Corpus et terminologie

Résumé

Les liens entre ressources termino-ontologiques et corpus intéressent la linguistique de corpus, la terminologie, l'informatique et les sciences de l'information. A partir de l'identification des difficultés théoriques et techniques communes, se dégage la nécessité d'approfondir la notion de genre, de mieux prendre en compte les questions d'usage et de maintenance ainsi que les modalités d'évaluation des ressources et des outils de construction.

***Mots-clés** : ontologie, terminologie, thesaurus, corpus, ingénierie des langues et des connaissances, sciences de l'information*

Abstract

The link connecting terminological and ontological resources with corpora may concern corpus linguistics, terminology, knowledge engineering or information sciences. Similar theoretical and technical issues underline the need to specify the notion of genre, to better integrate use and maintenance issues and to define evaluation protocols for the resources and their building tools.

***Keywords** : ontology, terminology, thesaurus, corpus, language and knowledge engineering, information sciences.*

1. INTRODUCTION

La thématique « Corpus et Terminologie » constitue un thème de recherche très interdisciplinaire qui permet aussi d'établir des ponts entre recherche et besoins en entreprise. En effet, la nécessité de représenter le contenu d'un texte sous une forme qui soit accessible, manuellement ou automatiquement (en l'occurrence, sous la forme de ressources terminologiques ou ontologiques - RTO par la suite -) rassemble des chercheurs de disciplines diverses et des professionnels pour qui la gestion de la documentation est un problème crucial. L'accès aux textes par des formes lexicales semble assez facile à réaliser d'un point de vue technique, mais beaucoup moins d'un point de vue sémantique. Il y a quelques années encore, seules des collaborations entre deux ou trois disciplines existaient (par exemple, pour la construction d'outils de

repérage de termes). Récemment, avec la montée en puissance d'internet, les besoins sont devenus plus pressants encore, ce qui a rendu nécessaire d'identifier les apports et les complémentarités des différentes disciplines concernées. Une réflexion s'est mise ainsi en place, en particulier au sein de l'Action Spécifique « Corpus et Terminologie »[1], (2002-2004), qui a impliqué des chercheurs des disciplines suivantes : Linguistique de corpus, Terminologie, Sciences de l'information, Traitement automatique des langues (TAL), Ingénierie des connaissances (IC), Recherche d'information (RI).

Ce chapitre rend compte des travaux menés dans l'Action Spécifique (Aussenac-Gilles, Condamines, 2003). Notre objectif consiste à montrer comment, à partir d'histoires et d'objectifs différents, ces disciplines ont pu parvenir à identifier des problématiques proches (méthodes, besoins) et à définir des perspectives communes. Le plan du chapitre suit précisément ce schéma de réflexion : la partie 2 présente les points de vue des différentes disciplines sur les RTO ; la partie 3 décrit les méthodes et les outils qu'elles proposent. Enfin, la partie 4 présente des perspectives communes identifiées pour développer la problématique de la construction de RTO à partir de corpus.

2. UNE PROBLEMATIQUE TRANSDISCIPLINAIRE MAIS DES HISTOIRES ET DES BESOINS DIFFERENTS

La représentation des connaissances sous la forme de listes de termes reliés par des relations (à la sémantique plus ou moins définie) est courante et ancienne. Des taxinomies utilisées en sciences naturelles aux XVIIe et XVIIIe siècles en passant par la classification universelle de Dewey (1876), les réseaux sémantiques de Quillian (1968) ou les ontologies de l'ingénierie des connaissances (Gruber, 1993), autant de modes de représentation qui mettent l'accent sur l'utilisation d'éléments lexicaux pour modéliser la connaissance. Ces représentations sont utilisées dans des systèmes informatiques où elles sont décrites à l'aide de langages comme les graphes conceptuels ou les logiques terminologiques. Toutefois, les possibilités de l'informatique ne doivent pas occulter plusieurs limites de ce mode de représentation, en particulier si l'on examine ses capacités à la fois à être construit à partir de textes et à

[1] http://www.irit.fr/ASSTICCOT ou http://rtp-doc.enssib.fr/article.php3?id_article=40

permettre l'accès à des textes (rarement ceux qui ont servi à l'élaborer). Ainsi, la représentation sous forme relationnelle introduit parfois une part d'implicite. Par ailleurs, la structuration d'un réseau conceptuel à partir de termes relève nécessairement d'une interprétation, c'est-à-dire d'une normalisation (Bachimont, 2000) (Rastier, 1995).

L'histoire et les besoins ciblés par chaque discipline permettent de comprendre son point de vue sur les corpus et les ressources termino-ontologiques.

2.1. Ressources termino-ontologiques

2.1.1. Terminologie textuelle : les bases de connaissances terminologiques

L'histoire de la terminologie prend sa source dans les travaux des taxinomistes des XVIIè et XVIIIè siècles qui visaient tout à la fois à nommer et à contrôler les éléments naturels. Foucault (1966) montre que dans les siècles qui ont suivi, et jusqu'à nos jours, une vision de la langue scientifique fonctionnant sur un mode particulier, différent de la langue « générale », en tout cas moins ambiguë et polysémique, s'est développée. On retrouve cette façon de voir dans les travaux de Wüster, ingénieur viennois qui, dans les années 1930, a développé la théorie générale de la terminologie (Cabré, 1998). Cette vision de la terminologie, par essence normalisatrice, a pour fondement l'idée que la langue dans des domaines spécialisés peut être un moyen de communication parfait ou en tout cas perfectible, d'où la nécessité de normaliser pour éviter les créations individuelles favorisant de mauvaises compréhensions. On comprend l'idéologie qui sous-tend cette option : améliorer les échanges entre industriels dans une même langue ou, mieux, dans différentes langues. Pour ce faire, la discipline terminologique est maintenue à l'abri des connaissances linguistiques et la construction de terminologies à l'écart de la réalité des discours des milieux professionnels (Wüster, 1981). Dans cette approche, la construction de terminologies se fait avant tout par interrogation des experts.

Le développement de l'informatique a provoqué une accélération dans la réflexion sur la terminologie. En effet, grâce à l'informatisation des textes et au développement d'outils pour les interroger, l'impasse sur la consultation de textes produits dans des situations réelles n'était plus tenable. Dans le même temps, la demande de la part des entreprises est devenue très prégnante. Ainsi, au début des années 1990, informaticiens et

linguistes ont collaboré pour interroger les modes de prise en compte des textes dans la construction de terminologies. Cette prise en compte a rapproché la terminologie de la linguistique (linguistique de corpus et analyse de discours), au bénéfice de la recherche dans ces disciplines. Dans le même temps, le mythe d'une langue scientifique et technique pure, ou en tout cas pouvant être purifiée, a été quelque peu écorné.

2.1.2. Sciences de l'information : les thesaurus

À la différence de la terminologie, les langages documentaires et les thesaurus revendiquent d'emblée le rapport aux textes, et plus largement aux documents, comme manifestant la connaissance. Ainsi, si la réflexion en terminologie est plutôt guidée, en tout cas à l'origine, par les besoins de la traduction, en documentation, elle est guidée par la nécessité de créer un lien entre documents et connaissance, et, de fait entre documents. Le thesaurus joue ainsi un rôle majeur à deux moments clés de l'interaction avec un document : lors de son indexation et lors de sa recherche. Dans la grande majorité des cas, le thesaurus est constitué sur des bases introspectives. C'est le cas de la Classification Décimale Universelle, censée organiser le savoir universel, et qui est très utilisée dans les bibliothèques.

L'informatisation des textes et, particulièrement, le développement d'internet ont eu, pour les sciences de l'information, un effet majeur. En effet, là où le travail d'indexation des documentalistes était jugé comme une aide pour l'accéder à des documents existant seulement sous forme matérielle, ce même travail peut être perçu comme un parasitage lorsqu'il vient s'intercaler entre l'usager et les corpus informatisés. Par ailleurs, l'existence d'internet a introduit la notion de commerce dans un domaine qui était plutôt considéré comme un service et un travail intellectuel. Des moteurs de recherche comme Google ont une visée lucrative qui vient dénaturer l'accès à l'information (Pédauque, 2005).

Confrontés, dès les années 1960, au problème de l'informatisation des textes, les documentalistes se sont trouvés impliqués (Chaumier, 2002) dans le traitement automatique des langues et de la représentation informatique des connaissances. Après une période de doutes sur leur rôle dans l'accès à l'information de textes automatisés, la réflexion semble maintenant s'apaiser. L'expérience de l'utilisation réelle d'internet a montré les limites d'un accès médiatisé par un intermédiaire commercial. Le rapprochement avec des disciplines proches, comme la terminologie textuelle, a permis de mieux situer l'automatisation des textes et, en retour, de proposer à ces disciplines sœurs de prendre en compte la notion

d'usage (Lainé-Cruzel, 2001). Les sciences de l'information ont identifié les mêmes tensions que la terminologie. Il y a nécessité à normaliser pour favoriser les échanges, dans une langue et entre les langues ; mais, normaliser, c'est imposer une vision du monde. Il faut donc trouver un point d'équilibre entre ces deux éléments. Ces disciplines ont aussi en commun le mode de représentation : des concepts reliés par des relations qui élaborent un système.

2.1.3. Ingénierie des connaissances : les ontologies

En ingénierie des connaissances, les ontologies ont été définies afin de mieux réutiliser les connaissances du domaine, de les gérer séparément des raisonnements (van Eijst, 1995) et, surtout, de faciliter l'échange de connaissances (Neches *et al*., 1991) et l'interopérabilité des systèmes les utilisant. Pour communiquer, ces systèmes requièrent des représentations du monde compatibles et cohérentes, traduisibles dans un même langage standard et simple (Gruber, 1991). Les ontologies reflètent ainsi la recherche d'invariants dans un domaine, d'une description générique des connaissances, au minimum consensuelle. Définie comme *« spécification normalisée représentant les classes des objets reconnus comme existant dans un domaine »* (Charlet, 2003), l'ontologie s'intéresse aux concepts en tant que futures primitives logiques, à leur définition via des relations sémantiques, mais aussi à leur pertinence pour la restitution de résultats aux utilisateurs. Ainsi, les concepts renvoient aux connaissances des individus, exprimées à travers le langage, et doivent être définis en tenant compte des termes du domaine et de leur sémantique.

S'inspirant d'abord du génie logiciel, les méthodes de construction d'ontologies n'ont problématisé la définition et l'organisation des concepts qu'après 1996 (Uschold *et al*., 1996) (Fernandez-Lopez et *al*., 1999). L'utilisation de textes comme sources de connaissances a pris son essor autour de 2000. Les potentiels de cette approche sont multiples : une automatisation partielle, grâce au traitement automatique des langues puis à l'apprentissage automatique ; une réduction du coût ; enfin, le renouvellement des hypothèses sur le statut des concepts et leurs liens avec les termes. L'étude de l'usage des termes dans les textes invite à abandonner une vue normative *a priori* pour prendre en compte usages et points de vue afin de normaliser ces concepts puis les formaliser en fonction d'un objectif particulier. L'utilisation des corpus se veut ici un moyen d'accéder aux connaissances d'un domaine en complément ou à la place de l'expertise humaine.

Le développement massif d'applications pour le web a ajouté de nouveaux enjeux, techniques et économiques, ayant des conséquences sur la forme mais aussi le fond des modèles attendus. Les architectures ou applications du futur web dit « web sémantique » font systématiquement appel aux ontologies, dont on attend sans doute trop : elles doivent fournir des représentations partagées par des agents logiciels, des méta-données pour annoter ou indexer des documents ou encore assurer la mise à disposition de connaissances consensuelles. Les ontologies y couvrent des réalités différentes alors que les liens avec les textes deviennent centraux.

2.1.4. TAL comme producteur et utilisateur de RTO

Le TAL est particulièrement concerné par l'étude de RTO en lien avec les corpus et ce, de deux manières complémentaires, en tant que producteur et consommateur de RTO. D'une part, le TAL peut produire des RTO, en général après une chaîne de traitements complexes faisant intervenir plusieurs niveaux d'analyse et donc plusieurs logiciels (Bourigault *et al*., 2004) (Habert et Zweigenbaum, 2002). Ainsi, de nombreux outils sont développés pour aider à extraire des terminologies à partir de corpus électroniques. D'autre part, les logiciels de TAL sont aussi consommateurs de ressources terminologiques, qui leur permettent d'obtenir de meilleurs résultats. Leur nature dépend des tâches réalisées par ces outils : ressources morphologiques, grammaticales, sémantiques, multilingues ou non , etc.

Cette situation est à mettre en regard avec la diversité des applications couvertes autant que des ressources produites ou utilisées par le TAL (Nazarenko, 2005). Les travaux habituels considèrent ces deux facettes (utilisation ou production) séparément. Il est primordial de les envisager conjointement dans le cas de la construction de RTO à partir de textes à cause de la complexité de ces ressources. Cette complexité entraîne plusieurs exigences : la possibilité d'enchaîner plusieurs traitements élémentaires ou non, la nécessité de superviser le tout, afin de produire des validations à chaque étape mais aussi d'orienter l'interprétation des résultats en fonction des objectifs de modélisation. Ce processus devient cyclique dès lors que les mêmes ressources produites par traitement automatique de corpus viennent alimenter une application relative à ces corpus (i.e. une application visant à les indexer, à y retrouver des informations, etc.).

La place des corpus dépasse celle d'entrée du processus d'analyse. Le plus souvent, ils co-existent avec les résultats pour les enrichir. Un logiciel de TAL n'est pas un « révélateur » de toute la sémantique des

textes, mais bien le moyen d'automatiser des recherches ciblées qui contribuent à reconstruire une sémantique (Nazarenko, 2005). Le corpus est à la fois l'objet sur lequel portent les traitements, la justification de leur pertinence et la source d'information qui contribue à mieux les interpréter et les exploiter.

3. DES OBJECTIFS ET DES METHODES PROCHES

Des histoires différentes ont pourtant conduit à un besoin commun : construire des ressources terminologiques ou ontologiques (RTO) à partir de textes. Malgré les limites de ce mode de représentation de la connaissance exposées ci-dessus, ce choix a une pertinence d'une part dans le cas des textes spécialisés, rendant compte en principe d'une connaissance plus consensuelle ; d'autre part, parce qu'il est intéressant de bénéficier de la puissance de calcul de l'informatique. Il s'agit donc de s'interroger sur les liens possibles entre des discours et des éléments lexicaux, étant entendu que l'on souhaite pouvoir utiliser les seconds pour accéder au contenu des premiers et que l'on souhaite rendre cette construction aussi systématique (voire automatique) que possible. En d'autres termes, la question qui a sous-tendu la réflexion de l'Action Spécifique est celle de savoir comment contrôler l'interprétation d'un texte (ou d'un corpus), à l'aide de différents outils, pour construire une RTO utilisable par des humains et/ou des machines, qui permette, en retour, d'accéder à des textes existants. La réflexion commune a souligné la nécessité d'une approche pluridisciplinaire pour mettre au point ces outils et les modalités d'interprétation de leurs résultats.

3.1. Logiciels pour identifier des éléments de RTO

L'analyse informatique de textes pour accéder au sens des mots qu'ils contiennent est en plein essor avec l'augmentation du nombre de documents électroniques disponibles. Ces analyses sont au cœur des procédés automatisés de construction de RTO. Elles s'intéressent avant tout aux éléments lexicaux et à leurs relations, à leur repérage exhaustif et à l'identification de leur sens plus qu'à leur représentation logique. Leur visée première relève de la désambiguïsation sémantique, ou de l'acquisition sémantique par la mise en évidence de relations ou de classes sémantiques afin de caractériser le sens.

Alors que la demande actuelle pousse à la définition d'outils qui donneraient accès directement à des concepts en vue de produire des ontologies, un regard interdisciplinaire oblige à nuancer cette hypothèse parfois naïve (Habert *et al.,* 2005). Ces logiciels d'analyse nécessitent une réflexion interdisciplinaire sur le rôle de supervision et d'interprétation des données tenu par l'utilisateur.

Acquisition de termes : Ces logiciels facilitent l'extraction, à partir de corpus analysé, de *candidats termes*. Sont candidats des mots ou groupes de mots susceptibles d'être retenus comme termes par un analyste, et de fournir des étiquettes de concepts. Pour donner des éléments de décision, ces outils s'appuient sur une ou plusieurs techniques (morpho-syntaxique, statistique, autres). Le recours à des ressources externes, comme des bases de données lexicales ou des textes étiquetés manuellement, est parfois envisagé par le TAL et l'IC, mais seulement d'un point de vue technique dans la perspective d'améliorer les résultats. Or la linguistique et les sciences de l'information soulignent l'importance de l'adéquation sémantique de ces ressources.

Structuration de termes et regroupement conceptuel : Dans cette classe, se situent des outils de repérage et classification de termes et des outils de repérage de relations, selon des approches statistiques ou lexico-syntaxiques. En effet, il s'agit là d'étapes vers l'*identification de concepts* et le *repérage de relations sémantiques* assurant leur mise en relation. Ces techniques, même lorsqu'elles incluent des analyses sémantiques, consistent à manipuler des formes de la langue pour en dégager du sens. Elles s'avèrent indispensables dans le cas où de gros volumes de documents ou de connaissances doivent être manipulés (comme en génétique). Mais la qualité et la couverture des résultats produits augmentent significativement dès que des professionnels de ce type d'analyse interviennent pour les retoucher.

Une des problématiques inter-disciplinaires est ici de définir ces logiciels comme des supports à une analyse outillée et supervisée par leurs utilisateurs, de préciser le rôle des utilisateurs et la place des logiciels dans un processus global de modélisation.

Depuis peu, une dichotomie oppose des logiciels de *construction d'ontologies* (*ontology design*), destinés à définir concepts et relations, aux logiciels *d'enrichissement d'ontologie* (*ontology population*) qui les complètent à l'aide de nouvelles instances de concepts. Grâce à des techniques d'apprentissage et d'extraction d'information, les contextes linguistiques de concepts peuvent être caractérisés. Ainsi, ces outils facilitent aussi *l'annotation* de documents à l'aide de méta-données et de

connaissances. Il s'agit là d'une nouvelle manière d'aborder l'annotation, dont la définition interroge les pratiques des disciplines qui maîtrisaient ces tâches jusque là (sciences de l'information et de la communication, recherche d'information, etc.).

3.2. Nécessité d'une interprétation des données

La double contrainte imposée par la constitution de RTO (normaliser tout en restant proche de l'usage) oblige à s'interroger sur les modes d'interprétation mis en œuvre. L'importance des relations dans les RTO fait que l'accent est mis sur la recherche d'éléments textuels pouvant être représentés sous cette forme relationnelle. Concrètement, deux approches très complémentaires sont possibles : l'approche par marqueurs et l'approche par interprétation des contextes distributionnels.

Initialement, les marqueurs de relations conceptuelles étaient formés d'éléments syntaxiques ou lexicaux auxquels on pensait pouvoir associer spontanément (hors contexte) une interprétation relationnelle. Ces marqueurs ont été étudiés depuis les années 1980 par des sémanticiens (Lyons, 1978), (Cruse, 1986) sur des bases introspectives. Les éléments considérés comme marqueurs devaient donc être interprétés *a priori* et pouvoir être utilisés de manière assez systématique par des outils automatiques. De nombreuses études ont montré que les contextes dans lesquels apparaissent ces marqueurs doivent être contrôlés pour limiter les phénomènes de polysémie, cette notion de contextes devant être entendue aussi bien comme co-texte (contexte linguistique) que contexte situationnel (contexte extra-linguistique) (Pearson, 1998), (Condamines, 2002), (Marshman *et al.,* 2002),(L'homme, 2004).

L'approche distributionnelle ne fait pas intervenir une interprétation *a priori.* Elle s'intéresse moins aux marqueurs de relations qu'aux termes et aux contextes dans lesquels apparaissent ces termes. L'idée à la base de l'approche distributionnelle est que la distribution d'un mot (l'ensemble des contextes dans lesquels il apparaît) détermine son sens. Or, la constitution de RTO revient à rendre compte du sens sous une forme relationnelle. Il y a donc une pertinence à essayer de retrouver des relations entre les termes en tirant parti de la récurrence des contextes dans lesquels ils apparaissent. Le distributionnalisme « à la Harris » a considéré que la prise en compte systématique des distributions d'un mot suffisait à faire émerger son sens. Des travaux plus récents ont reconnu que l'analyse de ces distributions ne donnait pas directement le sens mais supposait une interprétation (Habert et Zweigenbaum, 2002). De fait, le

choix de représenter le sens sous une forme relationnelle nécessite une interprétation qui doit être guidée à la fois par la prise en compte des contextes d'apparition et par l'objectif de la construction.

En réalité, les deux approches (interprétation *a priori* de marqueurs et interprétation *a posteriori*) sont complémentaires. En effet, l'interprétation se fait toujours par un ajustement entre connaissances langagières *a priori* et données du corpus et ce, jusqu'à ce que l'interprétation se stabilise. De fait, la notion de marqueur s'est affinée et il semble plus adéquat de parler de faisceaux d'indices de différentes natures permettant de construire des relations.

4. DES PERSPECTIVES COMMUNES

La définition d'une problématique interdisciplinaire a permis de repérer les convergences et de travailler à l'identification de perspectives de réflexion communes, qui tiennent en quatre points.

4.1. Développer et approfondir la notion de genre

La prise en compte de la notion de genre permet de constituer des catégories de textes censés avoir les mêmes caractéristiques extra-linguistiques et les mêmes régularités linguistiques. Un des objectifs qui apparaît consiste à affiner les descriptions pour les rendre mieux adaptées aux différents besoins d'analyse, en particulier à ceux du traitement des textes par l'informatique. Fondamentalement, la principale difficulté consiste à identifier les bonnes caractérisations, qui reflètent des points de vue complémentaires ou parfois contradictoires :

- *Prise en compte du contexte extra-linguistique :*
  - La production : caractéristiques des rédacteurs, des interlocuteurs visés, date, langue, niveau de compétences, objectifs …
  - L'utilisation : modes d'utilisation possibles, en fonction d'une demande particulière ; typiquement, regrouper des textes pour un usager d'un centre documentaire.
  - La diffusion : format ou support sous lequel le texte est disponible.
- *Prise en compte des régularités linguistiques* : Sur ce point, s'est développée l'analyse de corpus à l'anglo-saxonne (Biber, 1988) qui cherche à rapprocher des textes sur la base de fonctionnements linguistiques similaires, et ce de façon automatique. La limite est que

seuls sont repérés les fonctionnements identifiés par une forme : présence/absence d'éléments rendant compte d'un sens.

Trouver les modes de caractérisation des textes pour mieux anticiper les traitements efficaces sur ces textes sera, à l'évidence, l'un des défis à relever pour l'analyse de corpus en général. Pour la constitution de RTO à partir de corpus, les points de vue présentés ci-dessus sont pertinents et peuvent être déclinés. Le contexte de production des textes choisis doit être adapté au degré de spécialisation souhaité de la RTO. Si la RTO n'est pas construite par un spécialiste mais par un ingénieur de la connaissance ou un terminologue, le corpus devra être aussi compatible avec sa compétence, et complété de connaissances plus élémentaires. L'utilisation sera prise en compte via la définition de types de RTO : on n'utilisera pas les mêmes textes pour construire un thesaurus ou un modèle de connaissance pour un outil de raisonnement. Enfin, les caractérisations linguistiques peuvent guider dans le choix des outils et des techniques d'analyse. La description des marqueurs de relation peut ainsi être affinée en prenant en compte la notion de genre textuel (Condamines, 2002).

Une fois encore, les modes de prise en compte du TAL doivent être interrogés (Bourigault *et al.*, 2004). Les outils peuvent être utiles pour mettre au jour des régularités et des conjonctions de régularités, ce qui permet d'effectuer des rapprochements entre textes, parfois éloignés *a priori.* Dans un second temps, une fois les éléments de caractérisation identifiés, les outils peuvent permettre de repérer à quelle(s) catégorie(s) appartient un texte. Dans tous les cas, la prise en compte du genre, c'est-à-dire la compréhension de l'imbrication entre extra-linguistique et linguistique, relève déjà d'une interprétation.

4.2. Prendre en compte applications et usages

L'interprétation que nécessite la constitution d'une RTO à partir d'un corpus est influencée par la variation qui se situe à deux niveaux : dans les textes eux-mêmes (les usages langagiers) et dans les objectifs (les applications ou usages réels) visés par la RTO. La question est alors de savoir si cette variation est si présente qu'elle ne peut faire l'objet d'aucune systématisation, ou bien si on peut espérer la circonscrire afin de contrôler les modes d'interprétation et d'adapter les outils pour permettre cette interprétation. La variation conditionne également les capacités d'interprétation de la ressource elle-même, et donc sa maintenance.

La variation dans les usages langagiers n'est que le reflet de la difficulté qu'il peut y avoir à utiliser des textes, avec la diversité des

usages qu'ils peuvent manifester pour proposer une représentation de la connaissance sous forme relationnelle, qui soit relativement stable. La réflexion sur le genre, qui concerne à la fois les caractérisations linguistiques et extra-linguistiques, prend en compte la relation langue / connaissance. Elle devrait donc permettre de mieux comprendre s'il est possible de contrôler la variation dans les usages langagiers.

Le problème de la variation d'une application à l'autre semble *a priori* moins difficile à gérer. Les réflexions interdisciplinaires nous ont amenés à identifier un ensemble d'applications qui, sans être clos, recense les applications les plus courantes (constitution d'index, de thesaurus, de terminologies pour favoriser la communication, d'ontologies pour le TAL ou la RI …) et permet de repérer les éventuelles convergences. La prise en compte de l'application peut avoir un impact sur les choix méthodologiques et logiciels, au moins sur les quatre éléments suivants :

- Le profil de l'analyste. La constitution d'une RTO fait appel à des compétences multiples (analyse de textes, modélisation, informatique), chacune étant plus ou moins importante selon le type de ressource.
- La construction du corpus. Elle suivra des règles différentes en fonction de la nature de la RTO.
- Les outils de TAL. Deux types de besoins conduisent à des choix différents. Les besoins sophistiqués et ciblés, exigent de combiner des logiciels effectuant des traitements élémentaires. D'autres classes de besoins mieux identifiés permettent de développer des outils dédiés.
- Les outils de modélisation. Le degré de formalisation, la nécessité ou non d'une composante terminologique, de capacités inférentielles, etc. sont à déterminer en fonction de l'application ciblée, et orientent vers des plates-formes de modélisation particulières.

La prise en compte du rôle de l'application dans la construction de RTO est encore très récente et constituera un des enjeux du développement de la réflexion à venir.

4.3. Anticiper la maintenance des RTO

Les RTO ont vocation à rendre compte de connaissances considérées comme suffisamment consensuelles et stables pour pouvoir être partagées par différentes applications et utilisateurs, ou pour être mises à disposition au sein de communautés humaines. Or les conditions même de leur utilisation sont mouvantes et, avec elles, le vocabulaire et les conceptualisations en jeu. Paradoxalement, prévues pour figer des connaissances et des usages, les RTO doivent donner accès à des

connaissances qui évoluent dans des contextes dynamiques. Supposées fournir une norme ou une référence stable, elles peuvent être remises en question faute de répondre aux besoins, et devenir rapidement obsolètes. Les enjeux sont de taille. Ainsi, des données lexicales datées décrédibilisent la recherche d'informations. Au contraire, une indexation libre, par des termes fournis par les utilisateurs, développe une inflation de termes qui complique les futures recherches. Enfin, on n'est plus sûr de retrouver les bons documents avec d'anciens descripteurs. Sans cesse, les sciences de l'information sont confrontées à ces évolutions touchant les thésaurus.

On peut espérer qu'une maintenance régulière des RTO permette de conserver leur validité et leur pertinence. Mais se posent alors les questions classiques de la maintenance : Comment repérer de nouveaux besoins qui justifient des mises à jour ? Sur quoi les mises à jour doivent-elles porter ? Comment préserver la cohérence des éléments déjà consignés ? Comment s'assurer que les différentes utilisations de la ressource demeurent pertinentes ?

Aujourd'hui, ces problèmes sont mal maîtrisés, rarement abordés ou, lorsqu'ils le sont, traités d'un point de vue disciplinaire. Or, étroitement liée au mode de conception (à partir de textes dans ce cas), la maintenance se heurte aux mêmes obstacles. Au-delà de recommandations classiques comme la documentation de la ressource, la traçabilité des décisions liées à sa construction, plusieurs autres pistes sont envisageables au service d'une meilleure gestion de la maintenance : *capitaliser les expériences d'utilisation des logiciels et les démarches de modélisation, et les expériences de maintenance ; archiver ensemble corpus et ressources, voire même les outils spécifiques ayant servi à les construire ; décider si la ressource doit comporter une « épaisseur diachronique» ; outiller l'intégration des variations dans les ressources*.

Les recherches en apprentissage fournissent ici des repères et des formalisations 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 manière prometteuse d'envisager l'acquisition de RTO souligne l'intérêt croissant des collaborations entre TAL et apprentissage.

4.4. Définir des modes d'évaluation et de validation

Nous distinguons l'évaluation d'une RTO particulière construite dans un contexte donné, de celle d'un logiciel d'aide à leur construction de

l'évaluation des recherches. Dans les deux premiers cas, il s'agit d'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 des ressources ou de l'outil. Ces deux 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 donné aux résultats obtenus selon les méthodes identifiées que la reconnaissance des fondements scientifiques sur lesquels ils s'appuient.

En ce qui concerne les ressources, distinguons *validation*, qui renvoie à divers types de vérifications de la ressource à différents moments de sa construction, et *évaluation*, qui renvoie à la satisfaction du cahier des charges par la ressource terminée et dans son contexte d'usage. La *validation* fait intervenir, au cours du processus de modélisation, toute ressource garante des connaissances du domaine, en priorité les experts et le corpus. L'enjeu est de s'assurer que la conceptualisation élaborée par l'analyste n'est pas en contradiction avec ces connaissances de référence d'une part, et avec le rôle de la ressource auprès des utilisateurs dans l'application cible d'autre part. *L'évaluation* doit être réalisée selon les procédures de base du génie logiciel. La difficulté est que la RTO 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.

L'évaluation des logiciels et méthodes est donc étroitement liée à celle des ressources elles-mêmes, avec des difficultés propres : les logiciels sont des logiciels d'aide, et il est difficile de juger la part de l'intervention de l'analyste dans la pertinence des résultats ; ils sont rarement utilisés seuls, et la part de chacun des logiciels est difficile à évaluer.

Avec le souci d'afficher la qualité des RTO produites par les méthodes et logiciels issus de la recherche, la communauté scientifique commence à proposer des campagnes d'expérimentation permettant de mesurer la validité des ressources produites et des logiciels utilisés pour les construire. Les débats s'engagent sur les critères de validité à retenir pour dépasser des facteurs quantitatifs naïfs, entre le corpus de référence, les besoins des utilisateurs et une troisième série de facteurs, la capacité à réutiliser et maintenir cette ressource.

5. CONCLUSION

La problématique « corpus et terminologie » dans sa dimension interdisciplinaire a atteint un niveau de maturité évident. Le balisage effectué, conjointement à la dynamique de la recherche sur ces questions, a permis un éclairage qui devrait bénéficier à chaque discipline individuellement. De nombreuses questions, en partie dues à l'évolution très rapide de la numérisation, restent posées.

- Une première difficulté est liée à l'évolution très rapide des contextes, c'est-à-dire des besoins et des usages langagiers. Dans certains cas, il se peut que le temps de construction des RTO soit déjà plus long que la durée de l'application à laquelle elle serait destinée.
- La construction de RTO fait appel à l'interprétation humaine, qui garantit (souvent) sa pertinence mais qui a un coût financier et temporel. Des méthodes d'exploration à base quantitative seraient parfois peut-être plus adaptées.
- L'utilisation massive du web amène à s'interroger sur les possibilités de contrôler les textes qui s'y trouvent, par exemple en leur associant des méta-données. L'exploration non–contrôlée d'internet serait peut-être à envisager, ce qui remettrait en question la pertinence des RTO pour l'utilisation d'internet.

Malgré ces difficultés qui, dans certains cas, pourraient remettre en cause la pertinence des RTO, l'acquis en matière interdisciplinaire est indéniable. On le voit, la construction de RTO à partir de corpus s'inscrit dans des perspectives où les besoins, les modes d'utilisation, les modes de rédaction sont situés, c'est-à-dire anticipés. Or, c'est fondamentalement le besoin et l'envie de comprendre et d'interpréter ces situations qui ont rendu possible l'interdisciplinarité et qui ont contribué à stabiliser la problématique de l'accès aux textes. L'accélération numérique dont on ne connaît pas encore les conséquences n'affectera pas les acquis de cette interdisciplinarité et les perspectives qu'elle a permis de dégager.

6. REFERENCES

AMAR M. Les fondements théoriques de l'indexation : une approche linguistique. Paris : ADBS Éditions, 2000.

AUSSENAC-GILLES N., CONDAMINES A. *Corpus et terminologie. Rapport final de l'Action Spécifique ASSTICCOT*. Rapport IRIT/2003-23-R. Oct. 2003. 70 p.

AUSSENAC-GILLES N., CONDAMINES A., Documents électroniques et constitution de ressources terminologiques ou ontologiques. J. Charlet et J.-M. Salaun (eds) *Information, Interaction, Intelligence.* **4**, N°1, pp. 75-92. 2004.

BACHIMONT B., Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances. J.Charlet, M.Zacklad, G.Kassel, D.Bourigault (eds) : *Ingénierie des Connaissances, Evolution récentes et nouveaux défis*. Paris : Eyrolles. 305-324. 2000.

BIBER D., *Variation Across Speech and Writing*. Cambridge Univ. Press. 1988.

BOURIGAULT D., AUSSENAC-GILLES N., CHARLET J. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle (RIA*). PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. 18 (1) : 87–110. 2004.

CABRE M.-T., *La terminologie, Théorie, méthode et applications*. Paris : Armand Colin, Les Presses de l'Université d'Ottawa. 1998.

CHAUMIER J., *Les techniques documentaires au fil de l'histoire, 1950-2000*. Paris : ADBS Editions, 2002.

CONDAMINES A., Corpus Analysis and Conceptual Relation Patterns. *Terminology*, volume 8 number 1, 141-162. 2002.

CRUSE D.A., *Lexical Semantics.* Cambridge : Cambridge University Press. 1986.

FOUCAULT M., *Les mots et les choses*. Paris : Tel, Gallimard. 1966.

GRUBER, T. R.. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5: 199-220. 1993.

HABERT B., ZWEIGENBAUM P., Contextual Acquisition of Information Catégories : what has been done and what can be done automatically ? . Nevin B. (ed) : *The Legacy of Zellig Harris : Language and Information into 21st century*, volume 2. Amsterdam : John Benjamins. 2002 .

HABERT B., ILLOUZ G. FOLCH H., Des décalages de distribution aux divergences d'acceptation. in *Sémantique et Corpus* , dir. A. Condamines, 276-318. Lavoisier, 2005.

IBEKWE F., CONDAMINES A., CABRE T., 2005 : Application-driven Terminology Engineering. *Terminology* n°11-1, 2005.

JOUIS, C., *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 en informatique, EHESS, Paris. 1993.

LAINE-CRUZEL S., *Conception de systèmes de recherche d'informations : accès aux documents numériques scientifiques*.HDR, Université Claude Bernard Lyon 1. 2001. http://www.recodoc.univ-lyon1.fr/hdr_SLC.pdf

L'Homme M.-C..A Lexico-semantic Approach to the Structuring of Terminology. *Proceedings of Computerm'2004*, Coling 2004, Université de Genève. 7-14. 2004.

LYONS J., : *Eléments de sémantique*. Paris : Larousse Universités. 1978

MARSHMAN, E., MORGAN T., MEYER I.: French patterns for expressing concept relations. *Terminology*, volume 8 number 1. 1-30. 2002.

NAZARENKO A., Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? in *Sémantique et Corpus* , dir. A. Condamines, 221-244. Lavoisier, 2005.

PEARSON J., *Terms in Context,* Amsterdam: John Benjamins. 1998.

PEDAUQUE R.T., Le document : forme, signe et medium les re-formulations du numérique, STIC-CNRS. 2003, http://archivesic.ccsd.cnrs.fr/sic_00000511.html 2003

PEDAUQUE R.T., Le texte en jeu, Permanence et transformations du document, STIC-SHS-CNR,2005, http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/14/01/index_fr.html

QUILLIAN R., Semantic Memory, MINSKY M. (ed) : *Semantic Information Processing*,., Cambridge, Mass. M.I.T. Press, 227-70, 1968.

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

WUSTER E., 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. 55-108. 1981.