GESTION DU CORPUS DANS LA RECHERCHE TERMINOLOGIQUE

Lector univ. dr. Mihaela POPESCU
Universitatea „Transilvania", Braşov

Résumé

L'article propose une description des corpus en tant que représentations de connaissances. Les caractéristiques, les types de corpus, les critères de sélection des textes d'un ensemble de textes sont les éléments censés clé à aboutir à une possible gestion du corpus en terminologie tout en partant de l'idée que la saisie des relations entre les concepts et les termes, ainsi que les relations morpho-syntaxiques et paradigmatiques conduit à une gestion efficace des textes du corpus dans la recherche terminologique dans une ou plusieurs langues.

La représentation des connaissances sous la forme de liste de termes reliés par des relations est ancienne et courante. Nous vivons dans un environnement où les taxinomies utilisées dans les sciences naturelles au cours du XVII-ème et XVIII-ème siècles, ainsi que la classification universelle de Dewey (1876), les réseaux sémantiques de Quillian (1968) ou les ontologies de l'ingénierie des connaissances (Gruber, 1993) sont autant de modes de représentation qui mettent l'accent sur l'utilisation d'éléments lexicaux pour modeler la connaissance. Ces représentations sont soit employées dans des systèmes informatiques, soit constituent la base de langages de représentations (les graphes conceptuels), soit sont des logiques terminologiques. Les représentations sont fondées sur des systèmes relationnels. La structuration d'un réseau conceptuel à partir des termes relève d'une interprétation, d'une normalisation (Bachimont, 2000).

Dans ces conditions, nous avons besoin des textes, réunis dans un corpus à partir desquels nous devons construire les ressources terminologiques ou ontologiques. Ce mode de représentation de la connaissance est important dans le cas des textes spécialisés, d'une part, et de profiter du potentiel de l'informatique, d'autre part. Il s'ensuit qu'il faut s'interroger sur les liens entre des discours et des éléments lexicaux en utilisant les seconds pour arriver aux premiers. La question est celle de savoir comment gérer et contrôler l'information d'un corpus, étape principale dans la recherche terminologique, qui constitue la base de l'extraction des termes et des structures prêtes ensuite à être traitées automatiquement dans une ou plusieurs langues.

Le développement de l'informatique a contribué à l'accélération dans la réflexion sur la terminologie. L'informatisation des textes, le développement des outils pour les interroger, d'une part, la demande de la part de(s) entreprise(s), d'autre part, ont mené à une réflexion

différente sur le(s) texte(s) spécialisés. Déjà en 1990, informaticiens et linguistes se sont réunis (la frontière entre les sciences exactes et celles humaniste est franchie et une nouvelle approche sur interdisciplinarité commence à voir le jour) pour interroger les modes de prise en compte de textes dans la construction de terminologies. Cette rencontre a mené à la définition du concept de base de connaissances terminologiques (BCT), structure de représentation qui associe à un réseau de concepts des termes et des textes justifiant l'organisation du réseau (Meyer *et al*., 1992).

L'informatisation des textes, le développement et l'envergure de l'Internet, ont pour les sciences de l'information un effet important. La première nous aide à accéder à des données existant sous forme matérielle, le second a introduit la notion de commerce dans le domaine qui était plutôt considéré comme un travail intellectuel. À partir des années 1990, les documentalistes et les terminologues ont été confrontés aux questions érigées par les effets de l'informatisation des textes du traitement automatique des langues et de la représentation informatique des connaissances. L'informatique s'est approchée des disciplines comme la terminologie, et les tensions sont ressenties par les deux. Il est nécessaire de normaliser pour favoriser les échanges dans une langue ou entre les langues. Toutefois, normaliser signifie imposer une vision du monde. Un point d'équilibre doit être trouvé entre les deux disciplines, puisqu'elles ont un mode de représentation commun : des concepts reliés par des relations qui mènent à la construction d'un système.

Pour entreprendre une recherche terminologique, le terminologue réunit un ensemble de textes représentatifs du domaine étudié. Le corpus est l'ensemble constitué par ces textes. Un tel ensemble doit répondre aux conditions suivantes pour former un corpus (L'Homme, 2004) :

- il constitue un ensemble de données linguistiques (des mots, des phrases, des morphèmes etc.);
- les données linguistiques doivent apparaître dans un environnement naturel (des mots combinés dans des phrases, les phrases agencées dans des textes etc.); le corpus diffère des dictionnaires dans le sens que ceux-ci sont le résultat d'analyse faite par des spécialistes et reflètent un choix fait par eux;
- la sélection des textes doit reposer sur des critères explicites et permettra à un tiers d'interpréter les généralisations faites à partir du corpus;
- l'ensemble des textes est représentatif et doit être assemblé en fonction de l'élément à étudier, comporter un nombre suffisamment élevé d'occurrences de cet élément.

De nos jours, certaine entreprises, concernées par le traitement automatique des langues (TAL), ont confectionné des corpus dont certains atteignent une taille impressionnante. Certains d'entre eux peuvent être acquis, d'autres sont interrogeables par l'intermédiaire d'une interface Web.

Les corpus sont utilisés dans différentes communautés professionnelles, techniques et scientifiques. Chacun (littéraires, linguistes, terminologues, lexicologues et linguistes informaticiens) s'en sert pour obtenir les segments de textes correspondant à un thème. Il existe aujourd'hui une volonté de mieux définir et d'unifier les méthodologies de compilations de corpus pour l'observation de données linguistiques contribuant à la linguistique du corpus. Pourtant, chaque projet terminologique entraîne la confection d'un nouveau corpus, même si on peut récupérer une partie des textes ayant servi à un projet antérieur. En plus, les corpus de grande taille construits par les lexicologues contiennent des textes spécialisés, mais leur caractérisation n'est par assez raffinée pour être utile au terminologues.

«La valeur d'une recherche terminologique est directement fonction de la qualité de la documentation qui la fonde» (Dubuc, 2002). Le corpus doit constituer un ensemble représentatif de données linguistiques observables dans leur environnement naturel. En effet, toute la recherche terminologique s'organise à partir d'un corpus. La sélection rigoureuse des textes est garantie de la qualité de la recherche et il convient de passer un temps à structurer un corpus spécialisé.

En premier lieu, il faut sélectionner des textes spécialisés qui portent sur le domaine et qui contient les termes spécifiques. Les textes contenus dans le corpus doivent répondre à certains critères. Ils ont été définis par Marie-Claude L'Homme dans *La terminologie : principes et techniqu*es et nous les reprendront tel quels :

- *Domaine de spécialité* – les textes choisis doivent refléter le mieux possible le domaine ou le sous-domaine délimité au moment de la définition;
- *Langue(s)* – la sélection sera faite dans chacune des langues constituant l'objet de la description;
- *Langue de rédaction* – les textes du corpus ne doivent pas être des traductions, sinon, les traductions choisies doivent refléter l'usage réel dans le domaine;
- *Niveau de spécialisation* – il est défini en fonction de l'auteur du texte et des destinataires. Pearson (1998) a identifié les niveaux suivants : a) expert à expert (article dans une revue scientifique); b) expert à un expert dans un domaine connexe; c) didactique (texte s'adressant à des spécialistes en devenir; d)

vulgarisation (texte écrit par un expert ou un non-expert qui s'adresse à une personne ne possédant pas *a priori* les connaissances contenus dans le texte).

- *Type de document* – la forme de la publication est un reflet du niveau de spécialisation; on distingue des types de documents comme : manuel pédagogique, norme, catalogue, monographie, article scientifique, guide d'utilisation, rapport, actes et d'autres.
- *Suppport* – la recherche terminologique s'appuie sur des textes écrits, d'autant plus dans un contexte ou l'on fait appel a des traitements automatiques;
- *Date de parution* – les textes plus récents sont privilégiés;
- *Données évaluatives* – il existe des critères de nature évaluative, comme la renommée de l'auteur ou de la publication ou de la maison d'édition.

Nous pouvons constater que la sélection des textes repose sur des critères rigoureux, néanmoins, la taille et l'équilibre du corpus ne doivent pas être négligés. Les textes doivent contenir vraisemblablement les termes qui intéressent les terminologues ainsi que des renseignements sur ces termes. Comme le texte spécialisé porte sur un sujet ciblé, alors, il fait appel à un nombre limité de termes. Le corpus est équilibré lorsqu'il assure une certaine représentativité.

Pearson (1998) affirme que les types de textes qui offrent la meilleure explication des termes et les relations entre eux sont les textes qui assurent une communication de l'expert au spécialiste en devenir (novice), contrairement à la communication du type expert-expert où l'information peut rester implicite. La communication de l'expert envers son disciple tente de fournir toutes les notions pour une meilleure compréhension. Les textes écrits destinés à la communication de l'information contiendront un grand nombre de relations sémantiques entre les concepts (synonymie, hyperonymie, métonymie), relations exprimées d'une manière explicite.

Il existe aussi des corpus qui réunissent des textes en deux ou plusieurs langues, en d'autres termes, des corpus bilingues ou multilingues conçus pour des besoins de traduction. Ces corpus permettent aux terminologues de retrouver plus rapidement les correspondances interlinguistiques et font l'objet de traitements automatiques de langue spécifiques (TAL).

Les corpus multilingues peuvent être des *corpus alignés* et des corpus comparables. Les premiers réunissent des textes de plusieurs langues dont une partie constitue la traduction de l'autre. Leur réalisation repose sur l'établissement de correspondances entre les composantes formelles des textes. Les segments choisis sont alignés, l'un à côté de l'autre pour faciliter la consultation. La figure suivante montre comment deux courts textes sont alignés.

A **term** is the designation of a defined concept in a special language by a linguistic expression; it may consist of one or more words (i.e. simple term or complex term), or even may contain symbols. A **word** is the smallest linguistic unit conveying a specific meaning and capable of existing as a separate unit in a sentence; in a written text, it is marked off by spaces or punctuation marks before and after; affixes and endings are not words. (The simple term "vehicle" has one word; the complex term "police vehicle" has two words. After an unsatisfactory "term" search, a "word" search activates the search for a word — and its other-language equivalent — that might not be located at the beginning of a term but rather somewhere inside a complex term.)	Un **terme** est la désignation au moyen d'une entité linguistique d'une notion définie dans une langue de spécialité; il peut être constitué d'un ou de plusieurs mots (terme simple ou terme complexe) et même de symboles. Un **mot** est la plus petite unité signifiante qui peut exister de façon autonome dans une phrase; dans un texte écrit, il est délimité par des blancs ou par des signes de ponctuation; les affixes et les désinences ne sont pas des mots. (Le terme simple « véhicule » n'a qu'un mot; le terme complexe « véhicule de police » a trois mots. Après qu'une recherche par « terme » n'ait pu donner satisfaction, la recherche par « mot » peut offrir un mot — et son équivalent dans l'autre langue — qui pourrait se trouver non pas au début, mais à l'intérieur d'un terme complexe.)

Les corpus alignés peuvent être produits automatiquement grâce à des programmes nommés *aligneurs*. Ceux-ci s'appuient sur les frontières de la phrase (le point, le point d'interrogation et les retours) ou sur des éléments formels (les limites des paragraphes ou la numérotation des sections). L'alignement porte sur un seul couple de langues et donne lieu à un *bitexte.*

Les corpus comparables sont composés de deux ensembles de textes qui possèdent des caractéristiques communes. Ils peuvent appartenir à une seule langue, mais ceux en plusieurs langues sont plus utiles surtout pour les traducteurs. Ils se distinguent des corpus alignés car les textes qui les composent ne constituent pas de traductions, ni dans la première ni dans la seconde langue. La parenté des textes dans ce type de corpus est définie en fonction des critères différents : le même niveau de langue, la même tranche chronologique, la thématique abordée, à savoir le domaine de spécialité ou la subdivision d'un domaine générique. En vertu de la thématique commune, les textes doivent présenter un nombre très élevé de similitudes.

Les textes peuvent faire l'objet d'un enrichissement avant d'être interrogés ou exploités par d'autres formes de traitement automatique. Une technique couramment utilisée est *l'étiquetage*, qui consiste à attacher à une chaîne de caractères dans un texte, un renseignement de nature linguistique. Cette technique est très utile puisque les mots graphiques peuvent revêtir plusieurs sens et parfois joue le rôle de plus d'une partie de

discours. Dans ce cas, l'ambiguïté doit être enlevée mais, pour l'informaticien, cette tâche n'est pas des plus faciles. On marque alors le nom [N], le verbe [VB], l'adjectif ou participe passé [ADJ.], préposition [prep], déterminant du groupe nominal [DET], relatif [REL] ou d'autres étiquettes peuvent y être ajoutée. La forme d'étiquetage la plus courante est l'étiquetage morpho-syntaxique. On voit apparaître d'autres techniques qui essaient de décrire une partie de la structure syntaxique d'une phrase ou qui attachent de l'information sémantique aux mots.

La vision de la terminologie est par essence normalisatrice. Elle part de l'idée que la langue dans le domaine spécialisé peut être un moyen de communication perfectible, d'où la nécessité de normaliser pour éviter les créations individuelles menant souvent à de mauvaises compréhensions. Cette approche favorise les échanges entre industrie dans une même langue ou dans différentes langues. La construction des terminologies se fait par l'interrogation des experts, censés dresser les listes de concepts et de termes dans leur domaine de compétence.

L'utilisation des corpus est un moyen d'accéder aux connaissances d'un domaine en complément ou à la place de l'expertise humaine. L'utilisation des textes du corpus comme sources de connaissances a pris un grand essor à partir du début de ce millénaire. Les avantages de cette approche sont les suivants : une automatisation partielle (le traitement automatique des langues), une réduction des coûts, le renouvellement des hypothèses sur le statut des concepts et de leurs liens avec les termes. Cet usage des termes abandonne la vue normative et constructiviste des concepts, mais prend en compte des usages et des points de vue pour normaliser les concepts et les formaliser en fonction d'un objectif spécifique.

Grâce aux applications pour le web, les effets techniques et économiques se multiplient. Les architecture du «web sémantique» fait appel à des *ontologies* qui doivent fournir des représentations partagés par des actants logiciels, des méta-données pour annoter ou indexer des documents et assure la mise à disposition des connaissances consensuelles.

Les ontologies ont été définies pour mieux réutiliser les connaissances du domaine, de les gérer séparément et de faciliter l'échange des connaissances. Pour communiquer, affirmait Gruber, ces systèmes requièrent des représentations du monde compatibles et cohérentes, la recherche d'invariants dans un domaine, d'une description générique des connaissances.

L'ontologie est définie comme une spécification normalisée représentant des classes des objets reconnus comme existant dans un domaine et s'occupe des concepts, de leur définition par le biais des relations sémantiques et de leur pertinence pour la restitution des résultats aux utilisateurs. En précisant les ontologies, les concepts renvoient aux connaissances exprimées à travers le langage et doivent être définis en tenant compte des termes du domaine et de leur sémantique.

Le Traitement Automatique des Langues (TAL) est formé de logiciels qui peuvent produire des terminologies. Un logiciel de telle sorte n'est pas un révélateur de la sémantique des textes, au contraire, il est un moyen d'automatiser les recherches ciblées qui contribuent à reconstruire une sémantique. Le corpus est à la fois l'objet sur lequel portent les traitements automatiques, la justification de leur pertinence et la source d'information qui contribuent à les interpréter et exploiter.

Les logiciels de TAL peuvent accomplir les tâches suivantes :

- *Acquisition des termes*. Ils permettent l'extraction à partir du corpus analysé des termes candidats, en d'autres termes, des mots ou groupes de mots susceptibles d'être retenus comme termes et de fournir des étiquettes de concepts. Les types de techniques mise en œuvre sont soit morpho-syntaxiques, soit statistiques, soit d'une autre nature.
- *Structuration de termes et regroupement conceptuel*. Les outils de classification automatique de termes et les outils de repérage de relations se trouvent dans cette catégorie. La classification des termes est une méthode capable d'identifier les concepts ou l'association de termes à des classes. Le repérage de relations sémantiques permet la mise en relation des concepts. Les structures hiérarchiques sont les plus fréquemment rencontrées.

Nous ne pouvons pas conclure notre brève caractérisation du corpus sans nous arrêter sur le concept de *Knowledge-Rich Contexts* (KRC), Contextes riches en connaissances, concept introduit par Meyer *et al* (1998), fort important pour les terminologues, car ce type de contextes contient des termes dans un domaine spécialisé ainsi que des modèles (patterns) de connaissances. Cette relation entre le modèle de connaissance et les termes est d'un grand support pour la compréhension et l'établissement des relations conceptuelles dans lesquelles les termes apparaissent. Les textes sont considérés «de bons textes» lorsqu'ils sont riches en relations sémantiques, surtout en relations paradigmatiques. La richesse des textes sera d'autant plus grande si les modèles de connaissances apparaissent dans des contextes sémantiques simples et relevants pour le domaine.

En guise de conclusion, nous pouvons affirmer que la gestion du corpus dans la recherche terminologique a une dimension interdisciplinaire. L'évolution trop rapide des contextes modifie les besoins et les usages langagiers. Toute gestion de corpus, soit-elle automatique et réalisée par des logiciels performants, fait appel à une interprétation humaine, autrement, toute démarche terminologique tombera dans le piège de l'utilisation massive du web qui conduit à s'interroger sur les possibilités de contrôler les textes.

Bibliographie

BACHIMONT, B., *Engagement sémantique et engagement ontologique : conception et réalisation d'ontologie en ingénierie des connaissances. Évolution récente et nouveaux défis*, Eyrolles, Paris, 2000.

CABRÉ, M. T., *La terminologie. Théorie, méthode et application*, Armand Colin, Paris, Les Presses de l'Université d'Ottawa, 1998.

DUBUC, R., *Manuel pratique de terminologie*, 4ème éditions, Linguatech, Montréal, 2002.

GRUBER, T. R., « A translation approach to portable ontology specification » in *Knowledge Acquisition*, 5, 1993.

L'HOMME, M.-C., La terminologie : principes et techniques, Les Presses de l'Université de Montréal, Montréal, 2004.

QUILLIAN R., MINKI, M. (ed), «Semantic memory», in *Semantic Information Processing*, Cambridge, Mass. M. I. T. Press, 1968.

MEYER, T., DOUGLAS, S., BOWER, L., ECK, K., *Towards a new generation of terminological knowledge base*. Proceedings 16th International Conference on Computational Linguistics, COLING, Nantes, 1992.

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