Documents électroniques et constitution de ressources terminologiques ou ontologiques

Nathalie Aussenac-Gilles, Anne Condamines

Documents électroniques et constitution de ressources terminologiques ou ontologiques

Nathalie Aussenac-Gilles*, Anne Condamines†

* IRIT, Université Toulouse 3, 118, route de Narbonne, 31062 Toulouse Cedex, aussenac@irit.fr
† ERSS, Université Toulouse 2, Maison de la recherche, 5, allées Antonio Machado, 31043 Toulouse Cedex, anne.condamines@univ-tlse2.fr

Résumé

En s'intéressant à la notion de document électronique, cet article propose des éléments pour mieux comprendre les enjeux des sciences de l'information lors de la construction et de l'utilisation de terminologies en lien avec des corpus. La notion d'usage est un fil conducteur qui nous permet d'interroger à la fois l'utilisation des ressources construites au sein des applications et leur mode de construction, à partir de textes, eux-mêmes considérés comme la mise en oeuvre de régularités linguistiques, relevant d'usages du système.

Mots-clés *: terminologie, linguistique de corpus, ressources terminologiques, ontologies, traitement automatique du langage naturel, usages, sciences de l'information.*

Abstract

By focusing on the notion of electronic document, this paper suggests some items to better take into account the scope of information sciences when building and using corpus-based terminologies. The notion of usage is a guideline that we follow to investigate both the way designed resources are used within target applications and the way they are built up, from texts that are themselves considered as revealing linguistic regularities, that refer to the system use.

Key-words *: terminology, corpus linguistics, terminological resources, ontology, natural language processing, usages, information sciences.*

1 INTRODUCTION

La disponibilité sous format électronique des textes a eu comme effet majeur de mettre en lumière des convergences entre des disciplines qui, sans s'ignorer, n'avaient pas toujours établi des ponts entre elles. Ce constat est flagrant lorsqu'il s'agit de construire, à partir de textes, des ressources qui entretiennent des parentés évidentes comme les listes de termes, reliés ou non par des relations, que l'on retrouve dans de nombreuses disciplines et qui sont, de plus en plus souvent, construites à partir de textes. Cette problématique a motivé la fondation du groupe de recherche français Terminologie et Intelligence Artificielle (TIA)[1], qui rassemble des chercheurs en terminologie, linguistique de corpus, traitement automatique des langues (TAL) et ingénierie des connaissances (IC). Telle qu'elle est formulée par TIA, la question porte avant tout sur l'analyse des contenus et la théorisation des méthodes et des modèles, mais laisse de côté la question des supports numériques autant que celle de l'association entre contenu et support. Autrement dit, ces analyses ne s'intéressent pas directement à la notion de document.

La notion de document reste ainsi fondamentalement attachée aux sciences de l'information. Or de plus en plus, les sciences de l'information rejoignent la problématique du groupe TIA, d'une part, parce qu'elles sont aussi confrontées à la représentation des connaissances sous forme de concepts et de relations au sein de thésaurus et d'index, et d'autre part, parce que la numérisation a provoqué des bouleversements dans la pratique documentaire. Un des objectifs de l'Action Spécifique «Corpus et Terminologie»[2], que nous avons animée, a consisté précisément à établir les convergences entre ces disciplines autour de ce thème.

L'article aborde l'accès aux contenus documentaires par des applications informatiques faisant appel à des ressources terminologiques et ontologiques. Il se focalise sur le point de vue des sciences de l'information et sur l'enrichissement qui peut provenir de la confrontation des pratiques en sciences de l'information avec celles mises en œuvre

[1] http://www.biomath.jussieu.fr/TIA Groupe de travail du GDR I3

[2] Action spécifique ASSTICCOT (2002-2003) associée au RTP-DOC (33) du département STIC du CNRS, http://www.irit.fr/ASSTICCOT/

pour la constitution de ressources à partir de corpus. La présentation se centre ainsi principalement sur la notion d'usage, initialement empruntée aux sciences de l'information. On retrouve cependant cette notion aussi bien en ingénierie des connaissances, qui met ainsi l'accent sur l'utilisateur d'un système faisant appel à la ressource terminologique, que dans la linguistique de corpus ou le traitement automatique des langues, qui l'utilisent pour évoquer l'utilisation réelle de la langue dans des domaines et des discours particuliers. Dans les sciences de l'information, cette notion d'usage intervient à différents moments du processus (constitution de thésaurus, d'index ou recherche d'information). Autrement dit, les usages en ingénierie des connaissances relèvent des pratiques de construction de connaissances, tandis que pour la terminologie ou la linguistique de corpus, il s'agit de rendre compte des pratiques langagières inscrites dans les textes. Les sciences de l'information, quant à elles, ancrent cette problématique à la fois dans l'adaptation à la demande d'un utilisateur (recherche d'information) et dans la prise en compte de régularités rédactionnelles.

Ces divers points de focalisation d'une même notion nous amènent à présenter un point de vue global sur la construction de ressources terminologiques et ontologiques à partir de corpus. Ils nous permettent aussi d'expliquer pourquoi les deux axes proposés par R. T. Pédauque [20] «document comme signe» et «document comme medium» nous semblent non seulement complémentaires mais aussi étroitement liés.

Notre réflexion s'articule en trois parties. Nous replaçons notre questionnement dans le cadre de la problématique de l'accès au contenu des documents. Nous posons ensuite la question des contextes d'usage en lien avec les types de ressources et d'applications, avant d'aborder dans une dernière partie celle d'une confrontation des usages langagiers à une meilleure caractérisation des genres textuels.

2 PRATIQUE DOCUMENTAIRE ET CONSTRUCTION DE RESSOURCES TERMINOLOGIQUES

2.1 Constitution de ressources terminologiques à partir de corpus

La mise à disposition auprès d'utilisateurs de documents sous format électronique représente aujourd'hui un véritable enjeu scientifique. Cet

enjeu, associé à la demande sociale en lien avec le traitement des données textuelles contenues dans ces documents, a fait émerger une problématique nouvelle, visant à modéliser le contenu de documents sélectionnés ou formant une collection sous la forme de réseaux de termes pour permettre un meilleur accès à la connaissance. Ces modèles ou représentations peuvent être entre autres des thésaurus, des terminologies, des langages documentaires, des index ou des ontologies. Nous les appellerons par la suite des ressources terminologiques ou ontologiques (RTO). Jusqu'ici disparates car répondant à des problèmes et des besoins différents dans des contextes techniques variés, ces ressources tendent à acquérir des caractéristiques de plus en plus proches.

Depuis ces dix dernières années, un mouvement de convergence a conduit à fédérer les recherches relatives à la mise au point de ces structures de données, de manière à rendre plus rapide et plus pertinent leur contenu. Plusieurs disciplines, dont le matériau d'étude est constitué pour l'essentiel soit de textes, soit de représentations lexicales ou conceptuelles, se retrouvent dans cette problématique : la linguistique de corpus, la terminologie, et, en informatique, la recherche d'information, le traitement automatique des langues (TAL), l'ingénierie des connaissances (IC) et l'apprentissage pour la fouille de textes.

Ces disciplines s'interrogent sur les apports mutuels de leurs travaux aux problèmes de la construction et de la gestion des terminologies, bases de connaissances terminologiques et ontologies. En particulier, des outils et méthodes utiles à la construction de ces ressources à partir de textes, essentiellement basés sur des analyses terminologiques, morpho-syntaxiques et sur l'étude de leurs occurrences et distributions en corpus, ont été mis au point ou identifiés [16, 5, 10]. Mais, comme nous l'avons souligné en introduction, jusqu'à il y a peu de temps, ces analyses ne s'intéressaient pas directement à la notion de document en tant qu'il est à la fois support et contenu. En d'autres termes, le point de vue des sciences de l'information n'avait pas encore été intégré à la réflexion.

2.2 Trois tâches clés de la recherche d'information

Les sciences de l'information sont principalement concernées par trois tâches, qui ont lieu à des moments nettement distincts et qui sont souvent réalisées par plusieurs documentalistes :

1- construction de thésaurus ou de langages documentaires pertinents pour un fonds documentaire ;

2- indexation d'ouvrages à partir de thésaurus existants ;

3- recherche de textes pertinents en fonction d'une demande d'information particulière.

Ces trois moments font intervenir différents types d'acteurs. En 1, un documentaliste construit un thésaurus en direction d'un autre documentaliste (les deux rôles pouvant être tenus par la même personne). Cette tâche est surtout centrée sur la couverture du fonds, indépendamment d'utilisateurs particuliers. Il s'agit donc d'assurer que l'ensemble des usages langagiers (considérés comme autant de traces de connaissances) est pris en compte. En 2, un documentaliste utilise le thésaurus construit en 1 pour identifier des mots-clés pertinents à la fois du point de vue de l'ouvrage à indexer (usages langagiers) et de l'utilisateur pressenti (le documentaliste connaît les utilisateurs de son fonds et leurs besoins). En 3, la tâche est centrée sur un utilisateur particulier, auquel il faut fournir des textes pertinents et ce, même si, apparemment en tout cas, ces textes ne font pas partie d'un classement préétabli, *i.e.* n'obéissent pas forcément à une homogénéité langagière.

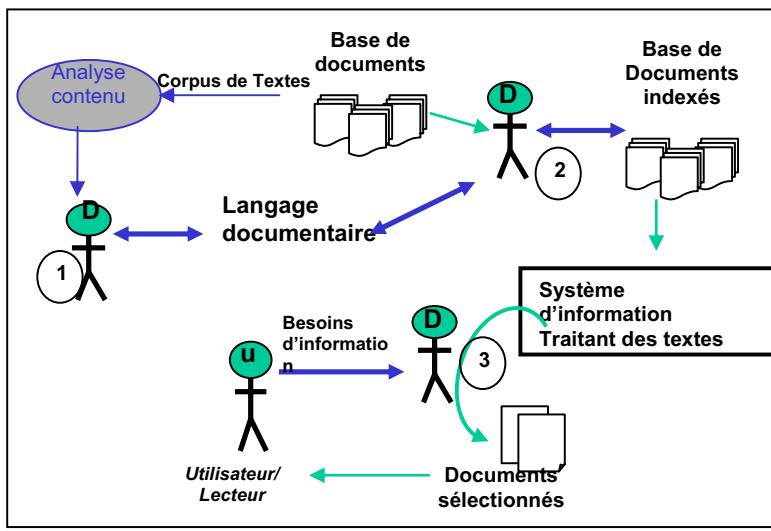

FIG. 1 – *Tâches du documentaliste, notées de 1 à 3 autour des personnages D (documentaliste) et U (utilisateur), pour préparer la recherche d'information en indexant les collections*

2.3 Convergences possibles

Les convergences entre ces disciplines (sciences de l'information d'une part, linguistique de corpus, ingénierie des connaissances et de la langue d'autre part) n'ont pu voir le jour que parce que la plupart des textes, objets de leurs investigations, sont désormais sous format électronique. En effet, avant ce phénomène de numérisation généralisée, la notion de document restait très cantonnée dans le domaine des sciences de l'information, le document prenant son statut parce qu'il faisait partie d'une collection organisée en fonction d'un point de vue relatif aux besoins en information (des utilisateurs). Parler de document, c'est ainsi, inévitablement, se situer du point de vue des sciences de l'information. La généralisation de la diffusion de documents sous format électronique, accompagnée de la mise à disposition de logiciels d'analyse ou d'exploration, a eu comme effet de réduire le temps entre la constitution d'une ressource de référence et la mise à disposition de textes auprès des utilisateurs [18]. Ce phénomène diminue aussi le nombre d'intervenants et, dans le cas de la recherche d'information sur le web par mots-clés, il supprime carrément tout intermédiaire (humain mais aussi matériel, au sens d'une terminologie prédéfinie) !

Le point de vue des sciences de l'information permet un éclairage original sur la problématique de la constitution de RTO à partir de textes. En effet, seule une partie des tâches relevant des sciences de l'information semble être prise en compte dans cet objectif. Si les autres tâches en paraissent plus éloignées, elles peuvent toutefois alimenter des réflexions qui intéressent l'ingénierie des connaissances et la terminologie textuelle. Ainsi, la construction de RTO examinée par ces deux disciplines englobe dans un même moment les tâches 1 et 3. Il s'agit en fait de construire une ressource équivalente à un thésaurus au sens où elle servira de référence. Mais cette construction est faite en fonction d'une utilisation définie d'emblée, c'est-à-dire dans une perspective qui prend en compte à la fois les tâches 2 et 3. La constitution de RTO peut alors être vue comme une sorte de compactage d'un ensemble de tâches réalisées à différents moments, par différentes personnes dans l'ensemble du processus mené par les sciences de l'information. Si ce compactage est devenu possible, c'est bien parce que les textes (corpus, documents) sont maintenant disponibles sous forme numérique.

Un autre mouvement de convergence porte sur la nature même de ces ressources terminologiques ou conceptuelles : thésaurus, terminologies, langages documentaires, bases de connaissances terminologiques ou ontologies. Jusqu'ici, les caractéristiques propres à chaque besoin avaient

conduit à définir des structures de natures différentes. Leur mise sous format informatique, des bases de données aux modèles désormais enrichis, autant que le statut des informations qu'elles contiennent tend à les rapprocher : de plus en plus de terminologies sont associées à un réseau conceptuel afin de définir les termes avec plus de précision et de souplesse. Cependant, chacune renvoie encore à une réalité différente qu'il est primordial d'avoir à l'esprit au moment de les construire puis de les utiliser. Par exemple, une terminologie renvoie soit à une liste de mots faiblement structurée hiérarchiquement, le plus souvent non formelle. Un langage documentaire est une ressource structurée conçue pour des utilisateurs (documentalistes indexeurs ou utilisateurs finaux), alors qu'une ontologie doit avant tout permettre à un logiciel d'effectuer des inférences.

2.4 Une vue unificatrice

La problématique que nous avons privilégiée est celle qui nous semble la mieux fédérer les réflexions des trois principales disciplines concernées : les sciences de l'information, la terminologie textuelle et l'ingénierie des connaissances. Les trois éléments qui caractérisent cette convergence sont d'abord le recours à des textes (organisés sous forme de corpus ou de collections), ensuite l'objectif de modélisation sous la forme de représentations relationnelles (les RTO) et enfin, la prise en compte de l'utilisation de ces RTO. Dans ce scénario fédérateur, les textes (au sens des contenus informationnels des documents) sont à la fois sources de connaissances pour construire des RTO et sources directes de connaissances présentées pour répondre aux besoins de l'utilisateur. En amont, ils sont choisis comme traces « objectivables » de connaissances supposées partagées et assez consensuelles, bénéficiant de l'intertextualité pour rendre accessibles ces connaissances. En aval, ils sont laissés à l'interprétation directe ou indirecte (par le biais des RTO) de l'utilisateur au sein d'applications. Dans tous les cas, le contenu du document est exploité via des analyses ou lectures transverses, qui ne respectent pas la linéarité de la présentation. Le document est ainsi détourné de la lecture prévue par son auteur.

La rencontre interdisciplinaire conduit à rapprocher les situations de construction et d'utilisation de ces ressources des tâches du documentaliste (Fig.2). D'une part, l'utilisateur est intégré comme un des acteurs majeurs du processus. Pratiquement, dans un contexte comme la consultation du web via un moteur de recherche, c'est lui qui réalise la tâche de recherche d'information (tâche 3), sans intervention du documentaliste. De plus, la tâche d'indexation (tâche 2) est le plus

souvent prise en charge par des modules de traitement automatique des langues (TAL2 sur la figure 2) au sein de l'application finale. D'autre part, la position des documentalistes dans leur tâche d'élaboration de langage documentaire (tâche 1) se rapproche de celle des autres constructeurs de RTO (linguistes, terminologues ou ontologues). Ce « déplacement » de rôle s'accompagne d'une intégration de l'expérience des sciences de l'information à la problématique de la construction de RTO, en particulier en ce qui concerne la prise en compte des utilisateurs. Elle permet aussi de théoriser le problème de manière plus homogène et plus consensuelle. L'absence du documentaliste en phase de recherche d'information rend d'autant plus fondamentale la tâche préparatoire de définition d'une RTO pouvant guider cette recherche.

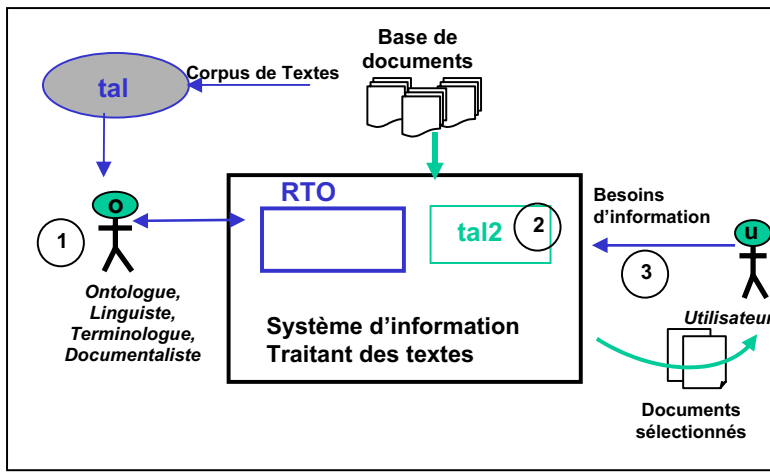

FIG. 2 – *Place des RTO dans l'accès au contenu de documents numériques*

Finalement, deux éléments apparaissent comme cruciaux dans le processus de constitution de RTO à partir de corpus. D'une part, l'utilisateur (ou l'utilisation) final(e) apparaît comme premier dans le processus. Il (elle) guide non seulement les choix de constructions mais aussi la constitution du corpus. D'autre part, le corpus justement joue un rôle de référence majeur puisque c'est à partir des fonctionnements langagiers (usages) qu'il manifeste que peut s'élaborer un type ou un autre de représentation. Ces deux éléments sont travaillés dans la prochaine partie à travers une réflexion sur la notion d'usage.

3 USAGES DES DOCUMENTS, USAGES DANS LES DOCUMENTS

En sciences de l'information, les usages ont plutôt été examinés sous l'angle de la prise en compte de l'utilisateur, qui se fait à un moment avancé du processus, les usages langagiers étant plutôt étudiés au moment de la constitution des thésaurus. En linguistique textuelle, les usages concernent plutôt la mise en œuvre réelle d'une langue, c'est-à-dire la prise en compte des *usages* (langagiers) *dans le document* (cf. 3.3). En IC et de plus en plus souvent en terminologie textuelle, les usages sont examinés du point de vue des applications et ils sont pris en compte très tôt dans le processus. Il s'agit donc d'interroger les *usages des documents*. Cette question des applications est examinée dans le prochain paragraphe (3.1). Parce que les besoins autant que les collections interrogées évoluent très rapidement, le paragraphe 3.2 concerne lui la question de la *maintenance des RTO* en accord avec de *nouveaux usages*.

3.1 Usages et types de ressources terminologiques

Le document électronique joue donc, dans notre perspective, le rôle de source de connaissances à double titre : il est consulté (plus ou moins directement) par des utilisateurs qui ont un besoin d'information, mais aussi par des médiateurs, comme les terminologues ou les cogniticiens, qui les analysent pour produire des ressources intermédiaires (les RTO) facilitant la recherche par l'utilisateur final. L'étude des contextes d'usages en consultation directe ou assistée par un logiciel souligne la diversité des besoins et des applications y répondant. La confrontation des expériences et recherches menées dans les différentes disciplines représentées dans ASSTICCOT a débouché sur le constat commun que des ressources terminologiques ne peuvent assister ce processus que si elles sont pensées et construites en fonction de ces usages.

Reconnaître cet état de fait traduit une position tranchée par rapport au courant fondateur de la terminologie, mais aussi par rapport à la vision classique des ontologies en IC et, plus encore peut-être, avec la vocation très normative des langages documentaires. Les évolutions du domaine de la terminologie illustrent bien ce débat qui a influencé tout le courant anglo-saxon et conduit aux ontologies actuelles [15]. Depuis les années 30, la Théorie Générale de la Terminologie défendait une vision unificatrice : le monde de la connaissance est découpé en domaines

stables, dont chacun est équivalent à un réseau fixe de concepts, les termes étant les représentants linguistiques de ces concepts. Or, le rapprochement récent entre terminologie et informatique débouche sur de nouvelles applications et sur un élargissement de la gamme des produits terminologiques qui vient bouleverser cette position [21]. Le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas une terminologie, mais autant de RTO que d'applications qui les utilisent. Selon l'application, ces ressources peuvent différer sensiblement quant aux connaissances auxquelles elles renvoient (unités retenues et leur description), mais aussi quant à leur degré de structuration ou de formalisation. Ces constats empiriques entraînent des changements de fond de la pratique terminologique, et appellent à un renouvellement théorique, que l'on peut rapprocher des changements de perspective touchant, en IC, la notion d'ontologie [13].

Le contexte d'usage d'une RTO a de multiples conséquences sur le processus qui va des corpus d'origine ou consultés à l'application en passant par la constitution de cette ressource [10]. Parmi les nombreux paramètres qui entrent en jeu, ont été identifiés les textes disponibles, les acteurs intervenant dans le projet (linguistes, experts du domaine, analystes ou cogniticiens), les utilisations et les utilisateurs visés [4]. Ces paramètres sont déterminants dans le choix de logiciels adéquats pour l'analyse des textes autant que pour la structuration et la représentation des connaissances. Chaque facteur doit être pris en compte pour espérer construire des produits utilisables et utilisés. Il s'agit alors de repérer, à partir d'expériences précises [1], des éléments pouvant guider, en fonction de l'application ciblée, le choix de ces paramètres : déterminer le profil de l'analyste qui construit la ressource, fournir des repères pour construire un corpus pertinent, caractériser les potentiels et contraintes des outils de TAL pour les utiliser au mieux, qualifier les exigences de structuration et de formalisation pour décider de l'outil de modélisation à retenir. Notons que quelques environnements de construction d'ontologies prévoient justement un couplage avec des logiciels de TAL pour s'appuyer sur des analyses de corpus [23] [25].

Ces réflexions dessinent des perspectives de recherches indispensables pour produire plus efficacement des ressources mieux adaptées [5]. Une première difficulté à dépasser concerne la méthode d'étude du problème. Alors que le champ théorique est relativement bien balisé [6], les avancées ne peuvent venir que de la mise en perspective de retours d'expérience, et de la confrontation des communautés d'utilisateurs avec les problèmes rencontrés et les solutions choisies. Un autre enjeu porte sur la disponibilité et la modularité des logiciels et des résultats, dont

l'intégration au sein de tâches plus complexes constitue encore un verrou. Enfin, cette étude fait ressortir les évolutions nécessaires des structures de représentation des RTO. Les coûts de développement élevés ne sont plus compatibles avec la dynamique des contextes d'utilisation, et appellent à imaginer des structures plus souples, plus faciles à élaborer, et qui conservent les outils de construction associés aux données.

3.2 Usages et maintenance des RTO

Dans le même esprit, parce que les contextes d'usage évoluent, se pose la question de la validité des ressources construites dans le temps et, par là, de leur maintenance. Alors que celles-ci étaient construites jusqu'ici avec une hypothèse forte de stabilité, justifiant un coût élevé de construction, l'accélération des modifications des contextes d'usage (enrichissement des bases documentaires consultées, évolutions terminologiques rapides ou encore arrivée de nouveaux besoins) requiert d'anticiper les problèmes d'évolution et de mise à jour des RTO. Ceci suppose de s'interroger sur les évolutions (des besoins, des documents ou du vocabulaire) qui justifieront des mises à jour, d'identifier ce sur quoi les mises à jour porteront (déterminer ce qui reste fixe par rapport à ce qui doit évoluer), de décider si la ressource doit comporter ou non une épaisseur diachronique et de juger de la cohérence à préserver ou non entre les différentes versions. Pour mieux anticiper ces problèmes de maintenance, il faut assurer une meilleure traçabilité et reproductibilité des processus de modélisation.

Les perspectives requises pour aller dans ce sens passent par la capitalisation des retours d'expériences d'utilisation des outils et méthodes. Elles supposent aussi de se donner les moyens d'archiver conjointement outils et ressources. Deux pistes sont envisageables pour outiller l'intégration des variations dans les ressources : un processus incrémental, s'appuyant sur le repérage d'écarts, permet des mises à jours régulières sans bouleverser la ressource ; un processus itératif, plus radical, consiste à reconstruire à partir de nouveaux textes une nouvelle ressource. Plus encore que la première, cette seconde voie nécessite de faire appel à des outils automatisant, même en partie, l'apprentissage de nouvelles connaissances [19].

3.3 Usages et genre

Travailler la notion de genre est une façon de prendre en compte les usages dans les deux sens que nous leur avons donnés : aussi bien du

point de vue des types d'applications que de celui des régularités de rédaction. Cette notion permet en effet d'inscrire les usages dans une perspective collective qui serait plus ou moins systématisable (c'est-à-dire plus ou moins prévisible). L'hypothèse est qu'il n'y a pas autant d'usages que d'individus mais bien des « genres » d'usages.

La notion de genre est utilisée par différentes disciplines, en lien avec celle d'usage. Selon les points de vue (analyse de discours, linguistique de corpus, TAL, sciences de l'information), ces deux notions, usages et genres, sont combinées différemment. Il importe de les resituer afin de comprendre les enjeux dans lesquels elles s'inscrivent et d'évaluer les complémentarités et les irréductibilités entre points de vue.

Premier constat : dans tous les cas où ces deux notions sont utilisées, il s'agit de regrouper des textes pour constituer des corpus homogènes. L'hypothèse commune est que c'est sur la base de similarités en lien avec l'appartenance à un même genre que pourrait s'organiser la catégorisation de textes (ou documents). La notion de genre faisant toujours intervenir la dimension extra-linguistique, c'est aussi une façon de penser, pour reprendre le texte de Pédauque, qu'un texte (ou un document) est certes un signe mais qu'il est toujours aussi simultanément un médium. Autrement dit, il n'y a pas de signe sans que ce signe soit inscrit de manière intime dans un contexte social. Ainsi, si un texte (ou un document) est bien un signe, c'est toujours un signe situé.

Mais une répartition se fait parmi les disciplines entre, d'un côté, l'analyse de discours, la linguistique de corpus et le TAL et, de l'autre, les sciences de l'information. En effet, pour les trois premières, la notion de genre s'institue autour du rassemblement de textes censés faire correspondre régularités langagières et situations extra-linguistiques de communication alors que pour la dernière, la notion de genre s'élabore d'abord autour de régularités de situations d'usages de documents.

3.4 Genres de discours et usages documentaires

3.4.1 Genres de discours

La notion de genre est majoritairement utilisée en analyse de discours. Elle est comprise comme un palier d'organisation de régularités langagières dépendant de régularités extra-linguistiques, situé entre l'énoncé individuel et les « formes de langue » [7]. Même si la notion de genre existe depuis l'antiquité, où elle était proche de celle de rhétorique

[11] c'est certainement Bakhtine qui, en URSS dans les années 1930-1940, l'a le premier approfondie .

« Tout énoncé pris isolément est, bien entendu, individuel, mais chaque sphère d'utilisation de la langue élabore ses types relativement stables d'énoncés, et c'est ce que nous appelons les genres de discours » ([7], p. 265).

Ainsi définie, cette notion pose un certain nombre de questions : présence de plusieurs genres dans le même texte, différences de points de vue d'analyse des genres, instabilité des situations de communication [12], [21]. En particulier, on distingue mal chez Bakhtine comment s'organise le lien entre les régularités extra-linguistiques et les régularités linguistiques ; le genre concerne-t-il le langagier, l'extra-langagier ou la covariance des deux ? On liste parfois des genres qui seraient repérés dans la langue comme l'article de journal, la lettre administrative, le manuel technique ... En réalité, on fait l'hypothèse qu'à ces situations de communication correspondent des régularités linguistiques. Et dans un certain nombre de cas, on peut trouver des indices linguistiques qui peuvent confirmer l'existence de ces genres. Par exemple, dans [8], il est confirmé que certains fonctionnements langagiers peuvent être caractéristiques du roman policier. Cela signifie aussi que tous les textes qui auront les caractéristiques réputées propres à ces discours pourront être considérés comme relevant de ce genre-là.

Inversement, des travaux de la linguistique anglo-saxonne comme ceux de Biber tendent à montrer qu'il faut distinguer caractérisation des situations extra-linguistiques, appelées « genre » chez Biber, et régularités linguistiques qui permettent de définir des « types » de textes qui peuvent remettre en question la caractérisation initiale faite sur des bases intuitives. Ainsi, l'objectif de Biber consiste à essayer de donner une assise linguistique à ce qui relèverait d'une classification intuitive [9].

Telle que l'entend Bakhtine, la notion de genre possède finalement les caractéristiques suivantes :

• Elle s'inscrit nécessairement dans une perspective dialogique du fonctionnement linguistique. Même s'il n'est pas un dialogue, tout texte s'inscrit comme la suite d'un dialogue instauré par les textes précédents, de soi ou d'autrui.

• Elle se met en place à l'insu des locuteurs. Lorsqu'ils obéissent à des régularités propres à un certain genre, les locuteurs ne font pas le choix de ces régularités. C'est parce qu'ils ont inconsciemment intégré

des fonctionnements langagiers qu'ils les mettent en œuvre dès qu'ils s'inscrivent dans une situation donnée.

- Elle a un caractère normatif très net ; comme le signale Todorov, « le genre forme un système modélisant qui propose un simulacre du monde » ([24], p. 128).

On verra que quel que soit le contexte dans lequel elle est employée, cette notion de genre conserve le plus souvent ces trois caractéristiques.

3.4.2 Typologie d'usages documentaires

La notion d'usages est centrale en sciences de l'information. C'est elle qui permet de rassembler des documents pour les mettre à disposition des utilisateurs sous forme de collections. La notion de similarité d'usage crée l'homogénéité. Mais, pour les sciences de l'information, la typologie des usages ne serait pas libre ; elle serait elle-même conditionnée par la notion de genre qui préexisterait à l'usage et qui constituerait des modes d'interprétation :

«[…] la fondation des genres s'appuie sur un usage visant à faciliter l'interprétation par le lecteur » « La catégorie des genres … constitue pour le lecteur un script. » [3].

Pour les sciences de l'information, la notion de genre semble ainsi majoritairement orientée vers le lecteur en tant qu'il bénéficie du classement effectué par les documentalistes. Il ne s'agit donc pas du lecteur direct d'un texte mais d'un lecteur tel qu'il est modélisé par les documentalistes. La notion de genre est ainsi moins associée aux régularités de rédaction qu'aux régularités d'interprétation [14]. Ainsi, lorsqu'elle est mise en œuvre par les documentalistes, la notion de genre est une méta-notion qui concerne à la fois une modélisation des textes (grâce à la notion de genre de discours) et une modélisation des usages (typologies de lectures). Cette double modélisation est celle qui préside à la constitution des collections mais aussi à l'attribution des mots-clés. Le choix de mots-clés pour des textes relève d'une interprétation dont le résultat se manifeste principalement par la représentation d'un texte sous la forme d'éléments lexicaux. Cette élaboration se fait à la fois sur le modèle qu'a le documentaliste du document à indexer et sur celui qu'il a de l'usage qui va en être fait, c'est-à-dire sur le genre de discours dont relève le document et sur le genre d'usage dont relève la lecture. Amar fait le même constat de cette double influence à propos de la pratique terminologique :

« une approche linguistique de la pratique terminologique […] - d'une part montre que les termes des terminologies sont indissociablement liés aux discours qui les instituent, - d'autre part, elle souligne que les termes ne sont pertinents que dans le cadre restreint d'une pratique technique ou scientifique donnée » ([2], p. 191).

On retrouve dans l'utilisation de la notion de genre qui est faite par les sciences de l'information plusieurs des caractéristiques décrites par Bakhtine :

- les genres préexistent à la tâche de documentation,
- ils se mettent en place à l'insu de ceux qui les utilisent,
- ils constituent une sorte de cadrage du monde.

La situation des documentalistes, mais aussi de tous ceux qui utilisent des textes pour construire des représentations censées être pertinentes pour des utilisateurs, se trouve ainsi au centre d'un ensemble de voix (au sens de Ducrot [17]) puisque s'y retrouvent :

- des locuteurs, rédacteurs de textes,
- les interlocuteurs de ces textes (lecteurs),
- des interprètes de ces textes (qui les constituent en corpus ou collections),
- des utilisateurs de ces corpus ou collections.

.Pour chacun de ces participants, la notion de genre est pertinente avec ses trois caractéristiques (modélisation du monde, non-conscience des locuteurs, dialogisme) mais il est probable qu'elle ne soit pas identique en fonction des protagonistes. *La question qui se pose alors est celle de savoir s'il est possible de constituer une théorie du genre dans cette situation précise qui consiste à construire des listes ou des réseaux de termes à partir de textes.*

3.5 Vers une unification de la notion de genre ?

La diversité des genres qui s'entrecroisent dans la construction de terminologies à partir de corpus peut laisser penser qu'une tentative de systématisation est vouée à l'échec. Toutefois, plusieurs éléments semblent plaider en faveur d'une stabilisation :

- Puisqu'il s'agit de constructions à partir de textes, le substrat langagier joue un rôle majeur : il est au centre d'un faisceau de points de vue et il peut constituer l'élément stable dans une situation très variable.

- Certains critères de classement, certes généraux, semblent assez récurrents d'une discipline à l'autre, par exemple, date et lieu de

production, niveau de compétence du/des rédacteur(s) et du/des lecteur(s), objectif initial, public visé…. Des projets européens, comme le projet PAROLE 3 ont d'ailleurs proposé d'utiliser ce type de critère.

- Le TAL offre des possibilités de tester des hypothèses très rapidement et d'organiser des textes en fonction, non d'une caractérisation a priori mais du repérage de régularités linguistiques. Cette ouverture, liée au format électronique des textes et au développement des outils d'analyse, joue un rôle fédérateur de différents travaux et offre une opportunité de mise en commun des réflexions.

- La construction de l'interdisciplinarité (en grande partie grâce au RTP-Doc) est maintenant bien avancée. Elle a certainement gagné à se focaliser sur certaines problématiques comme celle de la construction de RTO (Ressources Terminologiques et Ontologiques) qui ont en commun de consister en une représentation sous forme relationnelle de la connaissance, les termes étant les nœuds de ces réseaux.

4 CONCLUSION

Un regard pluridisciplinaire sur les liens étroits qui associent documents et ressources terminologiques fait émerger les situations d'usage des documents et des ressources comme des facteurs critiques pour déterminer l'adéquation entre contenu des documents et ressources d'une part, ressources et utilisations d'autre part. Pour mieux caractériser ces liens, nous retenons la nécessité d'approfondir et de développer la notion de « genre textuel », qui est une des pistes possibles pour mieux caractériser (par des régularités linguistiques) les contenus documentaires du point de vue de leur production mais aussi de leur interprétation ; de prendre en compte les applications utilisant des ressources pour mieux cibler les types de ressources adéquats et leurs méthodes d'élaboration à partir de documents ; et enfin, d'anticiper les besoins en maintenance au niveau méthodologique et représentation des connaissances (par la capacité à archiver les moyens de reconstruction de la ressource dans un nouveau contexte). En effet, les ressources terminologiques sont l'objet d'un paradoxe : elles doivent à la fois « normaliser » des connaissances, c'est-à-dire les figer à un moment donné et être utilisées pour accéder à des connaissances qui évoluent dans des contextes dynamiques.

[3] Preparatory Action for Linguistic Resources Organisation for Language Engineering

Pour atteindre ces différents objectifs, une approche interdisciplinaire est indispensable afin d'appréhender la grande variabilité des besoins en information d'une part, des ressources et applications envisageables pour y répondre d'autre part. Réflexion théorique (sur l'évolution des connaissances) et adaptation des outils (de diagnostic, d'intégration des évolutions, d'archivage…) et des méthodes (rôle des méthodes d'apprentissage par exemple) doivent continuer à se développer conjointement dans le cadre de projets interdisciplinaires.

REFERENCES

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

[2] AMAR M : *Les fondements théoriques de l'indexation, une approche théorique.* Thèse en sciences de l'Information et de la Communication, Université Lyon II, 1997.

[3] ARASZKIEWIEZ J. : Notes sur la théorie du genre ; Colloque sur le genre journalistique prévu en septembre 2004. 16 déc. 2003. Working paper http://archiveSIC.ccsd.cnrs.fr/sic_00000855.html

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

[5] AUSSENAC-GILLES N., CONDAMINES A.. *Terminologies et corpus. Rapport final de l'Action Spécifique ASSTICCOT*. Rapport Interne IRIT/2003-23-R70 p. http://www.irit.fr/ASSTICCOT/ . Oct. 2003.

[6] BACHIMONT B., A*rts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle.* Habilitation à diriger des recherches de l'Université Technologique de Compiègnes. Janv. 2004.

[7] BAKHTINE M. *Esthétique de la création verbale*. Paris : Gallimard, Tel. Beauvisage T. 1984.

[8] BEAUVISAGE T. Morphosyntaxe et genres textuels. *TAL* 579-608. vol. 42-n°2/2001.

[9] BIBER D. Variation Across Speech and Writing . Cambridge University Press. 1988.

[10] 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), Numéro spécial « Terminologie ». J.M. Pierrel, M. Slodzian (Ed.). Hermès : Paris. Vol. 16. N°1/ 2004. pp. 87–110.

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

[12] BRONCKART J.-P. *Activités langagières, textes et discours* . Lausanne : Delachaux et Niestlé. 1996.

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

[14] CONDAMINES A. Vers la définition de genres interprétatifs. *Actes de TIA'2003*, Université Marc Bloch, Strasbourg, PP. 69-79. 2003

[15] CONDAMINES A. *Sémantique et corpus spécialisés : Constitution de bases de connaissances terminologiques*. Mémoire d'habilitation à diriger les recherches. *Carnets de grammaire*,n°13, ERSS, Toulouse. 2003

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

[17] DUCROT O : *Le dire et le dit*. Paris : Editions de Minuit, 1984.

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

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

[20] PEDAUQUE R. T. (2003), *Document : forme, signe et medium, les re-formulations du numérique*. Article de travail .http://archivesic.ccsd.cnrs.fr/sic_00000413.html

[21] RASTIER F. *Arts et Sciences du texte*. Paris : PUF, formes sémiotiques. 2001.

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

[23] STAAB, S., MAEDCHE, A. Ontology Learning for the Semantic Web*, IEEE Intelligent Systems, Special Issue on the Semantic Web*, 16(2), 2001.

[24] TODOROF T. *Mikhaïl Bakhtine, Le principe dialogique* suivi de *Ecrits du Cercle de Bakhtin*e, Paris : Seuil, 1981.

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