Toulouse (F), July 6-7, 2001. 18-28.

Zweigenbaum P. (1999) Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé* 1999(23)